UNIVERSIDAD DE GRANADA
DEPARTAMENTO DE DIDACTICA DE LAS MATEMATICAS

SIGNIFICADO DE LA CORRELACIÓN Y REGRESIÓN PARA LOS ESTUDIANTES UNIVERSITARIOS

Tesis doctoral

Francisco Tomás Sánchez Cobo

Granada, Octubre de 1.998

UNIVERSIDAD DE GRANADA
DEPARTAMENTO DE DIDACTICA DE LAS MATEMATICAS

SIGNIFICADO DE LA CORRELACIÓN Y REGRESIÓN PARA LOS ESTUDIANTES UNIVERSITARIOS

Tesis doctoral

Francisco Tomás Sánchez Cobo

Granada, Octubre de 1.998

SIGNIFICADO DE LA CORRELACIÓN Y REGRESIÓN PARA LOS ESTUDIANTES UNIVERSITARIOS

Tesis doctoral

Memoria realizada bajo la dirección de la Dra. Dª Mª Carmen Batanero Bernabeu y del Dr. D. Antonio Estepa Castro, que presenta el Licenciado en Ciencias Matemáticas Francisco Tomás Sánchez Cobo para optar al Grado de Doctor.

Granada, 4 de Octubre de 1.998

Fdo.- Francisco T. Sánchez Cobo

V°B°
La Directora

Fdo.- MªC.Batanero Bernabeu

V°B°
El Director

Fdo.- A. Estepa Castro

Este trabajo ha sido realizado en el marco del Proyecto de Investigación PB97-0851, subvencionado por la Dirección General de Enseñanza Superior del Ministerio de Educación y Cultura.

A Mª José, Francis y Carlos
que soportaron mis desvelos,
calmaron mis inquietudes
y compartieron mi entusiasmo.

A mi querida madre María y a la grata memoria
de mi querido padre Francisco, con la íntima
satisfacción de la promesa cumplida.

Deseo hacer público reconocimiento de gratitud y afecto:

- A los Doctores Dª Mª del Carmen Batanero Bernabeu y D. Antonio Estepa Castro, porque con su estímulo me han dado fuerzas y me han hecho el trabajo ligero.

- A los Doctores D. Luis Rico Romero y D. Juan Díaz Godino, porque aupado en sus ideas he divisado horizontes más lejanos en el mundo de la Didáctica de la Matemática.

- A los profesores del Seminario de Investigación del Departamento de Didáctica de la Matemática de la Universidad de Granada, porque con sus aportaciones han enriquecido esta investigación.

- A los profesores del Departamento de Matemáticas de la Universidad de Jaén por su aliento, en especial al Profesor Dr. D. Francisco Javier Muñoz Delgado y a la Profesora Dª Mª Francisca Molina Alba cuya amistad es una ayuda inestimable.

- A mis hermanos José Manuel y Vicky, Mª Jesús y Felipe por su incondicional apoyo moral.

- A la grata memoria del Profesor D. Gonzalo Sánchez Vázquez, por su generosa amistad.

- A los profesores Dr. D. Luis Parras Guijosa, Dr. D. Jesús Navarro, D. Antonio Conde, Dª Antonia Oya, Dª Mª Rosa Fernández y Dª Esther García por su desinteresada colaboración.

ÍNDICE

Capítulo 7. Resolución de los problemas

Capítulo 8. Aportaciones y líneas de investigación abiertas

ÍNDICE DE TABLAS

Anexo VI. ***Tablas de respuestas de los alumnos a los items de opciones múltiples***

Anexo VII. ***Tablas de respuestas de los alumnos a los problemas***

Introducción

En la actualidad asistimos a un desarrollo notable de la estadística, lo cual supone que, entre otras aportaciones, se proporcione múltiples herramientas a todas las ciencias naturales y sociales. Puesto que uno de los pilares básicos de nuestra sociedad es la información, los nuevos currícula de matemáticas subrayan, como objetivo prioritario, el desarrollo de la habilidad del tratamiento de la información estadística, que permite una mejor comprensión de nuestro entorno y la participación efectiva en la sociedad moderna (Gal y Garfield, 1.997). La estadística tiene como objeto dar métodos para el tratamiento de las masas de datos y su aplicación a la toma de decisiones (Nortes Checa, 1.987). Es en este marco en el cual *"la estadística se vuelve tan importante, o incluso más, que las matemáticas tradicionales"* (Burrill y cols., 1.992, pág. 1), como se deduce del hecho de que, en los países desarrollados, se imparten cursos de introducción a la estadística y la probabilidad en el primer ciclo de la mayoría de las titulaciones universitarias.

Como consecuencia de las mayores demandas de formación, la investigación en educación estadística, cuya existencia es relativamente reciente (Hawkins, 1.991), está experimentando una gran expansión, que se muestra en las

publicaciones, asociaciones de profesores e investigadores y congresos propios de este área. Un núcleo de interés especial, dentro de la educación estadística, es el *razonamiento estadístico* (Schuyten, 1.991; Rubin, 1.989), entendiendo como tal *"la forma en que la gente razona con ideas estadísticas y dan sentido a la información estadística"* (Garfield, 1.998, pág. 781). Este tipo de razonamiento conllevaría la reducción y representación de datos, interpretación de resúmenes estadísticos, toma de decisiones y la realización de inferencias a partir de los mismos. Subyacente a estos procesos se encuentra la comprensión de conceptos muy relevantes, entre los cuales podemos destacar el de la asociación, por el cual nos interesamos en este trabajo.

La lógica del razonamiento estadístico puede tomar diferentes enfoques, que dependen del objetivo perseguido. En nuestro trabajo, nos interesa el estudio estadístico descriptivo, en el cual los resultados obtenidos se limitan a los datos que se poseen. Sobre ellos se podrá establecer un orden, investigar conexiones entre las principales características de orden cuantitativo, temporal, de simultaneidad, coincidencia, exclusión mutua, etc. *"De este estudio puede deducirse si un fenómeno acompaña a otro, si los cambios de uno guardan relación con los otros, etc. Una tarea no menos importante es el análisis de los medios lingüísticos de expresar conexiones empíricas y el paso de un lenguaje a otro; por ejemplo, el paso de una tabla a una fórmula o a un gráfico"* (Gutiérrez Cabria, 1.994, pág. 104).

En particular, nos interesamos por el estudio descriptivo de la correlación y regresión, y, más concretamente, por el *razonamiento correlacional* (Ross y Smyth, 1.995), que incluye los juicios e interpretación de una relación entre dos variables a partir de una descripción verbal, unos coeficientes numéricos, unas tablas o unos diagramas de dispersión. Este razonamiento cobra un papel fundamental en la investigación científica y técnica, así como en la toma de decisiones en los campos más diversos. Un aspecto a destacar es que, a pesar de esta importancia y de que las investigaciones psicológicas muestran que la habilidad de detectar e interpretar la correlación no es general en los sujetos adultos, el tema no ha recibido una atención adecuada por parte de los investigadores en educación matemática.

Es por ello que hemos elegido este tema para centrar nuestro trabajo, que se inscribe en el Programa de Doctorado del Departamento de Didáctica de la Matemática de la Universidad de Granada, y en la línea de investigación en educación estadística, continuando, por tanto, las investigaciones de Estepa (1.990, 1.994), Navarro-Pelayo (1.991, 1.994), Vallecillos (1.992, 1.994), Serrano (1.993, 1.996), Ortiz (1.996), Sánchez Cobo (1.996) y Cañizares (1.997).

Las preguntas que sobre este tema nos formulamos son las siguientes:

- ¿Cuál es el significado de la correlación y regresión que se presenta a los alumnos en un curso introductorio de estadística descriptiva?

- ¿Qué elementos de significado podemos identificar en esta presentación? ¿Qué propiedades, definiciones, ejemplos y ejercicios se muestran a los alumnos?¿Cuál es la metodología de presentación del tema?¿Podemos identificar sesgos o vacíos en este significado?

- ¿Qué significados construyen los alumnos sobre la correlación y regresión al finalizar un curso introductorio de estadística descriptiva?¿Cómo relacionan los conceptos de covariación, dependencia, covarianza, correlación, regresión? ¿Podemos identificar errores conceptuales sobre las nociones anteriores?

- ¿Son capaces los alumnos de estimar la correlación a partir de diversas representaciones de la misma (descripción verbal, tabla, gráfico, coeficiente)?¿Qué traducciones establecen entre estas diversas representaciones?¿Qué estrategias emplean?¿Cómo afectan la intensidad y signo de la correlación, el tipo de covariación y las teorías previas a estas estimaciones y estrategias?

- ¿Son capaces los alumnos de aplicar sus conocimientos a la resolución de problemas?¿Qué estrategias siguen?¿Qué dificultades se presentan en el cálculo, interpretación y predicción?

En esta Memoria tratamos de dar respuesta, al menos parcial, a las anteriores preguntas, conscientes de que se ha abordado una perspectiva muy amplia sobre el problema planteado.

La Memoria se ha organizado en los siguientes capítulos:

En el primer capítulo se describen los objetivos concretos planteados y la metodología utilizada en la investigación, así como el marco teórico que incluye los conceptos de significados personales e institucionales, juego de cuadros, representaciones y procesos de traducción. El primero de ellos tiene una proyección generalizada sobre toda esta Memoria, mientras que los últimos se centran más sobre los procesos de traducción de una representación de la correlación a otra, para lo cual tomamos como referencia los trabajos de Janvier (1.978, 1.987) sobre las funciones.

En el segundo capítulo hacemos una recensión de las investigaciones más significativas que se han realizado sobre el tema de la asociación en las cuatro últimas décadas. Se han expuesto, no sólo, las aportaciones que se han realizado desde el paradigma psicológico, sino, también, los escasos trabajos que se han desarrollado dentro del paradigma de la educación estadística, siendo realizados, principalmente, en nuestro Departamento y de los cuales esta Tesis es una continuación directa.

El capítulo tercero se dedica a la descripción de la enseñanza que han recibido los sujetos de la muestra en el curso en el que se tomaron los datos, y, además, la forma en que se introduce el tema en la enseñanza secundaria. Para ello hemos resumido y revisado los resultados de la investigación de Sánchez Cobo (1.996) sobre el análisis de la correlación y regresión en los libros de texto de bachillerato, tanto desde una vertiente de la presentación teórica del tema como desde el estudio de los ejercicios que se incluían en este tópico en los manuales. Asimismo, se han analizado los apuntes de clase del profesor y de dos alumnas del grupo en el que él da clase, con el fin de caracterizar la enseñanza que sobre la asociación reciben los alumnos de los primeros cursos universitarios. Como resultado se obtiene, igualmente, una lista de elementos de significado de la correlación y regresión, que nos ha servido de base para la construcción del cuestionario empleado en la evaluación de los conocimientos de los alumnos.

La construcción y diseño del cuestionario se aborda en el capítulo cuarto, describiéndose todas y cada una de las fases que se han seguido para su elaboración, así como las variables de tarea que han sido tenidas en cuenta. El

instrumento de evaluación consta de cuatro bloques bien diferenciados, que nos permitirán analizar la muestra estudiada y evaluar el conocimiento conceptual y procedimental, así como la capacidad de traducción de diversas representaciones de la correlación.

Los tres restantes capítulos describen los resultados de la evaluación llevada a cabo con este instrumento, sobre una muestra de 193 alumnos. El análisis comprende comparación de porcentajes de respuestas correctas e incorrectas, puesta en relación de elementos de significado correspondiente a diversos items, especificación de errores y estrategias de los alumnos, estudio cuantitativo del error en la estimación del coeficiente de correlación y análisis del efecto de las variables de tarea del cuestionario sobre estas estimaciones y estrategias de los alumnos. Creemos que podemos afirmar que un estudio tan pormenorizado como el que realizamos, no se ha efectuado anteriormente en las investigaciones sobre correlación, ni en el campo de la psicología ni, tampoco, en el de la educación matemática. Puesto que el número de investigaciones sobre la enseñanza de la estadística a nivel universitario es, todavía, insuficiente, pensamos que nuestros resultados proporcionan una información valiosa, tanto a los profesores de estadística como a los autores de libros de texto de esta disciplina en este nivel docente.

Finalizamos la Tesis presentando las principales aportaciones de nuestro estudio, respecto a los objetivos planteados. Las implicaciones para la enseñanza que se derivan, así como una reflexión sobre las líneas de investigación que quedan abiertas para los futuros trabajos que deseen continuar nuestra labor.

Tras la bibliografía utilizada en esta investigación, se concluye con la inclusión de los anexos sobre la muestra, los libros de texto empleados en la investigación, los apuntes de clase del profesor y de dos alumnas, el cuestionario utilizado en la investigación y las distribuciones de las respuestas de los alumnos a los items de opciones múltiples y a los problemas.

Capítulo 1

Problema de investigación y metodología

1.1. INTRODUCCIÓN

El presente trabajo se incluye en la línea de investigación en educación estadística del Departamento de Didáctica de la Matemática de la Universidad de Granada, y continúa las investigaciones sobre la didáctica de la asociación estadística llevadas a cabo por Estepa (1.990, 1.994), Batanero, Estepa y Godino (1.991, 1.995, 1.997, 1.998), Batanero, Estepa, Godino y Green (1.996), Batanero, Godino y Estepa (1.998), Estepa y Batanero (1.994, 1.995, 1.996), Batanero y Godino (1.998) y Sánchez Cobo (1.996).

La asociación es un concepto fundamental, no sólo por su relevancia en estadística, sino porque extiende la idea de dependencia funcional a situaciones aleatorias. Por otro lado, las investigaciones psicológicas muestran que el razonamiento correlacional intuitivo es, en general, pobre (Nisbett y Ross, 1.980). De ahí la importancia de realizar la enseñanza de estos conceptos y su influencia sobre el aprendizaje de los alumnos.

Son, no obstante, muy escasas las investigaciones, que desde un paradigma didáctico, están centradas en este tema. Nuestro objetivo general es contribuir a llenar este hueco, así como a la mejora de la planificación de la enseñanza del

concepto de asociación, en particular en los cursos de estadística descriptiva a nivel universitario.

Dentro de este objetivo general, en este capítulo describimos, con mayor detalle, los objetivos específicos formulados y la importancia que tienen para la Didáctica. Asimismo, fundamentamos el estudio mediante la descripción del marco curricular y teórico del estudio.

El capítulo se completa con la descripción de la metodología empleada en la parte experimental del estudio, la cual combina elementos cuantitativos y cualitativos.

1.2. OBJETIVOS DE LA INVESTIGACIÓN

Nuestro proyecto inicial era analizar el aprendizaje de la correlación y regresión al iniciarse el estudio del tema, que, teóricamente, se debería llevar a cabo en bachillerato. No obstante, aunque los planes de estudio anteriores a los actuales preveían que el estudio de la asociación estadística se iniciase, desde un punto de vista descriptivo, en el tercer curso de bachillerato, la realidad es que muchos profesores omitían su enseñanza, por diversos motivos. En el C.O.U., solamente, los alumnos de la opción C (Ciencias Sociales) y Opción D (Humanística y Lingüística) estudiaban este tema, y, únicamente, por espacio de unos 15 días, aproximadamente, dentro de la asignatura Matemáticas II, pero no el resto de las especialidades. Puesto que la mayor parte de las carreras universitarias incluyen una asignatura de estadística en primer o segundo curso, la asociación se estudia, desde un plano descriptivo, en todas las especialidades que cubren dicha materia. Podemos considerar, en consecuencia, que en la práctica el primer contacto de los alumnos con la correlación y regresión se lleva a efecto en el curso de estadística que suele incluirse en la mayoría de las titulaciones universitarias. Es en este nivel de enseñanza en el que se ha centrado nuestra investigación.

Los trabajos de Batanero, Estepa y Godino se circunscribieron a las concepciones iniciales de los alumnos y a su evaluación después de experimentos

de enseñanza basados en el uso de ordenadores. La evaluación de estas experiencias se centró, principalmente, en la capacidad que los alumnos participantes mostraban en la resolución de problemas y no en la comprensión de los conceptos teóricos. Por otra parte, la realidad es que los cursos universitarios de estadística basados en el uso de ordenadores son aún escasos y es más frecuente la enseñanza tradicional, apoyada en la exposición magistral del profesor y el trabajo personal de los alumnos. Mientras en los experimentos de Estepa y colaboradores el número de estudiantes era muy reducido (20-30 alumnos por clase), los cursos de estadística, a nivel universitario, suelen ser numerosos. Todo ello hace plausible que las dificultades de los alumnos en la comprensión de las nociones de correlación y regresión sean mayores que las detectadas por las investigaciones de Estepa, Batanero y Godino. Nuestro interés se centra en evaluar los conocimientos de los alumnos que asisten a este tipo de cursos sobre la correlación y regresión, una vez finalizado el curso introductorio de estadística en la universidad.

Más concretamente, como consecuencia de los citados estudios de Estepa y colaboradores y de nuestro propio estudio inicial, se identificaron varios tipos de cuestiones para continuar su investigación -que componen el objeto de este trabajo- y que destacamos a continuación:

a) Un primer fin es analizar los contenidos incluídos en el estudio descriptivo de la correlación y regresión tanto en Bachillerato, como en los cursos "típicos" de iniciación a la estadística en la universidad. Este estudio se llevará a cabo a partir del análisis de libros de texto y de la programación que ha sido suministrada por los profesores de las asignaturas y de los apuntes tomados en clase por los alumnos.

Juntamente, el análisis pretende identificar los elementos de significado asociados al mismo, en la línea iniciada en las investigaciones de Vallecillos (1.994) y Ortiz (1.996). Ello nos permitirá, asimismo, describir el significado institucional de la asociación estadística en un curso usual de Estadística descriptiva de nivel universitario y será la base para la construcción de los instrumentos de evaluación, ya que como indican Giménez y cols. (1.997) la evaluación no tiene sentido si no es dentro de un currículo. Además, puede servir

para el diseño de otros cursos sobre estos contenidos, así como materiales curriculares para la enseñanza.

b) Un segundo objetivo es caracterizar el significado personal que los alumnos universitarios dan a la correlación y regresión estadísticas al finalizar un curso de introducción a la Estadística en la universidad. En particular, estamos interesados en describir los errores conceptuales y procedimentales de estos estudiantes, la estimación que hacen los alumnos del coeficiente de correlación a partir de distintas representaciones de la correlación (verbal, tabla, gráfico) y la capacidad de traducción entre estas representaciones. En consecuencia hemos construido un cuestionario en el que se usan algunos de los items del instrumento empleado por Estépa (1.994), otros tomados de Morris (1.997), Jennings Amabile y Ross (1.982) , Cruise, Dudley y Thayer (1.984) y Tversky y Kahneman (1.982a) y otros de construcción propia.

El instrumento tiene interés en sí mismo, ya que puede utilizarse en la evaluación final de los conocimientos de otros alumnos o de los conocimiento iniciales en cursos más avanzados de Estadística.

Usando estos intrumentos, llevaremos a cabo un estudio de evaluación sobre una muestra amplia de alumnos universitarios y caracterizaremos sus conocimientos sobre la correlación y regresión. La evaluación se centra en los puntos siguientes:

- Comprensión de las propiedades más sobresalientes de la covariación, dependencia estadística, covarianza, coeficiente de correlación lineal y recta de regresión. Relaciones que los alumnos establecen entre los conceptos anteriores.
- Estimación de la correlación que hacen los alumnos a partir de diversas representaciones (verbal, gráfica y numérica).
- Interpretación de valores numéricos del coeficiente de correlación y construcción de situaciones asociadas, representadas en formal verbal y gráfica.

- Errores conceptuales asociados a elementos de significado relacionados con el coeficiente de correlación, covarianza, rectas de regresión, tipos de covariación y relaciones entre correlación y causalidad.

- Estrategias de los alumnos en el ajuste de una recta a un conjunto de datos, cálculo del coeficiente de correlación y estimación de valores de las variables.

- Interpretación del coeficiente de correlación y de las rectas de regresión en una situación problemática.

Todos estos puntos son componentes esenciales del significado matemático de las nociones de correlación y regresión. La evaluación que planteamos permitirá detectar desajustes entre el significado de los conceptos mostrados a los alumnos y el significado personal efectivamente construido. Como consecuencia se identificarán los puntos en los cuales es necesario reforzar la enseñanza del tema.

1.3. PERSPECTIVA CURRICULAR

El objeto de nuestro trabajo es la caracterización del significado que atribuye a la correlación y regresión una muestra de alumnos del primer curso universitario después de recibir instrucción sobre estos temas. Desde una perspectiva curricular, debemos tomar en consideración los elementos invariantes en toda reflexión o estudio sobre el currículum (Rico, 1.990):

a) Colectivo de personas al que se va a formar. En la sección 1.5 se especificará la muestra utilizada, junto con la población de la que se ha extraído dicha muestra.

b) Tipo de formación que se quiere proporcionar. En la sección 3.3 analizamos y en el anexo III presentamos los contenidos (respecto a la correlación y regresión) del curso de estadística impartido. Este tipo de cursos, coincide en nuestro caso, con los descritos por Gal y Garfield (1.997) como cursos genéricos y autosuficientes dirigidos a las aplicaciones de un sector del conocimiento. En

general, con estos cursos se cubrirían dos finalidades amplias, mediante las cuales los estudiantes lleguen a ser capaces de: i) Comprender y tratar con la incertidumbre, variabilidad e información estadística en el mundo que les rodea y participar de forma efectiva en una sociedad abrumada por la información, y, ii) contribuir y/o tomar parte en la producción, interpretación y comunicación de datos relativos a los problemas a los que se enfrentan en su vida profesional.

c) La institución social a través de la cual se llevará a cabo la formación. En nuestro caso es la Universidad de Jaén, por medio del Departamento de Estadística con su potencial docente y de recursos que le son propios.

d) Las necesidades a cubrir. En nuestro caso, este curso de estadística trata de formar a los alumnos de la Diplomatura de Ciencias Empresariales y de Enfermería desde una doble perspectiva: a) Que los alumnos adquieran los conocimientos de esta ciencia con el fin de que puedan ser aplicados a las Ciencias Empresariales (en el anexo III, se puede observar la sección dedicada a las aplicaciones económicas) y a todas aquellas materias de la Diplomatura de Enfermería que usen modelos estadísticos, y, b) que sirvan como fundamento para ampliar estudios en los que los conocimientos estadísticos adquiridos sean un prerrequisito conceptual.

e) El control al que va a estar sometido. El control a que está sometido este curso está bajo los auspicios del Departamento de Estadística con una doble motivo: La evaluación de los alumnos y la toma de datos para la reflexión y mejora de la enseñanza ofrecida curso tras curso.

1.4. SUPUESTOS TEÓRICOS

En esta sección describimos, muy brevemente, los conceptos teóricos usados en la investigación, que son los siguientes:

- Significados personales e institucionales de los objetos matemáticos
- Teoría de cuadros y dialéctica útil / objeto
- Representaciones y procesos de traducción entre las mismas

La actividad de resolución de problemas: significados personales e institucionales

Siguiendo a Godino y Batanero (1.994), consideramos la matemática como actividad de resolución de problemas, socialmente compartida, como lenguaje simbólico y como sistema conceptual lógicamente organizado. Estos autores caracterizan los problemas matemáticos por los "objetos matemáticos" (números, operaciones, ...) y las representaciones simbólicas que intervienen en el enunciado. La actividad matemática consiste en la búsqueda y generalización de las soluciones, la búsqueda de lo esencial en los diferentes contextos, las conexiones con otras situaciones, problemas y procedimientos y la actividad de simbolizar, formular, validar, generalizar, ... matematizar (Freudenthal, 1.991). Los problemas para las cuales puede ser válida una misma solución se pueden agrupar en *campo de problemas*.

En los procesos de resolución de problemas se emplean prácticas, entendiendo por tales cualquier actividad para resolver el problema, validar la solución, generalizarla o comunicarla a otras personas. Una práctica es significativa para una persona, cuando para dicha persona la práctica es pertinente para la resolución del problema.

Las diferentes instituciones matemáticas, las dedicadas a producir, utilizar o enseñar el saber matemático, comparten ciertas prácticas asociadas a diferentes campos de problemas, que son significativas, pertinentes y comúnmente aceptadas para resolverlos. Este tipo de prácticas se llaman *prácticas institucionales*. Por otro lado, se llaman *prácticas personales*, las que son significativas para una persona para resolver los problemas de un determinado campo. Algunas de estas suelen ser observables, otras suelen ser acciones interiorizadas no observables, por lo que necesitaremos indicadores empíricos de las mismas para su análisis.

Desde el punto de vista didáctico, esta diferenciación entre prácticas personales e institucionales permite recoger en una teorización común, tanto los procedimientos y las soluciones que en las instituciones de enseñanza del tema se consideran correctas, como las erróneas, inadecuadas o no totalmente correctas

que para los alumnos constituyen "buenas soluciones" de los problemas planteados.

A partir de la actividad de resolución de un campo de problemas y del sistema de prácticas asociado, se produce, en la institución, la emergencia progresiva de ciertos objetos, productos globales de las actividades que forman este sistema de prácticas. Estos objetos, en un momento dado son nombrados y se reconoce su entidad cultural. Sucesivamente son generalizados, dotados de propiedades, sistematizados y empleados en la resolución de nuevos problemas. El sistema de prácticas asociado al objeto se define como *significado del objeto institucional*.

Similarmente, el aprendizaje del sujeto se produce mediante la emergencia progresiva del *objeto personal* a partir del sistema de prácticas personales asociadas a una campo de problemas, las cuales constituyen el *significado personal del objeto*.

La diferenciación entre el objeto (emergente inobservable de un sistema de prácticas) y el significado del mismo (sistema observable de prácticas), que hacen estos autores, supone también el reconocimiento de la problemática de la evaluación de los conocimientos.

El sistema cognitivo del sujeto (su conocimiento conceptual y procedimental, sus intuiciones, representaciones, esquemas, ...), es decir la red de objetos personales construída en un momento dado, que constituye el significado personal, es una totalidad organizada y compleja, que, en general, no coincide totalmente con el significado institucional en una determinada institución. El estudio de las coincidencias entre el significado personal e institucional de un objeto matemático constituye el problema específico de la evaluación.

Ahora bien, el sistema cognitivo de un sujeto sobre un objeto matemático dado, en nuestro caso la correlación y la asociación, que constituye el significado personal, es una estructura conceptual compuesta de diversos elementos relacionados, conexionados y acomodados entre sí que denominaremos *elementos de significado*. Cada elemento de significado se caracteriza por un subconjunto de prácticas personales. Si este subconjunto de prácticas personales se identifica con el subconjunto correspondiente de prácticas institucionales diremos que el sujeto

tiene un conocimiento correcto del objeto matemático, en caso de que exista alguna discrepancia entre el subconjunto de prácticas personales e institucionales diremos que el conocimiento del sujeto es parcialmente correcto, o bien erróneo, según la intensidad de la discrepancia.

Teoría de cuadros y dialéctica útil / objeto

Otro marco teórico que utilizaremos en nuestro trabajo ha sido tomado de Douady (1.986). Para ella un cuadro está constituido por objetos de una rama de las matemáticas, las relaciones entre los objetos, sus formulaciones, -eventualmente diversas- y las imágenes mentales asociadas a estos objetos y relaciones. Las imágenes juegan un papel esencial en el funcionamiento como útil de los objetos del cuadro. Los cuadros más usuales son el cuadro algebraico, el cuadro numérico, el cuadro gráfico, el cuadro geométrico, etc. Dos cuadros pueden contener los mismos objetos y diferir por las imágenes mentales y la problemática desarrollada. Por otra parte, la familiaridad, la experiencia pueden conducir a conflictos entre lo que se espera y lo que se produce efectivamente y por consiguiente renovar las imágenes o hacerlas evolucionar. Por tanto, la noción de cuadro se puede concebir de manera dinámica. *Los cambios de cuadros* son un medio de obtener formulaciones diferentes de un mismo problema, que sin ser totalmente equivalentes, permiten un nuevo acceso a las dificultades encontradas y el uso de técnicas y procedimientos que no se imponen en la primera formulación. Cualquiera que sea la traducción de un cuadro a otro conduce a resultados no conocidos, a técnicas nuevas, a la creación de objetos matemáticos nuevos, en suma al enriquecimiento del cuadro origen y cuadros auxiliares de trabajo.

Una parte importante del trabajo de los matemáticos es interpretar los problemas que pretenden resolver, lo que conlleva observar, acometer, replantear el problema desde diversos puntos de vista.

Gran parte del trabajo del matemático consiste en plantear cuestiones y resolver problemas, lo que les lleva a producir *útiles* conceptuales. Terminado el trabajo y, por necesidad de transmisión a la comunidad científica, estos útiles se

descontextualizan y formulan de la manera más general posible, se integran en el cuerpo de conocimientos existente o sustituyen a algunos que existían con anterioridad, adquiriendo entonces el status de *objetos*. En consecuencia se puede decir que un concepto es un *útil* cuando centramos nuestro interés en el uso que se hace de él para resolver un problema. Un útil puede ser adaptado a varios problemas, varios útiles pueden ser adaptados a un mismo problema. Por *objeto* se entiende el objeto cultural, reconocido socialmente, que tiene su lugar dentro del saber científico, en un momento dado. Cuando un alumno resuelve un problema usa diferentes nociones matemáticas como útiles, cuando el alumno puede formularlos y justificar su empleo, se habla de útiles explícitos, en caso contrario de útiles implícitos.

Se dice que un alumno tiene conocimientos en matemáticas, cuando es capaz de provocar su funcionamiento como útiles explícitos en los problemas que se le plantean, haya o no en el enunciado del problema indicadores sobre la pertinencia de estos útiles, cuando es capaz de adaptarlos, aunque las condiciones habituales de empleo no se satisfagan completamente, y cuando los utiliza para proponer problemas o bien para plantear cuestiones con sus propias palabras.

Representaciones y procesos de traducción

Una investigación, desarrollada en otro ámbito matemático, que deseamos subrayar, por su utilidad para el presente trabajo, es la de Janvier. Dicho autor está interesado por las traducciones entre los diversos lenguajes de representación de las funciones, entendiendo por *"un proceso de traducción, el proceso psicológico involucrado al ir de un modo de representación a otro, por ejemplo, de una ecuación a un gráfico"* (Janvier, 1.987, p. 27). Un concepto fundamental en la anterior definición es el de *representación*, siendo factible divesas aproximaciones a dicho concepto (Rico, Castro y Romero, 1.996). En este sentido, estos autores indican que la noción de representación conllevaría dos entidades relacionadas y diferenciadas, entre las que podemos establecer una correspondencia, que son el *mundo representante* y el *mundo representado*, e implícitamente se presupone

algún tipo de conexión entre los objetos del mundo representante y el mundo representado.

Para Kaput (1.987) al especificarse una representación debemos descubrir las cinco entidades siguientes: a) El mundo representado, b) el mundo representante, c) los aspectos del mundo representante que lleva a cabo la representación, d) los aspectos del mundo representado que se representan, y, e) la correspondencia entre los dos mundos.

Para Duval (1.993) los objetos matemáticos nunca deben ser confundidos con la representación que se les hace, pues esto implicaría una pérdida en la comprensión. Pero si el objeto representado es el que importa, las representaciones son indispensables, ya que los objetos no son directamente accesibles a la percepción.

Una estrategia de investigación importante es emplear diversas representaciones para observar el aprendizaje de un concepto por los estudiantes y explicar, de forma satisfactoria, su construcción (Janvier, 1.978; Kaput, 1.987; Duval, 1.993, 1.995; Rico, Castro y Romero, 1.996).

Si primamos en la noción de función su faceta de dependencia entre variables, se puede apreciar los siguientes tipos de representaciones: i) Modelo físico o simulación, ii) descripción verbal, iii) tabla de valores, iv) gráfica, y, v) fórmula o ecuación, que se describen en Azcárate y Deulofeu (1.990). Cada una de ellas implica registros diferentes.

La descripción verbal nos ofrece un marco descriptivo y cualitativo de la dependencia funcional. Por el contrario, la tabla de valores enfatiza la dimensión cuantitativa de la función, eclipsando, por contra, los aspectos globales de ella. Sin embargo, la máxima información, tanto cualitativa como cuantitativa, que ofrece una función nos la facilita su representación gráfica o su fórmula o ecuación correspondiente, estando, cada uno de ellos, en concordancia con el lenguaje geométrico y algebraico, respectivamente.

La adquisición del concepto de función obligará, por tanto, a dominar, no sólo las distintas representaciones de esta noción, sino las traducciones que se

pueden establecer entre estos marcos. A este respecto, la Tabla 1.4.1, tomada de Janvier (1.987), hace visible la diversidad de tales procesos.

Tabla 1.4.1. Traducciones de las representaciones de la función

HACIA / DESDE	DESCRIPCIÓN VERBAL	TABLA	GRÁFICA	FÓRMULA
DESCRIPCIÓN VERBAL		Medida	Boceto	Modelo
TABLA	Lectura		Trazado	Aproximación
GRÁFICA	Interpretación	Lectura		Ajuste
FÓRMULA	Interpretación	Cálculo	Gráfica	

En este trabajo abordaremos el estudio de la correlación y regresión, que forma parte del concepto de asociación estadística. Esta noción amplia la idea de dependencia funcional, que puede considerarse un caso particular de la dependencia aleatoria.

Trataremos en el mismo de analizar la interpretación que los alumnos hacen de la noción de correlación a partir de diversas representaciones: La descripción verbal, representación numérica y gráfica y valor del coeficiente de correlación.

Para Duval (1.993) la conceptualización significa la coordinación de diversos registros de representación. Estos registros son complementarios en el sentido de que cada uno de ellos supone una selección de los elementos significativos del contenido que representan. La existencia de varios registros permite la economía en el tratamiento y trabajo matemático.

Las transformaciones de un registro de representación a otro no son, a veces, sencillas. Es más, algunos estudiantes ven a las diversas representaciones de un concepto como entidades diferentes que aludieran, por contra, a conceptos distintos, lo cual dificultaría, notablemente, los procesos de traducción (Sanz, 1.990).

Por otra parte, las traducciones pueden ser directas o indirectas, según que su desarrollo se efectúe en un único paso o que haya etapas intermedias. Un ejemplo de la primera clase sería la traducción de tabla a gráfica, mientras que de la segunda sería el pasar del coeficiente de correlación a la gráfica, que implicaría la realización de la tabla de valores. Es natural, por todo lo expresado con anterioridad, que las traducciones indirectas sean más complejas para los alumnos (Janvier, 1.987).

En el capítulo 4, dedicado a la construcción del cuestionario, volveremos sobre los procesos de traducción empleados en esta investigación.

1.5. DESCRIPCIÓN DE LA METODOLOGÍA

En esta investigación se ha empleado un paradigma metodológico mixto entre los métodos cuantitativos y cualitativos. El estudio teórico del contenido matemático y el análisis de los libros de texto y apuntes de los alumnos es de tipo cualitativo. Asimismo, analizaremos algunas variables cualitativas en las respuestas de los alumnos. Por otro lado, utilizaremos otras variables cuantitativas y análisis estadísticos de los datos.

La evaluación de los conocimientos de los alumnos se ha llevado a efecto a partir de sus respuestas escritas en un cuestionario. Podemos clasificar el método de recogida de estos datos como de medición, pues tratamos de obtener datos no directamente accesibles. Este cuestionario, sus objetivos, contenido y técnica de construcción se describen con detalle en el Capítulo 4.

Indicamos aquí, únicamente, que consta de tres partes:

a) Actividades de interpretación y estimación del coeficiente de correlación. Para asegurar la representatividad de los items en esta parte del cuestionario, hemos analizado las variables básicas intervinientes y hemos usado, para la construcción del cuestionario, un diseño experimental que ha tenido en cuenta estas variables. Son items de respuesta libre limitada.

b) Cuestionario de comprensión conceptual de algunos elementos de significado claves, formado por items de verdadero o falso con posibilidades de respuesta múltiple.

c) Dos problemas completos sobre correlación y regresión. Se trata de una prueba de ensayo y por tanto las respuestas son libres y de tipo cualitativo.

Población y muestra

Para la construcción de instrumentos de evaluación de los elementos de significado, se ha utilizado, en primer lugar, una muestra piloto de un número reducido de alumnos, con el fin de ajustar los cuestionarios y estudiar su fiabilidad.

Nuestro objetivo era caracterizar el conocimiento de los alumnos al finalizar la enseñanza sobre correlación y regresión en los primeros cursos universitarios. La población objetivo estuvo constituída por los estudiantes de las carreras de la Diplomatura en Empresariales y la Diplomatura en Enfermería de la Universidad de Jaén. La muestra de alumnos que cumplimentó el cuestionario estaba compuesta por 193 estudiantes de estas carreras, 104 (37 hombres y 67 mujeres) de la Diplomatura en Empresariales y 89 (20 hombres y 69 mujeres) de la Diplomatura en Enfermería. Por tratarse del primer curso universitario de estas Diplomaturas, la edad media de la muestra era de 20 años, teniendo el 87'9 % de la muestra 22 años o menos.

La forma de acceso a la universidad de esta muestra de alumnos ha sido la siguiente: 46 alumnos (24'0 %) de la opción A de COU, científico-tecnológica (ciencias puras); 42 alumnos (21'9 %) provienen de la opción C de COU, ciencias sociales (letras mixtas); 41 alumnos (21'4 %) de la opción B de COU, biosanitaria (ciencias mixtas); 21 alumnos (10'9 %) de Formación Profesional II, administrativa y comercial; 23 alumnos (12'0 %) de las distintas especialidades de Formación Profesional II Rama Sanitaria; 14 alumnos (7'2 %) de las diversas Especialidades del Bachillerato de la Reforma (LOGSE); 5 alumnos (2'6 %) tienen procedencia distinta a las anteriores, provienen de otras carreras o del acceso para mayores de 25 años.

En cuanto a los estudios de Estadística en cursos anteriores nos hemos encontrando que 117 alumnos (60'6 %) no habían estudiado Estadística en cursos anteriores; 32 alumnos (16'6 %) la habían estudiado en alguna de las especialidades de Formación Profesional; 16 alumnos (8'3 %) en C.O.U.; 14 alumnos (7'2 %) en Bachillerato; 14 alumnos (7'2 %) en otros curso, como el curso anterior - eran repetidores -, o bien en otros estudios cursados.

En el apartado C de las preguntas preliminares del cuestionario (véase anexo V) habíamos interrogado a los alumnos por el interés que tiene la asignatura de Estadística para su formación en la titulación que están estudiando, obteniéndose la Tabla 1.5.1. A partir de ella, podemos deducir que el interés que estos alumnos le dan a la asignatura de Estadística para la formación global en la carrera que cursan es elevado, sobre todo si se tiene en cuenta que no es una asignatura específica de la misma, sino que podíamos considerarla más bien una asignatura instrumental, dado que sus contenidos son herramienta para los conocimientos de dicha carrera.

Tabla 1.5.1. Frecuencia y porcentaje del grado de interés de los alumnos por la asignatura de Estadística

Grado interés	Frecuencia	Porcentaje	% acumulado
Mucho	12	6'3	6'3
Bastante	81	42'4	48'7
Suficiente	69	36'1	84'8
Poco	28	14'7	99'5
Muy poco	1	0'5	100'0
No responde	2		
Total	193	100'0	100'0

En cuanto al interés, que para estos alumnos, tiene el tema de la correlación y regresión para su formación en la titulación que estudian, los resultados son muy similares, aunque algo superiores, mostrándose en la Tabla 1.5.2.

Tabla 1.5.2. Frecuencia y porcentaje del grado de interés de los alumnos por la correlación y regresión

Grado interés	Frecuencia	Porcentaje	% acumulado
Mucho	15	7'9	7'9
Bastante	78	41'1	49'0
Suficiente	62	32'6	81'6
Poco	32	16'8	98'4
Muy poco	3	1'6	100'0
No responde	3		
Total	193	100'0	100'0

De lo anterior podemos inferir que, en general, los alumnos tienen bastante interés hacia la correlación y la regresión, a pesar de que estos temas no sean materias específicas de la carrera que cursan. Todos estos datos se presentan con mayor detalle en el Anexo I.

Variables

Variables dependientes

Puesto que se trata de una metodología que incorpora elementos cualitativos y cuantitativos, las principales variables dependientes de la investigación han sido de tipo cualitativo. Estas variables dependientes son las siguientes:

- Solución correcta o incorrecta de cada ítem de comprensión conceptual y errores específicos o comprensión de propiedades que ello implica.
- Estrategias de cálculo del coeficiente de correlación y de los parámetros de la recta de regresión.
- Discriminación entre las predicciones a partir de la recta de regresión de Y sobre X y de la recta de regresión de X sobre Y.

- Errores y sus tipos en el proceso de resolución de los problemas y en la interpretación de los coeficientes.

Entre las variables cuantitativas dependientes tenemos:

- Valor del coeficiente de correlación que el sujeto asigna a una distribución bidimensional dada mediante su descripción verbal, tabla numérica o diagrama de dispersión, o bien, valor que se deduce mediante el diagrama de dispersión que propone el alumno a partir de la descripción verbal o de un valor del coeficiente de correlación.
- Valores de los parámetros y estadísticos calculados en los problemas propuestos, incluyendo el coeficiente de correlación.
- Valor de las predicciones que se piden en los problemas.
- Errores absolutos en la estimación del coeficiente de correlación.

Variables independientes

Las variables independientes en las tareas de estimación e interpretación serán:

- Intensidad de la correlación: se consideran cinco intervalos que engloban desde la independencia hasta la dependencia funcional.
- Tipo de la función de ajuste: lineal o no.
- Signo de la correlación.
- Tipo de covariación: dependencia causal unilateral, interdependencia, dependencia indirecta, concordancia y covariación casual (Barbancho, 1.973).
- Teorías previas: coinciden o no con la correlación empírica de los datos.

Variables concomitantes

Se han controlado el contexto del ítem o del problema, que será familiar al alumno, la forma de presentación de los datos, el modo de plantear las preguntas, el número de datos y su tamaño y los distractores empleados en los items.

Fases de la investigación

La investigación ha constado de tres fases:

Primera fase

Análisis, desde un punto de vista matemático, de los temas sobre la correlación y regresión incluídos en los textos de bachillerato y en un curso universitario e identificación de elementos del significado institucional presentado a los alumnos. Esta fase se expone en el Capítulo 3.

Segunda fase

Construcción y validación de instrumentos de evaluación, con el fin de caracterizar los errores y concepciones de los alumnos sobre los elementos de significado caracterizados en la fase anterior y el resto de conocimientos que hemos mencionado. Esta fase se expone en el Capítulo 4.

Tercera fase

Toma de datos a partir de los instrumentos de evaluación construidos en la fase anterior, análisis de los datos obtenidos y redacción de conclusiones. Esta fase se expone en los capítulos 5, 6 y 7.

Análisis de datos

Los datos obtenidos de todos los instrumentos anteriores se han sometido a un proceso de análisis de contenido (Bardin, 1.986; Krippendorff, 1.990; Fernández y Rico, 1.992) para su análisis cualitativo y para su tratamiento estadístico se ha empleado el paquete SPSS y el BMDP.

Los análisis estadísticos han sido los siguientes:

a) Porcentajes de respuestas correctas, parcialmente correctas e incorrectas en los items de comprensión conceptual y análisis de diferencias por items.

b) Análisis de varianza de medidas repetidas del error absoluto en la estimación del coeficiente de correlación en función del tipo de tarea e intensidad del coeficiente de correlación.

c) Test t de diferencia entre los errores medios en la estimación del coeficiente de correlación en relación lineal / no lineal y en función de si las teorías previas coinciden o no con los datos.

d) Análisis de correspondencias (Nishisato, 1.980; Greenacre, 1.984; Greenacre y Hastie, 1.987; Lacasta y Brousseau, 1.995) de la tabla de contingencia obtenida al cruzar las estrategias en la estimación del coeficiente de correlación con las tareas, usando como variables suplementarias las variables independientes del cuestionario.

e) Para analizar la solución del problema, en sus distintos apartados, se ha realizado una primera aproximación, codificando cada una de las respuestas emitidas por los alumnos de la muestra. A continuación, se han agrupado las respuestas por su similitud en grupos, formando categorías de respuestas. Con estas categorías se ha realizado el análisis cualitativo (Miles y Huberman, 1.984; Huberman y Miles, 1.994). Finalmente se han examinado los porcentajes en cada categoría de respuestas.

Capítulo 2

Antecedentes

2.1. INTRODUCCIÓN

En este capítulo analizamos los antecedentes investigadores de nuestro trabajo, la mayor parte de los cuales se han llevado a cabo dentro del campo de la psicología. La habilidad para detectar la relación entre sucesos del entorno, cuando ésta existe, y el usar este conocimiento para hacer predicciones ha sido considerado como una componente básica de la inteligencia humana (Well, Boyce, Morris, Shinjo y Chumbley, 1.988), siendo la correcta valoración de la intensidad de la misma muy importante para el funcionamiento cognitivo y la conducta adaptativa de los individuos (Kareev, 1.995). El razonamiento correlacional es una componente esencial del razonamiento en sentido amplio, particularmente en las interacciones sociales en las cuales el ser capaz de aceptar las relaciones válidas y rechazar las que no lo son es una destreza fundamental (Ross y Smyth, 1.995). Ello es debido a la importancia que los juicios de asociación tienen para la toma de decisiones y la repercusión de ésta actividad en diversas disciplinas y ocupaciones del mundo actual, como economía, política o sociología.

Para realizar este estado de la cuestión hemos partido del trabajo de Estepa (1.994), completando su estudio con el resumen de las investigaciones que sobre este tema se han desarrollado en el período 1.994-1.998.

2.2. LOS JUICIOS DE ASOCIACIÓN COMO COMPONENTES DEL RAZONAMIENTO CAUSAL

El estudio de las relaciones que pueden existir entre dos sucesos es necesario para el progreso del conocimiento humano, ya que, como indica Crocker (1.981, pág. 272), *"conocer si los sucesos se relacionan y, con qué intensidad lo hacen, facilita a las personas explicar el pasado, controlar el presente y predecir el futuro".*

Pozo (1.987) considera que las investigaciones sobre asociación estadística pueden incluirse en un campo de investigación más amplio que versaría sobre el razonamiento causal, acerca del cual describe dos teorías principales. La primera de ellas es debida a Piaget y García (1.973), quienes consideran que el desarrollo en el sujeto del razonamiento causal sólo se produce cuando éste ha alcanzado la etapa de las operaciones formales. Para estos autores, la idea de causalidad se asimila a la vez a una transformación y a una conservación. La causalidad se ve como transformación, porque la causa produce un efecto que es diferente a la causa en sí misma. Por otro lado hay una conservación de la relación entre causa y efecto, relación que captamos por vía inferencial.

La idea de causalidad emerge a través de la interacción del sujeto y el objeto, que, de acuerdo con los principios piagetianos, se rige por los procesos de asimilación y acomodación. Lo específico del razonamiento causal es que la operación no sólo se aplica a los objetos, pues se considera que estos actúan unos sobre otros, la causa actúa sobre el efecto.

Otra teoría psicológica sobre razonamiento causal analizada por Pozo (1.987) es la de Kelley (1.973), el cual supone que para decidir entre las causas posibles cual es la efectiva, el sujeto actuaría empleando una versión ingenua del método usado en la ciencia, teniendo en cuenta tres dimensiones: Consistencia (la

causa es seguida por el efecto), discriminabilidad (otras causas no son seguidas por el efecto) y consenso (en otros casos se produce la misma causa con los mismos efectos).

Según Kelley la persona adulta ha adquirido un repertorio de ideas sobre la operación e interacción de factores causales. Cuando el sujeto carece de información sobre esa situación causal, recurre a esas concepciones o esquemas causales. Un ejemplo sería el esquema de causas múltiples suficientes, por el que en presencia de varias causas potenciales, el sujeto atribuye menor peso al efecto de cada una de las causas individuales (Pozo, 1.987).

Esta teoría diferencia dos problemas causales distintos, según la cantidad de información disponible para el sujeto: Si el sujeto tiene información que procede de observaciones múltiples se plantea un problema de covariación, entendida ésta como que los valores de dos variables tienden a cambiar de forma conjunta (Kareev, 1.995). Si el sujeto dispone de información de una sola observación se trata de un problema de configuración de causa. Las investigaciones sobre asociación estadística se encuadran en el primero de estos tipos de problemas.

Por su parte, Pozo (1.987) cree que el estudio del pensamiento causal debe contemplar las interacciones entre los aspectos representacionales de la causalidad (teorías causales) y los aspectos procesuales (reglas de inferencia). Una regla de inferencia ampliamente estudiada ha sido la detección de la covariación por parte de los sujetos. Para clasificar éstas diferenciaremos la situación experimental en función de las relaciones mantenidas entre causa y efecto. La relación de causalidad necesaria y suficiente es aquella en que la causa va seguida siempre del efecto sin que éste pueda producirse en su ausencia. La relación de causalidad suficiente es aquella en que la causa va siempre seguida del efecto, pero éste puede producirse en ausencia de la causa. La causalidad necesaria se produce cuando el efecto no puede producirse en ausencia de la causa, pero puede no producirse en su presencia y, por último, la causalidad contribuyente (ni necesaria ni suficiente) es aquélla en que la causa no va seguida siempre del efecto, que a su vez puede producirse en ausencia de la causa.

Podemos también diferenciar entre covariación simple y múltiple. La covariación simple se refiere a la causalidad necesaria y suficiente. La covariación

múltiple, donde se sitúa la investigación sobre la detección de la asociación estadística, a los otros tres tipos.

El razonamiento correlacional (Ross y Smyth, 1.995) es un campo de investigación que ha permitido analizar conjuntamente dos puntos de vista diferentes. Por un lado se han estudiado las estrategias utilizadas en los procesos de resolución de problemas, comparándolas con las que seguiría un individuo que aplicase procedimientos matemáticos al resolver estos problemas. Otro grupo de investigadores se interesa por la influencia del contexto y de las teorías previas sobre las soluciones de estos problemas. Los resultados son dispares, desde los de autores que consideran que la actuación de los adultos se ajusta a las normas estadísticas, por ejemplo Shaklee (1.979), a los que, como Nisbett y Ross (1.980), sostienen que el razonamiento sobre la asociación estadística es, en general, muy pobre. Posiblemente, sea aún necesaria mucha investigación sobre el tema para llegar a un acuerdo.

Para Crocker (1.981) el proceso para realizar un juicio de covariación entre dos variables consta de los seis pasos siguientes, en cada uno de los cuales pueden producirse errores que ocasionen un juicio inadecuado:

a) Decidir las variables que afectan a la relación de interés y, en consecuencia, los datos que se deben recoger.

b) Diseñar el proceso de selección de una muestra representativa de la población.

c) Clasificar los datos en función de la categorización de las variables que se ha establecido.

d) Resumir la información que se ha recogido, usando, por ejemplo, una tabla de contingencia y, a partir de ella, identificar las frecuencias de casos que confirman y contradicen la hipótesis.

e) Integrar los datos, produciendo el juicio de asociación

f) Usar el juicio obtenido como base de predicción o toma de decisiones

A continuación, resumimos, brevemente, las investigaciones sobre los puntos d) a f) que son los relacionados con nuestro trabajo. Finalizamos con el estudio específico de las investigaciones sobre correlación y regresión dentro de la educación matemática.

2.3. TRABAJO INICIAL DE INHELDER Y PIAGET

Inhelder y Piaget (1.955) son los primeros investigadores que se interesan por la formación de la idea de asociación y la detección de la misma. Estos autores consideran que esta idea es el último paso para comprender el concepto de probabilidad. Por tanto, para estos investigadores la comprensión de la idea de asociación implicaría la adquisición previa del razonamiento proporcional, probabilístico y combinatorio, que no se alcanzan hasta el período de las operaciones formales.

Inhelder y Piaget trabajaron sólo con niños a partir de 11 ó 12 años, a quienes preguntaron si, en una muestra dada, existía asociación entre el color de los ojos y el color del cabello. El dispositivo experimental usado consistía en un conjunto de tarjetas con dibujos de rostros con ojos y cabellos coloreados, a partir de las cuales debían establecer su juicio, usando los datos disponibles.

Este experimento es un ejemplo del problema más simple posible de asociación, que se presenta al clasificar una población o muestra respecto a dos variables estadísticas, cada una de las cuales presentan sólo dos valores posibles (presencia o ausencia de un atributo), como se muestra en la siguiente tabla:

	B	**no B**
A	a	b
no A	c	d

Para estudiar la existencia de asociación entre las dos variables se debe tener en cuenta los casos favorables ($a + d$), los casos desfavorables ($b + c$) y los casos posibles ($a + d$) + ($b + c$). Inhelder y Piaget observaron que algunos niños

sólo consideraban el caso a (positivo-positivo), sin tener en cuenta el caso d (negativo-negativo), que equivale, normativamente, al anterior; otras veces, relacionaban a con b o con c, sin percatarse de d. Según estos autores, la asociación se deduce a partir de la siguiente fórmula:

$$R = \frac{(a+d)-(b+c)}{(a+d)+(b+c)}$$

El sujeto después de comparar la probabilidad $\frac{a+d}{(a+d)+(b+c)}$ y la probabilidad $\frac{b+c}{(a+d)+(b+c)}$, debe comprender que la asociación es una función de (a + d) - (b + c) en relación con el todo. Estudios posteriores han mostrado que esta fórmula sólo es correcta cuando las frecuencias marginales son iguales (Shaklee y Tucker, 1.980).

En los trabajos de Inhelder y Piaget, los niños que se encontraban en el nivel de razonamiento III A (subestadios III A de 11-12 a 14-15 años) pueden estimar probabilidades simples, al calcular los casos favorables y los casos posibles, pero sólo comparan las celdas dos a dos, por lo que pueden comparar los casos a con los casos b o con los casos $a + b$.

Estos sujetos no consiguen relacionar los casos favorables (a + d) y los desfavorables (b + c) a la existencia de asociación con todos los casos posibles. Aunque comprendan que hay relación entre los ojos azules y el cabello rubio, y entre el cabello negro y los ojos negros, no llegan a comprender que a y d tienen la misma significación, ya que oponen los casos a a los b y los c a los d. Una segunda dificultad es que, una vez admitido que los casos a y d son favorables a la asociación y los casos b y c desfavorables a la misma, no los comparan entre sí o con el todo, sino que siguen comparando a con b y c con d.

Al finalizar el subestadio III A hay una etapa intermedia entre éste y el III B. El sujeto llega, de forma paulatina, a establecer las relaciones diagonales y a considerar (a + d) y (b + c) y a compararlas entre ellas o con (a + b + c + d), lo que nos indica el comienzo de la idea de asociación.

En el sujeto del nivel III B (subestadios III B a partir de los 14-15 años) la relación es espontánea entre los casos favorables a la asociación (a + d) y los casos desfavorables (b + c) con el conjunto de los casos posibles. Primero, se

comprende la reciprocidad entre los casos a y d, después, comprende la de los casos b y c. El nivel III B se diferencia del intermedio entre III A y III B en que, mientras en esta etapa intermedia se buscan las relaciones, en el nivel III B el sujeto relaciona de modo directo ($a + d$) con ($b + c$) para juzgar el grado de correlación existente, llegando al descubrimiento de los tres casos de correlación: positiva, nula y negativa.

2.4. ESTRATEGIAS EN LOS JUICIOS DE ASOCIACIÓN

Numerosos psicólogos se interesaron por el estudio de los juicios de asociación, continuando el trabajo pionero de Piaget e Inhelder. Resúmenes de estas investigaciones se encuentran en Crocker (1.981), Beyth-Marom (1.982) y Pérez Echeverría (1.990).

Un punto importante es el análisis de las estrategias, es decir, la forma de resolver los problemas de asociación, donde encontramos un amplio abanico de investigaciones. Distinguiremos tres tipos de problemas, según las variables estadísticas implicadas:

a) Juicios de asociación en tablas de contingencia. Forman el núcleo más abundante de investigaciones. Se trata de analizar la asociación entre dos variables cualitativas y el sujeto debe usar las frecuencias en la tabla de contingencia.

b) Juicio de asociación en diagramas de dispersión. Se trata de analizar la asociación entre dos variables numéricas donde la forma o dispersión del gráfico pueden ser útiles para evaluar la relación entre las variables.

c) Juicio de asociación en la comparación de dos muestras. Se trata de analizar la relación entre una variable numérica y otra cualitativa.

2.4.1. JUICIOS DE ASOCIACIÓN EN TABLAS DE CONTINGENCIA

Las aportaciones de la Psicología a la investigación sobre cómo las personas realizan los juicios de asociación entre variables es abundante en lo relativo a las tablas de contingencia 2 x 2 y bastante más exiguo en las demás formas de relacionar dos variables (Estepa, 1.995b). Para Arkes y Harkness (1.983) la diferencia en los resultados obtenidos en estas investigaciones se debe a la dificultad metodológica. Si se quiere conocer si los sujetos basan sus juicios en la casilla *a*, el investigador cambia el valor de esta casilla con lo que ha cambiado la relación de contingencia, pero si mantiene la relación de contingencia constante y cambia la casilla *a* también tiene que alterar el resto de casillas.

Un primer grupo de trabajos estudia las estrategias empleadas en la resolución de estos problemas por adultos (Smedlund, 1.963; Jenkins y Ward, 1.965; Abramson y Alloy, 1.980; Peterson, 1.980; Evans, 1.982; Shaklee, 1.983; Wasserman y cols, 1.983; Alloy y Tabachnik, 1.984; Wasserman y Shaklee, 1.984; Arkes y Rothbart, 1.985; Chatlosh y cols, 1.985; Dickinson y cols, 1.984; Vázquez, 1.987).

Pérez Echeverría (1.990) distingue siete tipos de estrategias hallados en estos trabajos, que describe como estrategias [a], [a-b], [a-c], [a, b, c], [a-b]/[c-d], [a+d]/[b+c] y [cualquier método de relación multiplicativa entre las cuatro casillas].

La estrategia [a] consiste en utilizar solamente la casilla (presente-presente) donde las dos variables contempladas coocurren. La relación será positiva, negativa o nula si el valor de a es mayor, menor o igual que el de las otras tres casillas. Inhelder y Piaget (1.955) ven esta estrategia como precursora de las operaciones formales, aunque algunos investigadores la han encontrado en sujetos adultos (Smedlund, 1.963; Shaklee y Tucker, 1.980; Shaklee y Mims, 1.982; Yates y Curley, 1.986).

La estrategia [a-b] consiste en comparar la diferencia de frecuencias absolutas entre la casilla *a* (presente-presente) y la casilla *b* (presente-ausente). Ha sido descrita por Inhelder y Piaget (1.955), Smedlund (1.963), Adi y cols. (1.978) y Shaklee y Mims (1.982). En uno de sus experimentos Arkes y Harkness (1.983)

encontraron que era la estrategia más utilizada. La estrategia [a-c] es similar a la anterior, pero usando la diferencia de las frecuencias absolutas de la casilla a con la c (ausente-presente).

La estrategia [a, b, c], en términos de un juicio causal, determina la necesidad y suficiencia de las causas a partir de las diferencias encontradas sin realizar cálculo matemático, solamente se infiere de manera subjetiva, a partir de las diferencias encontradas (Pérez Echeverría, 1.990).

La estrategia [a-b]/[c-d] se utiliza para calcular la razón entre las diferencias de las frecuencias de ambas filas (Pérez Echeverría, 1.990).

La estrategia [a+d]/[b+c] se utiliza para calcular la razón entre los casos que confirman la posible relación y los que la falsan. Una variedad de esta estrategia fue estudiada por Shaklee y Tucker (1.980), Shaklee y Mims (1.982), Arkes y Harkness (1.983) y Allan y Jenkins (1.983), que consiste en hallar la diferencia de la suma de las diagonales y se le suele llamar en la literatura δD = [(a+d)-(b+c)], Jenkins y Ward (1.965) observan la limitación de esta estrategia a los casos en que las frecuencias marginales son iguales, ya que en el caso de que sean desiguales se puede detectar asociación cuando existe independencia. Allan y Jenkins (1.983) encuentran que los juicios de asociación tienden a hacerse de acuerdo con la regla δD. Fue la estrategia más utilizada en el trabajo de Shaklee y Tucker (1.980).

En cuanto al último tipo de estrategia mencionado: [cualquier método de relación multiplicativa entre las cuatro casillas], se consideró cuando los sujetos relacionaban las frecuencias absolutas de las cuatro casillas utilizando la multiplicación. Una de tales estrategias es la llamada en la literatura

$$\delta P = \frac{a}{a+b} - \frac{c}{c+d}$$

propuesta por Jenkins y Ward (1.965) y estudiada entre otros por Shaklee y Mims (1.982) y que consiste en calcular la diferencia de las frecuencias relativas de las primeras casillas de las dos filas, o bien, en términos probabilístico, comparar las probabilidades condicionadas de un suceso, dados los valores alternativos de otro suceso. Según este criterio las variables serían independientes si δP = 0, siendo la asociación positiva o negativa según el signo de δP.

Pérez Echeverría (1.990) agrupa estas siete estrategias en cinco niveles según el número de casillas empleadas y el modo en que se integra la información:

Nivel 1, estrategia [a] (3'7 por ciento de uso)

Nivel 2, estrategias [a-b] y [a-c] (30'56 por ciento de uso)

Nivel 3, estrategia [a, b, c] (15'74 por ciento de uso)

Nivel 4, estrategias [a-b]/[c-d] y [a+d]/[b+c] (37'96 por ciento de uso)

Nivel 5, estrategia [cualquier método de relación multiplicativa entre las cuatro casillas] (12'04 por ciento de uso)

En su investigación con alumnos de C.O.U. Estepa (1.994) encuentra las mismas estrategias que hemos descrito para las tablas de contingencia, ampliando, además, los niveles de Pérez Echeverría (1.990) al caso de tablas de contingencia con más de dos filas o columnas (Estepa, 1.994; Batanero, Estepa, Godino y Green, 1.996). También, realiza una clasificación de las estrategias desde el punto de vista matemático, teniendo en cuenta, asimismo, los teoremas y concepto en acto implícitos en las mismas (Estepa y cols., 1.994; Batanero, Estepa, Godino y Green, 1.996). Entre las estrategias correctas encontró las siguientes:

S1: Comparación de todas las frecuencias relativas de las distribuciones condicionales $h(B_j / A_i)$ *de cada valor* B_j *para dos o más valores diferentes de* A_i. Los estudiantes que usaron esta estrategia, implícitamente se basaban en el hecho de que la dependencia de una variable B sobre otra variable A implica que la frecuencia relativa condicional $h(B_j / A_i)$ varía cuando A varía. Los papeles de A y B son intercambiables.

S2: Comparación de una frecuencia relativa condicional $h(B_j / A_i)$ *para un valor fijo de* B_j *para cada posible valor de* A_i *con la frecuencia relativa marginal* h_j. Los estudiantes que siguen esta estrategia implícitamente usan la propiedad de invarianza de la distribución de B cuando se condiciona por los valores de A.

S3: Comparación de casos a favor y en contra de B para cada uno de los valores de A. Intuitivamente se está empleando la razón de posibilidades que es una medida de asociación para variables dicotómicas.

Entre las estrategias incorrectas las más importantes son las que se basan en frecuencias absolutas o sólo emplean una parte de los datos en la tabla de contingencia. Algunas de estas estrategias son las siguientes:

S4: Comparación de dos o más distribuciones de frecuencias absolutas condicionales. Esta estrategia es parcialmente correcta porque los alumnos usan toda la información pertinente; sin embargo, la comparación se hace en términos de frecuencias absolutas, en lugar de comparar frecuencias relativas.

S5: Comparación de una frecuencia absoluta condicional con la frecuencia marginal correspondiente.

S6: Comparar la suma de las diagonales en la tabla. El sujeto reconoce los casos favorables y desfavorables a la asociación, pero no llega a relacionarlos con el total de casos.

S7: Usar sólo una distribución condicional; usar sólo una celda. Sólo se tiene en cuenta parte de la información disponible.

2.4.2. JUICIOS DE ASOCIACIÓN EN DIAGRAMAS DE DISPERSIÓN

Aunque algunos autores estudian la exactitud de los juicios de asociación a partir de diagramas de dispersión, por ejemplo Erlick y Mills (1.967) y Lane y cols. (1.985), únicamente en la investigación de Estepa, hemos encontrado un análisis de las estrategias seguidas por los estudiantes.

En esta investigación, se utilizaron items en los que el juicio de asociación debía realizarse a partir de diagramas de dispersión y en comparación de muestras relacionadas o independientes. Para cada uno de estos tipos de items, se

identificaron una serie de estrategias que no habían sido descritas en trabajos anteriores.

Respecto a las nubes de puntos las principales estrategias correctas identificadas fueron las siguientes (Estepa y Batanero, 1.994; 1.995; 1.996):

S1: Basarse en la forma creciente, decreciente o constante del diagrama de dispersión. El alumno, intuitivamente, usa la idea de que la dependencia es positiva si al aumentar una variable la otra aumenta, es negativa si al disminuir una variable la otra aumenta y hay independencia si no se observa tendencia de aumento o disminución.

S2: Comparar globalmente la dispersión de la nube, para evaluar la asociación, usando la idea intuitiva de que a mayor dispersión la asociación es menor.

Entre las estrategias parcialmente correctas o incorrectas se encontraron las siguientes:

S3: Compara la nube de puntos con un patrón dado; por ejemplo con una línea recta. Esta estrategia es parcialmente correcta porque podría existir relación de tipo no lineal entre las variables.

S4: Deducir la asociación sólo a partir de puntos aislados

S5: Esperar una relación de tipo determinista o interpretar la existencia de otras variables como falta de asociación

2.4.3. JUICIOS DE ASOCIACIÓN EN LA COMPARACIÓN DE MUESTRAS

En cuanto a la comparación de muestras (Estepa y Sánchez Cobo, 1.996a) se diferencian dos tipos de problemas: Comparación de muestras independientes (una variable numérica medida en dos muestras diferentes) y comparación de

muestras relacionadas (una variable numérica medida dos veces en la misma muestra).

Entre las estrategias correctas encontradas en este tipo de problema por los autores se hallan:

S1: Comparar las medias o totales de las dos muestras. Los estudiantes, implícitamente, se basan en la idea correcta de que si hay diferencias en las dos muestras los estadísticos de las mismas deben ser diferentes.

S2: Hallar la diferencia de pares de valores o comparar pares de valores en muestras relacionadas, calculando posteriormente el total o media de las diferencias. Esta estrategia es correcta y coincide con el método general de comparación de muestras relacionadas.

También se encontraron estrategias incorrectas, como, por ejemplo:

S3: Basar la comparación sólo en los máximos y mínimos o en parte de la distribución

S4: Esperar una uniformidad en las diferencias, por ejemplo, siempre creciente o siempre el mismo valor de las diferencias.

Las investigaciones de Estepa y colaboradores muestran, también, la riqueza de conceptos y procedimientos relacionados con el estudio de la asociación, así como la complejidad del significado de este concepto.

2.5. INFLUENCIA DE LAS TEORÍAS PREVIAS

Un punto muy importante en las investigaciones sobre la detección de la asociación es si el sujeto se basa en los datos del problema o se guía por sus creencias sobre el tipo de relación que debe existir entre las variables. Las experiencias y el ambiente cultural donde el sujeto se desenvuelve contribuyen a la

formación de una serie de teorías que utiliza para interpretar los hechos cotidianos. La fuerza de estas creencias depende de la experiencia de las contingencias entre las acciones sobre el entorno y sus resultados. Estas teorías o expectativas previas que tenemos sobre el mundo están presentes cuando un sujeto realiza un juicio de asociación entre variables.

Jennings, Amabile y Ross (1.982) compararon los juicios de correlación que realizaban los sujetos, en situaciones donde existían fuertes teorías previas (relacionar variables tales como número de horas de estudio y éxito en un examen) y donde no existían (relacionar una serie de números y letras). Llegaron a la conclusión de que, cuando los sujetos poseían teorías previas, se sobreestimaba la correlación, mientras que, en caso contrario, era necesaria la existencia de una fuerte correlación entre los datos para detectar la asociación, y, aún así, se subestimaba.

Un concepto relacionado con las teorías previas es el de *"correlación ilusoria"*, que consiste en percibir correlación basándose en nuestras propias expectativas, sin ningún hecho empírico que la sustente (Murphy y Medin, 1.985; Chapman y Chapman, 1.969). Tversky y Kahneman (1.982b) consideran que la correlación ilusoria se puede explicar por el heurístico de accesibilidad, por el cual los casos de clases numerosas se recuerdan mejor y con más seguridad que los casos de clases menos frecuentes; los sucesos probables son más fáciles de imaginar que los improbables y las conexiones asociativas entre los hechos se fortalecen cuando los hechos coocurren con cierta frecuencia (Tversky y Kahneman, 1.982b) .

Wright y Murphy (1.984) suponen que los juicios de correlación dependen de las teorías previas que tengan los sujetos y de los datos presentados. Realizaron varios experimentos con variables continuas y con coeficientes de correlación 0'10, 0'50, 0'90, en diferentes contextos. Concluyen que, en presencia de teorías en concordancia con los datos, los sujetos dan una buena aproximación de la correlación aún cuando ésta sea baja. En cambio, cuando éstas no existen la estimación de la asociación es más dispersa y se aparta más de la realidad objetiva.

Alloy y Tabachnik (1.984) afirman que *"para percibir el grado de covariación entre dos sucesos son relevantes dos fuentes de información: la información sobre la situación de la contingencia objetiva entre los sucesos proporcionada por el entorno y las teorías previas o creencias del sujeto respecto a los sucesos de covariación en cuestión"* (pág. 114). Si la percepción de la covariación coincide con las teorías previas y con la información de la situación, se realizaría una atribución o percepción de la covariación con una confianza extrema. Si, por el contrario, se encuentran en desacuerdo, se estaría en un dilema cognitivo. En este caso la relativa fortaleza de las dos fuentes de información determina la naturaleza y exactitud de la percepción de la covariación.

2.6. EL CONTEXTO Y LA PRESENTACIÓN DE LA INFORMACIÓN

Aunque tiene menor importancia que las teorías previas, hemos encontrado investigaciones que muestran el efecto de las instrucciones dadas a los sujetos antes de comenzar la prueba sobre los resultados obtenidos (Shaklee y Tucker, 1.980). Por este motivo se incluyen en los cuestionarios de la presente investigación instrucciones estandarizadas y han sido siempre administrados por el investigador, quien explicó con claridad a los sujetos lo que se esperaba de su respuesta.

La presentación de la información varía en las distintas investigaciones realizadas, si bien las podemos resumir en dos modos: a) en forma de tabla de doble entrada, como, por ejemplo, Arkes y Rothbart (1.985), Tversky, Sattath y Slovic, (1.988), Pérez Echeverría (1.990), Ortega Martínez (1.991) y Price y Yates (1.995), y, b) de manera secuencial, donde la información se presenta por medio de tarjetas, diapositivas o medios electrónicos (pulsar un botón, juegos de ordenador, etc.), como, verbigracia, Inhelder y Piaget (1.955), Wasserman, Chatlosh y Neunaber, (1.983), Dickinson, Shanks y Eveden (1.984), Shanks (1.989), Chapman y Robbins (1.990), Bolger y Harvey (1.993), Vallée-Tourangeau, Baker y Mercier (1.994), Kareev (1.995), Klinger y Greenwald (1.995) y Anderson y Fincham (1.996). Ward y Jenkins (1.965) comparan sistemáticamente la presentación secuencial de los datos y mediante tablas, y comprueban que los

juicios de asociación son mejores cuando se presentan los datos en forma de tabla de doble entrada. Wasserman y Shaklee (1.984) llegan a la misma conclusión.

La presentación de la información y los contextos utilizados en las tareas propuestas a los sujetos sobre juicios de asociación en las diferentes investigaciones pueden influir en los resultados y conclusiones obtenidas (Arkes y Harkness, 1.983; Shaklee y Mims, 1.982). En este sentido, Trolier y Hamilton (1.986) estudian la influencia de las variables en los juicios sobre correlación. Investigan los efectos de tres tipos de variables sobre la capacidad de los individuos al valorar relaciones de asociación: i) La forma en la cual la información era presentada (por ejemplo, en forma continua o binaria), ii) la intensidad real de la correlación presentada en la información, y, iii) las expectativas de los sujetos respecto a la relación en cuestión. Los resultados indican que las estimaciones de los sujetos sobre la correlación estaban significativamente influídas por estos tres factores. Los juicios de los sujetos eran sensibles a la diferencia entre correlaciones altas y bajas presentes en la información, pero reflejaban que esta diferencia era debida, en gran parte, a que la información era mostrada en forma binaria antes que en forma continua. También, los sujetos hacen estimaciones más altas de la correlación cuando ellos esperan que las variables estén correlacionadas que en caso contrario.

Por último, según Beyth-Marom (1.982), la presentación de la información por medio de variables simétricas o asimétricas y el tipo de instrucción también tienen influencia en los juicios sobre correlación emitidos por los sujetos. Una variable simétrica es aquella en que los valores que puede tomar tienen el mismo peso para el individuo perceptor, por ejemplo la variable sexo (masculino, femenino). Una variable asimétrica es aquella cuyos valores no tienen el mismo peso para el sujeto perceptor, por ejemplo padecer una enfermedad (padece la enfermedad, no la padece) sería su expresión asimétrica. En las variables asimétricas se discrimina entre la ocurrencia y no ocurrencia de los sucesos, entre el caso positivo y el negativo. Beyth-Marom (1.982) encontró una mejor percepción de la correlación cuando se utilizan variables simétricas.

2.7. CONCEPCIONES DE LOS ESTUDIANTES E INFLUENCIA DE LA ENSEÑANZA

Las estrategias utilizadas por los sujetos en los problemas sobre juicios de asociación no es lo único que influye en la solución que obtienen, no siendo infrecuente que aplicando estrategias que son, en general, inadecuadas, alcancen, no obstante, soluciones correctas en problemas específicos. Por tanto, la tasa de aciertos en problemas propuestos en estas investigaciones está condicionada por el problema correlacional presentado (Shaklee y Tucker, 1.980). Por otro lado, la solución depende de la concepción subyacente sobre la asociación.

En los trabajos de Estepa, Batanero y Godino (Estepa, 1.994; Batanero, Estepa y Godino, 1.996; Estepa y Batanero, 1.995, 1.996) se han descrito cuatro concepciones incorrectas diferenciadas sobre la asociación estadística. Estas concepciones, que fueron identificadas al comparar las estrategias y juicios de asociación de los alumnos, mediante el análisis de sus argumentos, son las siguientes:

1. *Concepción determinista de la asociación*. Algunos estudiantes no admiten más de un valor de la variable dependiente para cada valor de la variable independiente. Cuando esto ocurre, consideran que no hay dependencia entre las variables. Es decir, la relación entre las variables debe ser, desde un punto de vista matemático, funcional. Por ejemplo, en un problema planteado a la muestra de alumnos en el que se les da una tabla con la presión sanguínea tomada a 10 mujeres antes y después de un tratamiento médico, un alumno argumenta: *"El tratamiento no tiene mucha influencia en la presión sanguínea, ya que a algunas mujeres les aumenta la presión sanguínea, mientras que a otras les disminuye".*

2. *Concepción unidireccional de la asociación*. Se percibe la dependencia solamente cuando es positiva (asociación directa), considerando la asociación inversa como independencia. El siguiente ejemplo ilustra un caso de asociación inversa interpretada como independencia, se refiere a un problema en el que se da a los alumnos una tabla de contingencia 2 x 2, en la que las frecuencias

presentadas muestran independencia entre las variables fumar (fuma, no fuma) y padecer trastornos bronquiales (padece, no padece): *"Personalmente creo que no hay dependencia, porque si tú miras la tabla hay mayor proporción de personas con trastornos bronquiales entre los no fumadores"*. Esta concepción fue también encontrada, con posterioridad, por Morris (1.997, 1.998).

3. *Concepción local de la asociación*. Utilizan, únicamente, parte de los datos proporcionados por el problema para emitir el juicio de asociación. Si la parte de datos utilizados confirma un tipo de asociación, adoptan ese tipo de asociación en sus respuestas. En el mismo problema del caso anterior: *"Existe dependencia entre fumar y padecer trastornos bronquiales porque si observamos la tabla hay más fumadores con trastornos bronquiales que no fumadores 90 > 60"*.

4. *Concepción causal de la asociación*. Algunos estudiantes sólo consideran la existencia de asociación entre variables si se puede atribuir una relación causal entre ellas. Este tipo de concepción se encontró, particularmente, en un problema en el que se pedía que dos jueces puntuaran a un conjunto de individuos, siendo un ejemplo de ella la respuesta dada por un sujeto que dice: *"Porque un juez no puede influir en el otro. Cada uno tiene sus preferencias no puede haber mucha relación entre las puntuaciones otorgadas por cada uno"*.

Evolución de las concepciones con la enseñanza

Sin duda, el principal trabajo llevado a cabo sobre el diseño de secuencias de enseñanza de la asociación estadística basado en el uso del ordenador y la evaluación de su impacto sobre las concepciones de los alumnos es el llevado a cabo por Batanero, Estepa y Godino durante los cursos 1.990 a 1.998. Los resultados de este proyecto se describen brevemente en Batanero, Godino y Estepa (1.998) y se utilizan como base para la reflexión sobre el papel del ordenador como recurso didáctico y como instrumento en la resolución de problemas en Batanero, Estepa y Godino (1.998).

El punto de partida de los autores es la consideración de que el análisis exploratorio de datos puede ser una actividad apropiada para reforzar en los

estudiantes la comprensión de la asociación (Batanero, Godino y Estepa, 1.988), puesto que la mayor parte de las actividades de exploración de datos giran, de un modo u otro, alrededor de este concepto. Asimismo, *"la exploración de datos es el primer paso para encontrar relaciones y sugerir hipótesis"* (Noda y Espinel, 1.992, pág. 29). La disponibilidad de tecnología adecuada, además, pone a disposición de los alumnos herramientas potentes de cálculo y sistemas múltiples de representación de la asociación entre variables que, a priori, pueden ayudar al alumno a construir o ampliar el significado del concepto, respecto a lo que sería factible sin estos útiles.

El primero de los experimentos de enseñanza se orientó, en general, al aprendizaje del análisis exploratorio de datos (Estepa. 1.990), poniéndose a punto algunos ficheros de datos y actividades basadas en los mismos que recogiesen las principales variables del campo de problemas correspondientes (Godino y cols., 1.991). Una de las primeras consecuencias de esta experiencia fue apreciar las estrategias intuitivas de los alumnos y sus dificultades en relación con la asociación estadística (Godino y cols. , 1.990; Batanero y cols., 1.991; Estepa, 1.995a).

En Estepa (1.994) se describe un segundo experimento de enseñanza, orientado, esta vez, al tema, específico, de la asociación estadística. Participaron en el mismo 20 alumnos y el aprendizaje fue evaluado mediante la comparación de los resultados en dos cuestionarios paralelos (pre-test y post-test) tanto sobre las estrategias empleadas por los alumnos como por la resistencia de sus concepciones iniciales incorrectas (Batanero, Estepa y Godino, 1.996; Batanero y cols. , 1.997; Batanero, Godino y Estepa, 1.998; Batanero, Estepa y Godino, 1.998). Una primera consecuencia es que se produce, en general, una mejora en las estrategias de los alumnos, en particular en las referidas a los juicios de asociación a partir de diagramas de dispersión. Sin embargo, la mejora no es homogénea en los diferentes alumnos y se observa, asimismo, una permanencia en la concepción causal de la asociación, probablemente, porque en la secuencia de enseñanza no se diseñaron actividades específicas destinadas a la superación de esta concepción.

Por otro lado, los autores realizaron la observación de una pareja de alumnos a lo largo del proceso de aprendizaje, a partir del registro de su interacción

con el ordenador, sus conversaciones y debates y los resultados escritos de su trabajo en 7 sesiones de prácticas. A partir de la observación realizada se identificaron una serie de "actos de comprensión" del concepto de asociación, los cuales han sido descritos en los trabajos citados, y que son los siguientes:

1. La comparación de dos o más muestras, con objeto de estudiar la posible relación entre dos variables debe efectuarse en términos de frecuencias relativas. En la primera sesión los alumnos comienzan comparando frecuencias absolutas de la distribución de una variable en dos muestras. Este error es advertido por el profesor al final de la clase, pero se presenta de nuevo en las sesiones 2, 3 y 5. A partir de ahí los estudiantes parecen haberlo superado.

2. La posible existencia o no de diferencias en la distribución de una variable entre dos o más muestras se deduce a partir de la comparación de toda la distribución de la variable en cada una de las muestras y no de una parte de la misma. Los estudiantes, sin embargo, comienzan con la comparación de valores aislados, al estudiar las dos muestras. Por ejemplo, en la primera sesión, los estudiantes solamente comparan los valores de máxima y mínima frecuencia en ambas muestras; aunque estas diferencias apuntan a la existencia de posible asociación, este modo de proceder es insuficiente para cuantificar la intensidad de la misma. Esta dificultad vuelve a aparecer en las sesiones 2, 3 y 5, desapareciendo en las sesiones posteriores.

3. A partir de una misma frecuencia absoluta pueden deducirse dos frecuencias relativas condicionales diferentes, según la variable que se emplee como condición. El papel de condición y condicionado en la frecuencia relativa condicional no es intercambiable. Numerosos autores, como, por ejemplo, Falk (1.986), señalan la dificultad de interpretación de una probabilidad condicional porque los alumnos no diferencian, a veces, el papel jugado por la condición y el condicionado, con lo que pueden confundir $P(A|B)$ con $P(B|A)$ o no llegar a discriminarlas. Muchos alumnos, en este estudio, mostraron confusiones similares a ésta en el estudio de las concepciones previas, las cuales siguieron manifestándose durante el proceso de instrucción y que de forma persistente continúan exhibiéndose al finalizar la

enseñanza. Apareció en la sesión 5 y se superó con la ayuda del profesor. No apareció en el resto de las sesiones.

4. Dos variables son independientes si la distribución condicional de una de ellas no cambia cuando se varían los valores de la otra variable. Hasta llegar a la sesión 5, los estudiantes no descubren que una condición para la independencia es la invarianza de las distribuciones relativas condicionales, cuando varía el valor de la variable condicionante.

5. En la determinación de la asociación entre dos variables, éstas juegan un papel simétrico. Por el contrario, en el estudio de la regresión, las variables desempeñan un papel asimétrico. Hay dos rectas de regresión diferentes, según cual de las dos variables actúe como variable independiente. El hecho de que en la correlación no se distinga entre la variable explicativa y la variable explicada, mientras que en la regresión esta diferencia sea esencial (Moore, 1.995) provocó gran confusión entre los estudiantes. Cuando necesitaron seleccionar la variable explicativa para calcular la línea de regresión en las sesiones 5, 6 y 7, no supieron qué variable elegir. Por ejemplo, para calcular la línea de regresión del peso sobre la altura, los estudiantes se desconcertaron por el hecho de que existía mutua dependencia entre las dos variables, debatieron largamente sin llegar a una solución aceptable para ellos. El profesor no se dio cuenta del problema y finalmente los estudiantes calcularon la línea de regresión eligiendo la variable explicativa al azar. Al final del período de enseñanza estos estudiantes aún no habían descubierto que se pueden calcular dos líneas de regresión diferentes.

6. Una correlación positiva indica dependencia directa entre las variables. Aunque en la sesión 6, los alumnos pudieron interpretar la magnitud del coeficiente de correlación, no discutieron el tipo de asociación (directa o inversa). Al final de la sesión aunque llegan a indicar que "al aumentar una variable la otra aumenta" no identifican este hecho con la idea de relación directa entre las variables. Nunca llegan a emplear la idea de "relación o dependencia directa".

7. Una correlación negativa indica dependencia inversa entre las variables. En la sesión 6, los alumnos se sorprenden al encontrar, por primera vez, un coeficiente de correlación negativo, hasta el punto de preguntar al profesor si ello es posible. Asimismo, aparece la duda en la comparación de dos coeficientes de correlación

negativos, ya que, en este caso, un número menor corresponde a mayor intensidad en la asociación. Así, el conocimiento adquirido sobre el orden de los números negativos dificulta ahora la comprensión del signo negativo del coeficiente de correlación; se convierte en obstáculo epistemológico para dicha comprensión. Consideramos que ello se debe al fenómeno de inversión de la relación de orden (González y cols., 1.990). En realidad, aunque ayudados a veces por el profesor, han observado que el signo negativo del coeficiente de correlación se corresponde con una pendiente negativa en la recta de regresión y que al aumentar los valores de x disminuyen los de y, en el resto de la sesión, no llegan a utilizar el término "dependencia inversa". No alcanzan a diferenciar los dos tipos de asociación al término del aprendizaje

8. El valor absoluto del coeficiente de correlación es indicativo de la intensidad de la asociación. Aunque en las primeras actividades los alumnos asocian un alto valor del coeficiente con una dependencia fuerte, hasta la sesión 6 no identifican, en principio, la idea de "intensidad de la asociación" con el coeficiente.

Dos conclusiones de los trabajos anteriores son las limitaciones del software disponible y que el inconveniente de atender a una diversidad en la formación básica del alumno contribuye, en ocasiones, a la persistencia de las dificultades de comprensión de los alumnos al finalizar la instrucción. Para mostrar esta problemática Batanero y Godino (1.998) presentan los datos recogidos al finalizar un curso básico de análisis de datos, en una prueba que evalúa la capacidad final de los alumnos para la determinación de la asociación entre variables.

En concreto, en el curso citado, de 80 horas de duración, los alumnos trabajaron 2 sesiones semanales a lo largo de un curso en el laboratorio de informática (1 ó 2 alumnos por ordenador) con algunos de los procedimientos del paquete estadístico Statgraphics. Los 32 alumnos, divididos en 2 grupos, disponían, en su mayoría, de conocimientos estadísticos básicos, aunque nunca habían trabajado con un paquete estadístico. El análisis de los autores pone de manifiesto que la actividad de "análisis de datos", incluso a nivel exploratorio, es muy sofisticada y requiere el conocimiento de los conceptos subyacentes en las diferentes representaciones de la asociación (gráficas y numéricas, resumidas o no,

descriptivas e inferenciales). Precisa, también, de la selección óptima de la representación en función de los datos del problema, la flexibilidad en el cambio de sistema de representación -cuando lo requiera el análisis-, la interpretación adecuada de los resultados y su puesta en relación con las preguntas de la investigación. Aunque los alumnos mostraron, en general, buenos resultados, se observa, asimismo, las dificultades en cada uno de los pasos del proceso descrito.

Los autores señalan que la investigación sobre la enseñanza de la estadística es aún muy incipiente y se concentra, preferentemente, en el estudio de las concepciones de los alumnos sobre conceptos elementales. No hay apenas investigación sobre las concepciones de los alumnos respecto a conceptos estadísticos avanzados, probablemente, porque este tipo de conceptos no se han incluído en el currículo de secundaria hasta muy recientemente. Más escasos son, todavía, los estudios de experimentos de enseñanza con o sin ordenadores.

Por otro lado, evaluar este tipo de experimentos es muy laborioso por la cantidad de datos generados y la ausencia de modelos previos para el análisis e integración de los mismos. La misma evaluación del trabajo de los alumnos con el ordenador plantea problemas de investigación específicos, porque es difícil transferir los resultados de la investigación sobre evaluación usando métodos tradicionales. Además, la variedad de parámetros a tener en cuenta en el análisis de un experimento hace que éstos sean difícilmente reproducibles en otros contextos, con otro software u otro tipo de alumnos. Todo ello muestra la necesidad de proseguir la investigación y reforzar las conexiones de ésta con la práctica docente.

Otros trabajos que hemos encontrado que estudien la correlación desde un punto de vista didáctico son los de Morris (1.997, 1.998) y Truran (1.997). Morris (1.997) utiliza una muestra de 20 estudiantes de Psicología a los que propone cumplimentar un cuestionario y una serie de tareas relacionadas con la correlación. Este estudio se compone de cuatro partes:

a) Estudio sobre si los estudiantes encuentran la estadística difícil o es una materia poco útil.

b) Concepciones alternativas de los estudiantes en el tema de la correlación

c) Dificultades que encuentran los estudiantes al llevar a cabo diversos procedimientos estadísticos.

d) Estudio de las tareas más adecuadas para valorar la comprensión de los estudiantes en el tema de la correlación.

Comentaremos los resultados encontrados en los apartados b) y d) que son los que se relacionan con nuestro tema. No todos los estudiantes de la muestra tienen una comprensión clara de la correlación. Algunos estudiantes infieren causalidad a partir de la correlación. Además, encontró que los estudiantes tienen dificultades y confusiones con la correlación positiva, negativa y la ausencia de correlación. Así algunos estudiantes creen que la correlación positiva siempre es más fuerte que la negativa y que una correlación negativa es indicativa de la existencia de no correlación entre las variables. Como se puede observar estos resultados coinciden con los obtenidos en los trabajos de Estepa (1.994), Batanero, Estepa y Godino (1.996) y Estepa y Batanero (1.995, 1.996).

En cuanto al apartado d), este estudio no encuentra un tipo particular de tareas que se puedan usar para valorar la comprensión de los estudiantes del tópico de la correlación, ya que algunas de las tareas utilizadas eran válidas porque evocaban una variedad de concepciones de los estudiantes, no obstante, otras tareas producían un tipo de respuestas de difícil categorización y análisis.

Truran (1.997) evalúa los aprendizajes en sendos cursos impartidos en la universidad de Adelaida (Australia) y en el Sepang Institute of Technology (Malasia), sobre los coeficientes de correlación y determinación, la interpretación de la ecuación de regresión, de la pendiente y punto de corte de la recta de regresión con el eje de ordenadas y de las posibles restricciones en las predicciones. La diferencia entre las respuestas de los alumnos de ambos centros es pequeña. Casi todos los alumnos identifican la correlación moderada y negativa y la cuarta parte de los alumnos adopta la respuesta errónea, relación lineal.

En relación con los resultados sobre la interpretación del coeficiente de determinación, observa un aprendizaje rutinario en las respuestas de los estudiantes malayos y falta de comprensión del concepto en los australianos.

Encuentra la concepción determinista de la asociación estadística, descrita por Batanero et al. (1.996), como él mismo indica. En cuanto a las limitaciones en las predicciones, algunos estudiantes argumentan de manera razonable las reservas que se deben hacer en las extrapolaciones, teniendo en cuenta tanto si la correlación es moderada, como en este caso (r = -0.57), como si tamaño de la muestra es pequeño. Termina el trabajo dejando abierto el interrogante de si el enfoque tradicional por el cual se estudia solamente la correlación lineal en los primeros cursos universitarios es adecuado.

Capítulo 3

Descripción de la enseñanza

3.1. INTRODUCCIÓN

Este capítulo se dedica a la descripción de la enseñanza recibida por los alumnos que forman parte de la muestra, tanto durante su etapa de educación secundaria como en el curso en el que fueron recogidos los datos. Comenzamos presentando un resumen de nuestro estudio previo, que ha sido publicado con mayor detalle en Sánchez Cobo y Estepa (1.996), Estepa y Sánchez Cobo (1.996b), Sánchez Cobo (1.996), Sánchez Cobo y Estepa (1.997a,b, 1.998) y Estepa y Sánchez Cobo (1.998). El análisis de textos de bachillerato es llevado a cabo desde una doble perspectiva: El estudio de la presentación teórica del tema de la correlación y de la regresión y de los ejercicios insertados en este tópico. Este trabajo nos ha servido de base para el posterior análisis de la programación del profesor que imparte la asignatura de Estadística, así como de los apuntes de dos alumnas de un grupo al que da clase el mencionado profesor. Con ello aspiramos alcanzar uno de nuestros objetivos, cual es el de caracterizar el conocimiento de los alumnos de primeros cursos universitarios al finalizar la enseñanza sobre la correlación y la regresión.

Hemos pretendido con ello, además, comparar el análisis de los contenidos incluídos en el estudio descriptivo de la correlación y de la regresión efectuado por Sánchez Cobo (1.996) en bachillerato, con los cursos "típicos" de iniciación a la Estadística en la universidad, con lo cual se intenta identificar los elementos de significado asociados al mismo, en la línea iniciada en las investigaciones de Vallecillos (1.994) y Ortiz (1.996). Ello nos permitirá describir el significado institucional de la asociación estadística en un curso usual de introducción a la Estadística descriptiva de nivel universitario y será la base para la construcción de los instrumentos de evaluación. También, puede servir para el diseño de otros cursos sobre estos contenidos, así como materiales curriculares para la enseñanza.

3.2. LA ENSEÑANZA DE LA ASOCIACIÓN EN BACHILLERATO: ANÁLISIS DE TEXTOS

En los cuestionarios oficiales vigentes cuando estos alumnos cursaban el Bachillerato, el tema de asociación se introducía en el 3º de B.U.P., aunque la enseñanza efectiva del tópico no se llevaba a cabo en todos los centros. En la investigación de Sánchez Cobo (1.996) se analizaron las nociones de la correlación y de la regresión en una muestra de 11 manuales de Matemáticas de 3º de B.U.P., que contemplaba las editoriales más representativas en los institutos de Jaén y provincia, con el fin de caracterizar esta enseñanza. La muestra de libros de texto se eligió en forma "intencional". Por lo tanto, son los investigadores los que, en el momento de seleccionar la muestra, utilizan los criterios que consideran oportunos (Azorín y Sánchez Crespo, 1.986) de modo que ésta sea representativa. A los textos se les ha asignado un código consistente en el año de edición y una letra, los cuales se explicitan en el Anexo II.

El estudio de los libros de texto, que resumimos a continuación, se abordó desde una doble perspectiva:

i) el análisis de la exposición teórica del tema

ii) el análisis de los ejercicios que se incluían en cada tema.

3. 2. 1. ANÁLISIS DE LA EXPOSICIÓN TEÓRICA DEL TEMA

El análisis de la exposición teórica de los contenidos sobre Correlación y Regresión en los textos estudiados se organizó en torno a los siguientes apartados:

I. Objetivos del tema y metodología de la presentación

II. Contenidos incluídos y su organización

III. Presentación de las distribuciones dobles

IV. Estudio de la correlación

V. Estudio de la regresión

I. OBJETIVOS Y METODOLOGÍA

Objetivos

Los objetivos juegan un papel significativo en la enseñanza, pues, como destacan Araújo y Chadwick (1.988), *"si no se sabe con certeza adónde se va, se puede llegar a cualquier parte"* (pág. 95), siendo *"una guía para el desarrollo del material de aprendizaje"* (Araújo y Chadwick, 1.988, pág. 95). Como manifiesta Orton (1.990), la importancia de los objetivos educativos es enfatizada por los siguientes tres argumentos:

"a) proporcionan al profesor unas orientaciones para el desarrollo de materiales de instrucción y del método docente;

b) permiten al profesor concebir medios de valorar si se ha realizado lo que se pretendía; y

c) proporcionan una dirección a los alumnos y les ayudan a realizar esfuerzos mejores para el logro de sus metas" (pág. 59).

En consecuencia, el conocimiento de los objetivos a conseguir por parte del alumno es elemento favorecedor de sus aprendizajes, puesto que el estudiante podrá enfocar su trabajo y dosificar su esfuerzo según la dirección que éstos determinen. En nuestro caso, sólo 2 de los 11 libros estudiados contienen los objetivos al principio del desarrollo del tema.

Metodología de la presentación del tema

En primer lugar, para estudiar el modo de presentación del tema, se analizaron los libros que comienzan con ejemplos y los que empiezan por teoría, resultando que de los 11 libros estudiados 2 inician el tema con exposiciones teóricas (77C y 77D) y 9 con ejemplos. Esto nos lleva a considerar que en el primer caso se tiene una concepción más formalista de las matemáticas y en el segundo más constructivista. De los manuales que inician la presentación del tema con uno o varios ejemplos, hemos detectado que luego solamente continúan con esa tendencia, aproximadamente, la mitad de los mismos. En concreto, únicamente los libros de texto (78A), (81A), (86A), (87A) y (88A) emplean, de forma generalizada, los ejemplos como vía preparatoria al desarrollo del epígrafe. Los restantes textos utilizan este procedimiento sólo muy esporádicamente a lo largo de la exposición.

Número de ejemplos en el desarrollo del tema y de ejercicios al final del mismo

Otro aspecto que hemos analizado, en relación a la metodología, ha sido el número de ejemplos que el autor del libro expone en el desarrollo del tema y el número de ejercicios que se proponen. Los ejercicios aparecen al final del tema, a excepción de dos casos en los que aparecen incluídos dentro de la exposición del tema, aunque también hay algunos al final. Algunas veces se da un enunciado y a continuación se pide realizar varias tareas, como pueden ser: dibujar la nube de puntos, hallar la recta de regresión, el coeficiente de correlación, etc. El número de ejercicios varia de 4 a 67, siendo su media de 24.8.

Referencias históricas

A menudo los conocimientos matemáticos se presentan de una manera totalmente alejada de los problemas y motivaciones que dieron lugar a su

nacimiento y desarrollo. Se oculta todo el proceso de construcción, las motivaciones, las intuiciones personales y conjeturas que llevaron a su autor a estudiarlos, los caminos seguidos que se revelaron poco apropiados. Es decir, los saberes matemáticos se presentan destemporalizados, descontextualizados y despersonalizados. Sin embargo, en la enseñanza es preciso buscar nuevos contextos problemáticos a partir de los cuales el alumno construya el saber puesto en juego. El papel del enseñante es, en cierta medida, inverso al del productor del saber matemático, debe por tanto realizar una recontextualización y una repersonalización del saber (Brousseau, 1.986).

Didácticamente tienen gran interés las referencias históricas para la enseñanza de los conceptos, porque sitúan a éstos en el contexto donde nacieron, lo que proporciona una mejor comprensión de los mismos, además de incentivar la motivación. A pesar de que *"cada vez es mayor el interés por emplear la motivación histórica en las clases de matemáticas, aún no se ve esto reflejado en el trabajo diario en el aula"* (Giménez, 1.986, pág. 57). De los libros estudiados 6 no presentan referencias históricas y 5 hacen referencia sucinta a los trabajos de Galton. Estas referencias históricas que hemos encontrado son, en el sentido de Boero (1.989), "notas históricas" que se incorporan para situar un contenido nuevo al abordar su estudio.

II. PRESENTACIÓN DE LOS CONTENIDOS

Contenidos matemáticos expuestos en el tópico

Como puede observarse en la Tabla 3.2.1.1, existe una cierta homogeneidad en los contenidos que se presentan en el apartado correspondiente a la Correlación, aunque haya también otros que muestran un tratamiento muy variable según que texto. Uno de estos contenidos es el de *Dependencia funcional y aleatoria* e *Independencia aleatoria*. A pesar de que es un concepto básico y de que se encuentra en su hábitat natural, solo tres libros de texto lo incluyen - (77A),

(81A), (90A) -. Ello dificulta la posibilidad de una integración conveniente de la correlación, a la vez que se prima a la regresión.

De todos los manuales analizados sólamente uno - (77C) - aborda otro aspecto fundamental sobre la correlación la *covariación*, detallando los tipos de ella que hay, que coinciden con los descritos por Barbancho (1.973). Es importante este aspecto, pues conecta y distingue la dependencia y la causalidad, nociones frecuentemente identificadas por los alumnos como equivalentes, como muestra las investigaciones de Estepa (1.994). Asimismo, los diversos tipos de covariación permiten una mejor comprensión de la dependencia aleatoria, evidenciando una riqueza de situaciones que escapan a la dependencia estrictamente funcional.

Tabla 3.2.1.1. Contenidos sobre correlación incluídos en los libros

		77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
Dependencia	Funcional	X					X					X
	Aleatoria	X					X					X
Independencia aleatoria		X					X					X
Diagrama de dispersión			X	X	X	X	X	X	X	X	X	X
Covariación				X								
Correlación: concepto			X	X		X	X	X	X	X	X	X
Correlación (tipos)	Directa		X	X			X	X	X	X	X	
	Inversa		X	X			X	X	X	X	X	
	Independencia		X	X				X	X	X		
Correlación (medida)	Covarianza	X	X			X	X	X		X	X	X
	Coef.correlación	X	X	X	X	X	X	X	X	X	X	X
	Otras	X	X			X						
Propiedades coef. correlación			X	X	X	X	X	X	X	X	X	X

Los distintos tipos de correlación suelen mostrarse en la mayoría de los manuales analizados. A pesar de la importancia de distinguir la correlación directa e inversa, hay cuatro textos - (77A), (77D), (78A) y (90A) - que no la incluyen dentro de los contenidos desarrollados en la unidad correspondiente. En más de la mitad

de los textos al estudiar la correlación no se presentan situaciones en las que haya independencia de las variables, con lo cual se podría inducir la creencia de que únicamente existen dos posibilidades: correlación directa e inversa.

La mayoría de los manuales emplean dos parámetros para determinar la correlación que hay entre las dos variables: la covarianza y el coeficiente de correlación. Los libros de texto (77A), (77B) y (78A) mencionan otras medidas de la correlación. En el texto (77A) se expone que: *"Las dos rectas de regresión no suelen coincidir. El ángulo que forman nos proporciona una información de gran interés respecto a la dependencia o relación entre las dos variables"* (pág. 308).

En cuanto a los contenidos incluídos en el apartado sobre la Regresión, esencialmente se conforman alrededor de la determinación de las rectas de regresión, de Y sobre X o de X sobre Y, encontrándose en la práctica totalidad de los libros de texto estudiados. Las otras nociones examinadas, como puede observarse viendo la Tabla 3.2.1.2, se presentan en menos de la mitad de dichos textos. Así sucede con *Ajuste lineal, Regresión: concepto, Derivadas parciales* y *Centro de gravedad*. Un caso a subrayar es el del *Coeficiente de regresión* que únicamente aparece en los textos (77B), (78A) y (81A).

Tabla 3.2.1.2. Contenidos sobre regresión incluídos en los libros

		77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
Ajuste lineal					X	X	X	X				X
Regresión: concepto				X	X		X				X	X
Derivadas parciales					X	X		X				X
Rectas de regresión	y sobre x	X	X	X	X	X	X	X	X	X	X	X
	x sobre y	X	X		X	X	X	X	X	X	X	X
Coeficiente de regresión			X			X	X					
Centro de gravedad		X			X	X	X	X	X			

Organización general del tema

El primer aspecto que hemos tenido en cuenta sobre la organización del tópico es la secuenciación puesta en juego a la hora de presentar los conceptos de la correlación y de la regresión, habiendo encontrado en los manuales cuatro posibilidades: i) Tratamiento de forma casi única y/o muy destacada de la regresión; ii) Tratamiento de forma casi única y/o muy destacada de la correlación; iii) Tratamiento de la regresión y de la correlación - en este orden -, y, iv) Tratamiento de la correlación y de la regresión - en este orden -.

En el presente epígrafe, nos vamos a interesar en profundizar en el análisis de otros aspectos básicos de la organización del tema: la definición de los conceptos, los ejemplos y las demostraciones. Un libro de texto de matemáticas recoge, dentro de si, una gran cantidad de conceptos, que comunica, en gran medida, a través de sus definiciones correspondientes. Podemos, en primer lugar, preguntarnos por la utilidad de las definiciones (Skemp, 1.980), pudiendo subrayarse las dos siguientes:

- Nos dice dónde comienza y dónde termina un concepto
- Nos capacita para relacionarlo con otros conceptos

Es decir, *"las definiciones pueden verse como una vía de añadir precisión a las fronteras de un concepto, una vez formado; y establecer explícitamente su relación con otros conceptos"* (Skemp, 1.980, pág. 30). En relación con la comprensión de los conceptos, el propio Skemp ha distinguido entre *"comprensión relacional"* y *"comprensión instrumental"*, considerándose la comprensión relacional como la que se produce cuando se integran y relacionan los conceptos con contenidos matemáticos más generales, mientras que cuando el énfasis se pone en la memorización de rutinas de las que desconocemos su por qué estaríamos en una comprensión instrumental (Cockcroft, 1.985). Nosotros hemos establecido, basándonos en las anteriores ideas expuestas, una taxonomía de definiciones:

- *Definición relacional*: aquélla que pretende mostrar las conexiones del concepto que se está definiendo con otros conceptos con los que está relacionado. Por ejemplo, en el texto (81A) se presenta la definición: *"Llamaremos correlación a la teoría que trata de estudiar la* dependencia *que existe entre las dos variables que intervienen en una distribución bidimensional"* (pág. 441).

- *Definición instrumental*: aquélla que muestra únicamente cómo determinar procedimentalmente el concepto que se está definiendo. Desde un punto de vista matemático sería una definición de tipo constructivo. Así, por ejemplo, en el libro (77C) se presenta la definición: *"En el caso de dos variables X, Y y de covariación lineal, se define el llamado coeficiente de correlación lineal, mediante la expresión*

$$r = \frac{S_{xy}}{S_x S_y}$$ " (pág. 303).

- *Definición instrumento-relacional*: Serían aquéllas que incluyen tanto las relaciones con otros conceptos, como los algoritmos para su determinación. Por ejemplo, en el libro de texto (88A) se presenta la definición: *"Para medir la correlación existente entre dos variables definimos el coeficiente r de correlación lineal:*

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$ " (pág. 164).

En la Tabla 3.2.1.3 se recoge el examen correspondiente de los libros de texto respecto a los distintos tipos de definiciones, referidos exclusivamente a la correlación y a la regresión.

Tabla 3.2.1.3. Frecuencia con que los diversos tipos de definición aparecen en los libros

	77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A	Total
Defínción relacional		2	1	5		5	3	2	2	2	4	26
Definición instrumental	2	3	1		1	1	1	2	2		2	15
D.instrumento-relacional							1		1	1		3

Como puede observarse existe un notable deslizamiento hacia las definiciones relacionales - casi el doble - frente a las instrumentales, siendo las primeras utilizadas para explicar las nociones de correlación y de regresión, no usándose nunca para este caso las de tipo instrumental. Igualmente, las definiciones instrumentales se emplean mayoritariamente para explicar los coeficientes de correlación y de regresión y la covarianza, lo que por otra parte parece bastante natural al tratarse de parámetros estadísticos. Sin embargo, consideramos conveniente que también se diera una definición relacional, es decir, que se explicitaran de forma instrumento-relacional. Las definiciones instrumento-relacionales son altamente escasas, y en algunos casos los autores aunque las utilicen parecen querer subrayar alguno de estos dos aspectos.

Ejemplos y contenidos matemáticos

Consideramos que los autores de los textos, en alguna medida, reflejan tanto la importancia como la dificultad de los contenidos matemáticos a través de los ejemplos que contienen los temas. La Tabla 3.2.1.4 muestra el cruce entre nociones y los ejemplos de la correlación, significándose con una "c" que dicho contenido está incluído en el libro de texto, y con una "e" que hay un ejemplo en él.

Como puede observarse en ella hay un número significativo de manuales - (77D), (78A), (86A) y (90A) - que no ejemplifican ninguno de los contenidos matemáticos expuestos. Los textos (77C), (81A), (82A) y (88A) ilustran con ejemplos entre el 25 y el 50 % de los contenidos desarrollados. Solamente tres libros de texto - (77A), (77B) y (87A) - muestran sus contenidos ejemplificados en una proporción igual o superior al 50 %, siendo el libro (77B) el que presenta mayor índice, con dos de cada tres conceptos con su ejemplo correspondiente. Una proporción tan pequeña de ejemplos por contenido es una dificultad que se añade al trabajo de los alumnos.

Otra característica que podemos destacar es que hay varios contenidos, *Correlación: concepto, Otras medidas de la correlación* y *Propiedades del coeficiente de correlación,* que no están ejemplificados en ninguno de los manuales de la muestra, aunque en el caso de *Propiedades del coeficiente de correlación*, como veremos posteriormente, se sustituyen los ejemplos por gráficas, ya que sirven de ejemplificación de una noción más que de ilustraciones de ella.

Únicamente el *Coeficiente de correlación* se presenta con ejemplos en más del 50 % de ellos, lo que induce a considerar que dicha noción es la que se estima más importante por los autores de todas las concernientes a la correlación.

Tabla 3.2.1.4. Contenidos y ejemplos sobre correlación que incluyen los textos

		77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
Dependencia	Funcional	ce					ce					c
	Aleatoria	ce					ce					c
Independencia aleatoria		ce					c					c
Diagrama de dispersión			ce	c	c	c	ce	ce	c	c	ce	c
Covariación				ce								
Correlación: concepto			c	c		c	c	c	c	c	c	c
Correlación (tipos)	Directa		ce	c			c	c	c	ce	ce	
	Inversa		ce	c			c	c	c	ce	ce	
	Independencia		ce	c				c	c	ce		
Correlación (medida)	Covarianza	c	ce			c	c	ce		ce	c	c
	Coeficiente correlación	c	ce	ce	c	c	c	ce	ce	ce	c	ce
	Otras	c	c			c						
Propiedades coeficiente correlación			c	c	c	c	c	c	c	c	c	c

Un cruce análogo hemos llevado a cabo entre los contenidos matemáticos y los ejemplos de la regresión incluídos en los manuales de la muestra, recogiéndolo en la Tabla 3.2.1.5.

Todos los textos presentan ejemplos para algún contenido, pero la proporción de conceptos ejemplificados alcanza cuando mucho el 40 %, siendo sólo dos manuales, (82A) y (90A), los que llegan a esta cota, lo que nos parece, como en el caso de la correlación, una cantidad bastante insuficiente. Igualmente existen dos contenidos, *Recta de regresión de y sobre x* y *Recta de regresión de x sobre y*, que se encuentran ejemplificados en más de la mitad de los libros en lo que están presentes.

Tabla 3.2.1.5. Contenidos y ejemplos sobre regresión que incluyen los textos

		77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
Ajuste lineal		e			c	c	c	c				c
Regresión: concepto				c	c		ce				c	c
Derivadas parciales					ce	c		c				c
Rectas de regresión	y sobre x	ce	ce	ce	ce	ce	ce	ce	ce	ce	ce	ce
	x sobre y	c	c		c	ce	c	ce	ce	ce	ce	ce
Coeficiente de regresión			ce			c	c					
Centro de gravedad		c			c	c	c	c	ce		e	

Posición de los ejemplos respecto al contenido

En su obra "Psicología del aprendizaje de las matemáticas" Skemp (1.980) exponía su primer principio del aprendizaje de las matemáticas: *"Los conceptos de un orden más elevado que aquellos que una persona ya tiene, no le pueden ser comunicados mediante una definición, sino solamente preparándola para enfrentarse a una colección adecuada de ejemplos"* (pág. 36); añadiendo a continuación: *"La gran mayoría de los libros de texto, pasados y presentes quebrantan el primero de estos principios. En casi todos se ven nuevos temas, introducidos no a base de ejemplos, sino por definiciones de la más admirable brevedad y exactitud para el profesor (que ya posee los conceptos a los cuales se refieren), pero ininteligibles para el estudiante"* (p.36).

En consecuencia, es importante la ubicación de los ejemplos cuando se introduce un concepto nuevo (Contreras y Sánchez, en prensa). De todos los textos estudiados solamente en el (88A) todos los ejemplos preceden a los conceptos que se van a desarrollar; otros tres - (77A), (77B) y (81A) - muestran algunos casos en que los ejemplos anteceden a los contenidos. En el resto todos los ejemplos van a continuación. Esto podría ser debido a una típica presentación del modelo teoría-práctica, dado que los autores se encuentran persuadidos por *"una metodología de enseñanza donde el conocimiento a enseñar, primero se formaliza,*

luego es traducido algebraicamente, para pasar a aplicarlo, por último, a la resolución de ejercicios" (Ruiz Higueras, 1.991, pág. 181).

Las gráficas como ejemplos

Finalmente hemos considerado las gráficas cuando se emplean como ejemplos y no como representación gráfica de un ejemplo concreto. Los autores de los manuales se sirven de ellas para ejemplificar, por ejemplo, el tipo de dependencia (positiva, negativa, independencia). En este sentido, podemos considerar que la gráfica cumple un papel de *ideograma*, ya que *"es un signo gráfico que representa una idea"* (Lacasta, 1.995, pág. 130), teniendo, esencialmente, *"una función de mera comunicación"* (Lacasta, 1.995, pág. 131). Los resultados se dan en la Tabla 3.2.1.6. Podemos destacar de esta tabla que hay dos textos - (77B) y (78A) - que no presentan ninguna gráfica como ejemplo. Es, además, ostensible que los manuales (77A), (87A) y (90A) son los únicos que poseen más de la mitad de gráficas con respecto a las nociones que poseían alguna.

Tabla 3.2.1.6. Gráficas utilizadas como ejemplos en los textos

		77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
Dependencia funcional	Parabólica	X										X
	Lineal				X		X	X				X
Diagrama de dispersión		X							X			X
Correlación	Fuerte	X						X	X	X		X
	Débil	X						X		X		X
Correlación	Positiva			X	X		X		X	X		
	Negativa			X	X		X		X	X		
Independencia		X		X	X		X	X	X	X	X	X
Recta de regresión		X						X		X	X	
Regresión no lineal				X						X	X	
Recta mejor ajuste	Mínimos cuadrados	X										
	Otras	X										

Puede observarse que los contenidos *Correlación fuerte, Correlación positiva, Correlación negativa e Independencia* presentan gráficas en más de la mitad de la muestra de libros seleccionada y, especialmente, el último se halla en todos los textos. Si comparamos con la Tabla 3.2.1.5 puede advertirse como algunos textos, los (77C), (77D), (81A), (86A) y (87A), sus autores han preferido usar las gráficas como ejemplos cuando se trata de las nociones de *Correlación positiva, Correlación negativa e Independencia* y no emplean otros tipos de ejemplos, dado que, en esta situación, la gráfica nos da toda la información necesaria para determinar qué clase de correlación es la presente.

Análisis de las demostraciones incluídas

Según la tradición fue Tales de Mileto (600 a. de C.) el primero que realizó algún tipo de demostración para un teorema - *"Todo círculo queda dividido en dos partes iguales por un diámetro"* - (Boyer, 1.986, pág. 76), y, desde entonces, una nota distintiva de todo trabajo matemático será la inclusión, en mayor o menor medida, de demostraciones que validen las proposiciones que se presentan. Por ello podemos considerar que *"la matemática es, pues, la disciplina con demostraciones"* (Davis y Hersh, 1.988, pág. 117). Incluso hay filósofos de las matemáticas, como Lakatos (1.986), que desplazan el centro de gravedad del progreso matemático de las demostraciones formales al proceso recursivo conjetura-refutación.

En educación matemática existe una abundante literatura sobre el análisis de las demostraciones. En los *"Estándares curriculares y de evaluación para la educación matemática"* de la Nacional Council of Teachers of Mathematics (NCTM, 1.991) se hace hincapié en el importante rol que desempeña la demostración en los currícula de matemáticas, recogiéndose bajo el título de "Las matemáticas como razonamiento -estándar 3 para los niveles 9 / 12-" las siguientes recomendaciones:

"En los niveles 9-12, el currículo de matemáticas debe incluir experiencias numerosas y variadas que refuercen y amplíen las destrezas de razonamiento lógico para que todos los estudiantes sean capaces de

- elaborar y comprobar conjeturas;

- formular contraejemplos;

- seguir argumentos lógicos;

- juzgar la validez de un argumento;

- construir argumentos sencillos válidos;

y para que, además, los futuros universitarios sean capaces de

- construir demostraciones para enunciados matemáticos, incluyendo demostraciones indirectas y demostraciones usando el principio de inducción" (pág. 147).

Más adelante, dentro del segundo y tercer objetivo del estándar, vuelve a subrayar la importancia que tiene, para los alumnos que continúen estudios universitarios, el aprender métodos de demostración más formales, básicos en un nivel matemático superior, así como establecer como centro de interés el de la demostración por inducción dada su relevancia dentro de la matemática discreta.

Una de las características típicas de los libros de texto de matemáticas correspondientes no sólo a nivel universitario sino también al de la enseñanza secundaria, que podemos circunscribir casi exclusivamente a los manuales de esta disciplina, es la presencia dentro de su organización de las demostraciones. En los textos que hemos analizado, como refleja la Tabla 3.2.1.7, existe una gran variabilidad entre ellos en cuanto a la presentación de demostraciones. Puede observarse que casi todos los libros de texto recogen alguna demostración. Dentro de la regresión hemos detectado un fuerte sesgo en las proposiciones que se tratan de demostrar , siendo la determinación de la recta de regresión la que, al encontrarse en 9 textos, recibe la mayor atención de los autores.

En el lado opuesto están la covarianza y el coeficiente de regresión que sólo se hallan en 1 texto, (82A) y (77B) respectivamente. Posiblemente, esto se deba a que, al tratarse de parámetros estadísticos, los autores consideren como esencial su definición, que, como citamos en la sección correspondiente, son de carácter instrumental, es decir, a través de una fórmula.

Tabla 3.2.1.7. Demostraciones presentadas en los textos

	77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A	Total
Covarianza							X					1
Coeficiente correlación		X					X	X			X	4
Propiedades coef correlación			X	X	X			X				4
Recta de regresión	X	X	X	X	X	X	X	X			X	9
R. regresión(m. simplificado)	X					X	X	X			X	5
Coeficiente de regresión		X										1

Todas las demostraciones desarrolladas son deductivas, lo cual, en un segmento educativo como el de la enseñanza secundaria, pudiera ir en detrimento de destrezas investigadoras (MacNab y Cummine, 1.992). Dada la dependencia que los estudiantes tienen respecto de los libros de texto, sería conveniente la inclusión en los mismos de *demostraciones en dos columnas* (NCTM, 1.991). No hemos encontrado ningún manual en el que se haga utilización de semejante herramienta didáctica.

Uno de los aspectos que recientemente se ha analizado con más profundidad es el de la utilidad o funciones de las demostraciones. Ya Bell (1.976) realizó una primera categorización de las funciones de la demostración, y destacó tres tipos: I) verificación; II) iluminación, y, III) sistematización. Una aportación más pormenorizada es la llevada a cabo por De Villiers (1.993), que amplía hasta cinco las funciones de toda demostración:

1. La demostración como medio de verificación/convicción
2. La demostración como medio de explicación
3. La demostración como medio de sistematización
4. La demostración como medio de descubrimiento
5. La demostración como medio de comunicación

La demostración como medio de verificación/convicción subraya que la demostración es el argumento que proporciona certeza absoluta de la veracidad de

lo enunciado. Era, y prácticamente sigue siéndolo, la función fundamental, aunque el binomio demostración-convicción en realidad debería establecerse en orden contrario, convicción-demostración, puesto que de la práctica de los matemáticos se infiere que se tiende a demostrar lo que previamente se está convencido de su validez (Davis y Hersh, 1.989). La demostración como medio de explicación indica que la función radica en la profundización en por qué es verdad, sirviendo la demostración como iluminación o clarificación. La demostración como medio de sistematización pretende la organización de varios resultados dentro de un sistema de axiomas, conceptos fundamentales y teoremas, lo cual la involucra estrechamente con procesos de axiomatización y definición a posteriori (Krygowska, 1971). Aparece únicamente a niveles muy avanzados. La demostración como medio de descubrimiento hace hincapié en la obtención de nuevos resultados en el proceso de alcanzar dicha demostración. Finalmente, la demostración como medio de comunicación enfatiza el hecho de que *"la argumentación matemática no es mecánica ni formal, ..., es un intercambio entre humanos basado en significados compartidos, no todos los cuales son verbales ni formulísticos"* (Davis y Hersh, 1.989, pág. 56).

En el análisis realizado de las demostraciones incluídas en los manuales hemos encontrado que son utilizadas como herramientas de convicción y de explicación, aunque el papel verificador de la demostración no tiene sentido para los alumnos (De Villiers, 1.993), siendo las demás funciones ignoradas.

III. PRESENTACIÓN DE LAS DISTRIBUCIONES DOBLES

Dos variables estadísticas unidimensionales

Son los libros que toman los pares de valores de una variable bidimensional como pertenecientes a variables unidimensionales y estudian la posible relación existente entre estos últimos. Es decir, comienzan el estudio de la regresión partiendo de dos variables unidimensionales. Los alumnos de secundaria cuando se enfrentan por primera vez con el tópico de correlación y asociación ya han

recibido nociones sobre dependencia funcional, ocupándose en este análisis de dos variables, la variable independiente y la variable dependiente, solamente ligadas por la relación de dependencia estudiada. Al comenzar a trabajar la dependencia aleatoria, que es una extensión de la dependencia funcional, parece natural y lógico que algunas opciones didácticas omitan los conceptos de distribución bidimensional y traten las variables estadísticas del mismo modo que se hacía al estudiar la dependencia funcional, tal y como se expresan en este sentido Hawkins y cols. (1.992),

> *"Hay una escuela de pensamiento que cree que es más razonable quedarse a un nivel introductorio con los temas que impliquen ocuparse de una distribución unidimensional (regresión) y diferir para después la parte del curso, o texto, cuya materia requeriría ocuparse de la distribución bidimensional conjunta (correlación)"* (Hawkins y cols., 1.992, pág. 52).

Este es el caso de los libros (77A), (77B), (77C) y (87A), en los que en algunos de ellos, aunque brevemente, se hace alusión a las distribuciones bidimensionales, de forma efectiva lo que se estudia es el grado de relación existente entre dos variables unidimensionales.

En todos ellos se omite el estudio de las frecuencias relativas dobles, momentos dobles, ..., siendo interesante reseñar cómo un texto, el (87A), a pesar de que realiza una introducción al tema repasando conceptos de Estadística sobre variables estadísticas unidimensionales, tampoco hace referencia a los hechos estadísticos anteriormente mencionados respecto a una variable estadística bidimensional.

Variables estadísticas bidimensionales

Cuando se introduce el tema, partiendo de las distribuciones de frecuencias bidimensionales o variables estadísticas bidimensionales, es decir, parte de una muestra donde se han recogido datos sobre dos características de la misma y construye una tabla de frecuencias doble (tabla de contingencia). A partir de la

tabla se estudia la asociación que pueda existir entre las dos características (libros 77D, 78A, 81A, 82A, 88A y 90A).

Tabla 3.2.1.8. Conceptos incluidos sobre variables estadísticas y aleatorias bidimensionales en los textos analizados

	TEXTOS											
	77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A	Total
Diagrama dispersión		X	X	X	X	X	X	X	X	X	X	10
Distribución conjunta				X		X	X					3
Frecuencias dobles				X		X	X			X		4
Frecuencia marginal				X		X	X			X		4
Frecuencia condicional				X								1
Momentos dobles				X								1
Momento condicional												
Momento marginal												

Con respecto a la opción precedente, los libros de texto anteriormente citados despliegan, de forma implícita o explícita, una serie de conceptos nuevos conectados a la idea de variable estadística bidimensional: frecuencia relativa doble, distribuciones marginales, momentos dobles, etc. El manual (78A) podemos considerarlo, desde el punto de vista de esta clasificación, como fronterizo, puesto que dedica mucho espacio a la exposición y cálculo de los estadísticos de las variables unidimensionales y, a partir de ellos, obtener la relación existente, en caso de que la haya, entre dichas variables. En este sentido el autor se expresa de la siguiente manera: *"El estudio así realizado da idea de que el estudio de las variables bidimensionales se reduce al estudio de dos variables simples. Esto no es así, podemos relacionar éstas por otra medida que enlaza ambas. Se trata de la covarianza"* (pág. 201). Sin embargo, con posterioridad hacen una presentación desde un enfoque propio de variable estadística bidimensional, aunque con un desarrollo de muy escasa profundidad, como también le ocurre al texto (90A).

En uno de los libros (77D) se calculan los momentos ordinarios y centrales. Asimismo, también expone este manual la idea de distribución condicional, destinando un epígrafe a la representación gráfica de una distribución de frecuencia bidimensional - estereograma -. No presenta ningún ejemplo de asociación entre variables cualitativas.

Variables aleatorias bidimensionales

Uno de los libros (86A), lleva a cabo una introducción al tópico a través del concepto de variable aleatoria bidimensional. Con tal motivo, expone las nociones de variable aleatoria bidimensional, función bivariable de probabilidad, función de probabilidad discreta bivariable y gráfica de la función de probabilidad de dos variables discretas. Obviamente, estos conceptos no se incluían en los textos de los apartados precedentes.

Otro manual, el (77D), después de introducir el tópico a partir de las distribuciones de frecuencias bidimensionales, añade otro tema sobre las variables aleatorias bidimensionales, estudiando en ellas los momentos, la correlación y las rectas de regresión. Este tema supone una mayor complejidad ya que se añaden los conceptos de función de distribución y densidad conjunta, relación entre ambas, distribuciones condicionales de probabilidad, momentos de las distribuciones conjuntas, marginales, etc.

Diagrama de dispersión

El diagrama de dispersión es la representación en coordenadas cartesianas de los valores de una variable estadística bidimensional cuantitativa. Mediante su observación podemos apreciar, intuitivamente, si en un conjunto dado de datos existe o no relación entre las variables (directa, inversa, independencia) y su forma (lineal o no). En general, los textos utilizan, al principio del tema, esta manera de aproximar al alumno a la noción de asociación. Hemos analizado en los libros de texto las características de los diagramas de dispersión presentados.

En la Tabla 3.2.1.9 se expone el número de diagramas de dispersión en cada uno de los libros estudiados, bien con sólo los puntos o también con una o las dos rectas de regresión. También se han incluído las representaciones de sólo una o las dos rectas de regresión. En dicha tabla se puede observar que el número de representaciones gráficas varía de un libro a otro, desde 20 para el (77A) a 3 del (78A). En dicha tabla analizamos el tipo de asociación presentada en estos diagramas.

Un hecho a tener en cuenta es la gran desproporción que existe en estas representaciones entre la asociación directa y la inversa o la independencia. En general, si consideramos las tres primeras filas de la Tabla 3.2.1.9, se puede observar que de las 103 representaciones, 72 se refieren a la asociación directa, 18 a la inversa y 13 a la independencia. Este hecho podría llevar a los alumnos a considerar que la asociación directa es el tipo de correlación estudiado más importante, quedando los otros dos como excepcionales.

Tabla 3.2.1.9. Frecuencia del tipo de asociación presentado en los diagramas de dispersión y/o rectas de regresión en los textos estudiados

TIPO DE ASOCIACIÓN	LIBROS DE TEXTO											
	77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A	Total
Directa	16	9	3	5	2	4	8	6	10	5	4	72
Inversa	1	2	1	2	1	4	1	1	3	2		18
Independencia	2	1	1	1		1	1	1	2	2	1	13
Otros: un punto, no lineal	1	1	2				1		1	2	1	9
Total	20	13	7	8	3	9	11	8	16	11	6	112

En consecuencia, este hecho debería tenerse en cuenta en la elaboración de nuevos manuales. Además, el grado de dependencia que suelen mostrar los diagramas de dispersión suele ser fuerte y sólo cuatro muestran una dependencia no lineal.

IV. ESTUDIO DE LA CORRELACIÓN

Dependencia funcional y dependencia aleatoria

Asociación o covariación

La idea de covariación como variación conjunta o sincronizada de dos variables es un concepto importante en el desarrollo del tema y es el punto de partida en el estudio de la correlación. Hemos contabilizado este extremo y, únicamente, en el libro 77C se define explícitamente - *"En muchos ejemplos puede apreciarse que ciertas variables presentan una sincronización más o menos intensa. Esta variación conjunta o sincronizada es lo que llamamos covariación"* (pág. 294) -. En otros, como el 77B y el 78A, aparece de manera implícita, es decir, se ofrecen las tablas de valores de las variables y se representan en el plano cartesiano, definiendo la nube de puntos y su utilidad para apreciar la relación entre las variables implicadas.

Dependencia aleatoria

En segundo lugar, se ha estudiado si en el desarrollo del tema el libro hace referencia a la diferencia entre dependencia funcional y dependencia aleatoria, ya que esta diferencia se debe tener en cuenta en el estudio de este tema por varias razones. La primera de ellas es que en cursos anteriores se ha estudiado, con exclusividad, la dependencia funcional. Este hecho, junto a la naturaleza de los contenidos matemáticos estudiados hasta aquí por los alumnos, pudiera proporcionar a los estudiantes una concepción determinista de la asignatura de matemáticas (la probabilidad se ha estudiado antes de tercero de B.U.P. muy poco). Por otra parte, la dependencia aleatoria extiende la dependencia funcional, ya que ésta es un caso particular de la primera (cuando la correlación es perfecta). Por tanto, creemos que la distinción entre ambas es fundamental al comienzo de este tema.

Hemos contabilizado los libros donde se hace referencia a estos dos conceptos, obteniendo que en 4 de ellos se hace alusión a los dos tipos de dependencia y en 7 no. En este último caso, se pierde la motivación subyacente a

la teoría de la regresión y de la correlación, que es la búsqueda de modelos los cuales permitan expresar la dependencia no determinista entre variables, que es la situación más frecuente cuando se trabaja con problemas reales.

Tipos de covariación

Por las experiencias personales anteriores, es posible que los estudiantes identifiquen a priori la asociación con la causalidad. Como es sabido asociación y causalidad no siempre son coincidentes, aunque uno de los pasos en la búsqueda de relaciones causales es estudiar la covariación de las variables (Pozo, 1.987). En Estepa (1.994) y Batanero, Estepa y Godino (1.996) se mostró que los alumnos tienen, en una proporción muy importante, dificultades en diferenciar asociación y causalidad. Cuando encuentran una asociación importante entre dos variables, tenderán a pensar que una de ellas provoca o es causa de la variación de la otra. Por todo ello, es necesario al comienzo del tema distinguir los distintos tipos de covariación. Según Barbancho (1.973), los tipos de covariación son:

1. *Dependencia causal unilateral*: Cuando la ocurrencia de una variable X influye en la ocurrencia de Y, pero no al contrario. La variable X se le llama variable independiente o causa, y la variable Y variable dependiente o efecto. Por ejemplo, la posición de la Luna influye en la altura del agua del mar en la marea, pero no al contrario.

2. *Interdependencia*: Cuando la ocurrencia de una variable X influye en la ocurrencia de una variable Y y viceversa, por ejemplo, la altura y el peso de un grupo de personas.

3. *Dependencia indirecta*: Dos variables pueden mostrar cierta covariación debido a la variación de una tercera variable que está correlacionada con ambas, produciendo una asociación aparente. Esta tercera variable podría no ser tenida en cuenta. La naturaleza de la relación observada entre dos variables puede, por consiguiente, cambiar radicalmente cuando tomamos en consideración otras variables que están ocultas en la situación.

4. *Concordancia*: Correlación producida por la ordenación de un conjunto de datos por dos personas de forma independiente. Por ejemplo, en el caso de que

dos jurados diferentes den una clasificación sobre los participantes en un concurso literario, con total independencia, puede interesar saber si existe concordancia o no entre las clasificaciones establecidas por ellos.

5. *Covariación casual*: Cuando parece que en la covariación de dos variables hay cierta sincronía, lo que se podría interpretar como la existencia de asociación entre ambas, sin embargo, la covariación es casual o accidental. Por ejemplo, cuando en unas competiciones deportivas se observa que existe cierta relación entre el color de la camiseta y los primeros puestos en las pruebas deportivas.

Covarianza

Otro concepto, relacionado con la parte fundamental de este tema, es el de covarianza. La covarianza o momento mixto de segundo orden viene dada por la expresión

$$\mathrm{cov}\,(X,Y) = \frac{1}{N} \sum_i \sum_j (x_i - \bar{x})\,(y_j - \bar{y})\,f_{ij}$$

cuando se introduce a partir de la distribución estadística conjunta y como

$$\mathrm{cov}\,(X,Y) = \frac{1}{N} \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

cuando se introduce a partir del estudio de dos variables estadísticas unidimensionales. Este estadístico nos proporciona un primer coeficiente de la asociación para variables cuantitativas. El signo de la covarianza nos indica el tipo de asociación (directa o inversa), como puede deducirse, fácilmente, de la expresión que la define. Esta expresión se transforma en la siguiente (simplificada)

$$\sigma_{XY} = \frac{1}{N} \Sigma\, x_i\, y_i\, n_i - \bar{x}\,\bar{y}$$

En caso de independencia, la covarianza toma un valor nulo.

No obstante, la covarianza no es independiente de las unidades de medida, lo que hace necesario definir un coeficiente con esta propiedad, que es el coeficiente de correlación.

En 2 de los libros analizados no aparece y en los 9 restantes se define este concepto de manera algorítmica como $\sigma_{XY} = \frac{1}{N} \Sigma (x_i - \bar{x})(y_i - \bar{y})$. De estos 9 libros, en tres de ellos sólo se define con la fórmula anterior; en otros tres se

demuestra la fórmula $\sigma_{XY} = \frac{1}{N} \Sigma x_i y_i n_i - \bar{x}\,\bar{y}$, que facilita los cálculos; en otros se propone como ejercicio demostrar la última fórmula; finalmente, en otros dos, se trata la covarianza teniendo en cuenta los momentos, que previamente se han definido. Un aspecto que nos parece significativo es que ninguno de los textos que incluyen el estudio de la covarianza relaciona el signo de ésta con la dependencia existente, a pesar de que en varios de los manuales - en concreto cuatro - se muestran ejemplos con covarianzas de signo tanto positivo como negativo.

Correlación

Definiciones de correlación en cada uno de los libros

Una vez examinados los conceptos de dependencia aleatoria y funcional, covariación y covarianza, nos centramos en el estudio de la correlación. Comenzamos analizando las definiciones de la misma.

Algunos manuales, simplemente, no definen el término. Es el caso del texto (77A), que no presenta la noción de correlación, incluyendo, en cambio, el concepto de coeficiente de correlación lineal, el cual es relacionado con las rectas de regresión, definiéndolo en la forma siguiente: *"se llama coeficiente de correlación lineal al número* ***r*** *:* $r = \sqrt{b \cdot b'}$ *"* (pág. 311) donde b y b' son las pendientes de la recta de regresión de y sobre x y de x sobre y , respectivamente.

Un segundo grupo de textos definen la correlación como el estudio de la dependencia entre variables, como, por ejemplo, el manual (81A): *"Llamaremos correlación a la teoría que trata de estudiar la dependencia que existe entre las dos variables que intervienen en una distribución bidimensional"* (pág. 441).

En este sentido amplio, la correlación englobaría todo estudio de la relación entre variables estadísticas y no sólo la determinación de un coeficiente o coeficientes que midieran la intensidad o la forma de la relación. Bajo esta acepción es como se usa este término en algunos manuales universitarios como, por ejemplo, Smirnov y Dunin-Barkowskij (1.978): *"La aplicación fundamental que encuentra la teoría de la correlación se refiere a la solución del problema de predicción aumentada, es decir, de la indicación de los límites en los cuales, con*

una seguridad dada por adelantado, se hallará la magnitud que nos interesa si otras determinadas magnitudes con ella relacionada, obtienen determinados valores" (pág. 358); indicando más adelante: *"En la teoría de la correlación se analizan primeramente estas dos características (medias y dispersiones) de las distribuciones condicionales"* (pág. 360), refiriéndose tanto a la regresión de X sobre Y como a la de Y sobre X. En este sentido amplio, el concepto o la teoría de la correlación incluiría también el estudio de la regresión, mientras que en un acepción más restringida la correlación se ocuparía del estudio de la intensidad de esta relación y la regresión del ajuste de un modelo a los datos.

Por ejemplo, Lóbez Urquía y Casa Aruta (1.975) indican que *"se hace necesario complementar el análisis de la regresión con la obtención de unas medidas o coeficientes que permitan calibrar el grado de dependencia estadística existente entre dos variables, o dicho de otro modo, el grado de representatividad o bondad de la función analítica ajustada a los datos obtenidos empíricamente por observación"* (pág. 106).

Tabla 3.2.1.10. Tipos de definiciones sobre la correlación en los textos analizados

TIPO DE DEFINICIÓN	TEXTOS
No incluye	77A, 77D
Relación causal	77B
Dependencia / Relación entre variables	81A, 82A, 87A, 90A
Medida de la intensidad de la relación	77C, 78A, 86A, 88A

El primero de estos puntos es el presentado, por ejemplo, por la definición de correlación en el texto 77C (*"Una de las técnicas para el estudio de la covariación de dos variables es la de la correlación. Este procedimiento consiste en buscar números que indiquen la intensidad de la variación conjunta que tienen las variables"*, pág. 303). En este punto no se menciona la idea de bondad del ajuste, aunque alguno, el texto 77A, la introducirá posteriormente. Destacamos, también, el texto 77B, pues presenta una definición de correlación que mueve a equívoco, pues

indica que *"La relación causal que pueda existir entre dos características (variables), se denomina correlación entre ambas variables"* (pág. 317). Claramente se hace referencia a una relación de tipo causa-efecto entre las variables, cuando lo que existe es una dependencia de tipo aleatoria entre ellas.

Como resumen incluímos la Tabla 3.2.1.10 con los tipos de definiciones sobre la correlación en los textos analizados.

Definición del coeficiente de correlación

Respecto al propio coeficiente de correlación, también hemos encontrado una variedad de definiciones:

i) Directamente, a partir de la fórmula $r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$, como, por ejemplo, el texto 77C, en el cual se expone: *"..., se define el llamado coeficiente de correlación lineal, mediante la expresión*

$$r = \frac{S_{xy}}{S_x S_y}$$ " (pág. 303)

Esta definición tienen la ventaja de relacionar los signos de la covarianza y el coeficiente de correlación. Sin embargo, ningún manual hace hincapié en esta característica, presentándolo como una modificación sobre la covarianza, para obtener un coeficiente adimensional. Por ejemplo, en el libro de texto 87A se indica el inconveniente de utilizar la covarianza como medida de la correlación entre dos variables pues *"sus unidades están relacionadas con las unidades de las variables de partida, ..., lo cual impide la posibilidad de comparar correlaciones entre variables bidimensionales expresadas en unidades diferentes"* (pág. 347). Respecto al método de cálculo, sería sencillo, una vez obtenidas la covarianza y las desviaciones típicas de las variables.

ii) Como raíz cuadrada del producto de las pendientes de las rectas de regresión. Este modo de definir al coeficiente de correlación es el utilizado por los textos 77A, 82A, 86A y 90A. Por ejemplo el manual 77A indica que *"si b, b' son los valores hallados en las ecuaciones de las dos rectas de regresión, se llama coeficiente de correlación lineal al número r: $r = \sqrt{b \cdot b'}$ "* (pág. 311). Inclusive, el libro de texto 82A presenta una definición errónea: *"La correlación se mide mediante el coeficiente de correlación lineal, que se define así*

$$r = \pm\sqrt{\frac{m_y}{m_x}}$$

m_y = *pendiente de la recta de regresión de* $y_x = \dfrac{\sigma_{xy}}{\sigma_x^2}$

m_x = *pendiente de la recta de regresión de* $x_y = \dfrac{\sigma_y^2}{\sigma_{xy}}$ " (pág. 311 y 312)

Creemos que esta definición es errónea, dado que al indicar los autores que la pendiente de la recta de regresión X sobre Y es $\frac{\sigma_y^2}{\sigma_{xy}}$, confunden la variable independiente con la variable dependiente. Asimismo, puede inducir a los alumnos a considerar que el signo del coeficiente de correlación es doble, positivo y negativo, ya que el mismo procede del signo de la raíz cuadrada.

iii) En los manuales 77A, 77B, 86A y 87A se muestran otras definiciones sobre el coeficiente de correlación, complementarias de las expuestas anteriormente en dicho texto, dado que las consideran más operativas para la determinación de este parámetro. Por ejemplo, en el 86A se dice: *"... el coeficiente de correlación lineal viene definido por la expresión*

$$r = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sqrt{\left[\sum_{i=1}^{n} x_i^2\right]\left[\sum_{i=1}^{n} y_i^2\right]}}$$ "

donde $x_i = x_i - \bar{x}$, e , $y_i = y_i - \bar{y}$, indicando a continuación:

"Observa que el signo de r depende del signo de $\sum_{i=1}^{n} x_i y_i$ " (pág. 260)

En el texto 87A se significa que para un cálculo más cómodo del coeficiente de correlación dicho concepto se define como:

$$"r = \frac{\sum x_i\, y_i - N\bar{x}\,\bar{y}}{\sqrt{\sum x_i^2 - N\bar{x}^2}\,\sqrt{\sum y_i^2 - N\bar{y}^2}}$$ " (pág. 348)

puesto que dicho algoritmo se puede determinar con una tabulación conveniente.

En el manual 81A se da la siguiente definición:

"Llamaremos coeficiente de correlación lineal, y se designa por ***r****, a la media geométrica de los coeficientes de regresión lineal. Es decir:*

$$r = \sqrt{\frac{S_{xy}}{S_x^2}\frac{S_{xy}}{S_y^2}} = \frac{S_{xy}}{S_x S_y}$$ " (pág. 443)

Esta definición, al incluir el concepto de media geométrica, añade un plus de dificultad a la noción de coeficiente de correlación, ya que dicho concepto no se encuentra explicitado, en ninguna parte, en este libro de texto.

Hemos recogido en la Tabla 3.2.1.11 los distintos tipos de definiciones del coeficiente de correlación presentes en los libros analizados, asi como si indican alguna relación entre el signo de éste con el de la covarianza o el de la pendiente de las rectas de regresión.

Tabla 3.2.1.11. Tipos de definiciones del coeficiente de correlación en los textos analizados

TIPO DE DEFINICIÓN	TEXTOS										
	77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
$r = \sigma_{XY} / \sigma_X \sigma_Y$	X	X	X	X	X	X			X	X	
$r = (b_{XY} b_{YX})^{1/2}$	X						X	X			X
Otras	X	X						X	X		
Relación signo con covarianza											
Relación signo con pendientes		X									
Relación signo con $\Sigma x_i y_i$								X			

Propiedades e interpretación del coeficiente de correlación

Para utilizar correctamente el coeficiente de correlación es necesario conocer y comprender algunas de su propiedades. Aparte de las ya señaladas en relación con el signo de la covarianza y con las pendientes de las rectas de regresión, podemos clasificar las demás propiedades en tres grandes apartados: I) aquéllas que hacen referencia a alguna característica de la correlación, II) aquéllas que hacen referencia a alguna característica de la regresión, y, III) aquéllas que aluden a alguna cualidad del coeficiente de correlación.

Dentro del primer bloque están la mayoría de las propiedades presentadas en los manuales. Normalmente se utilizan para definir algún tipo de correlación (correlación directa fuerte, correlación inversa fuerte, correlación perfecta, no existe correlación) en base a un determinado valor que alcanza el coeficiente de

correlación. Por ejemplo, en el libro 81A se dice: *"Si* $0 < r < 1$ *la correlación será más intensa a medida que r se aproxima a* 1, *y tanto más débil a medida que r se aproxime a cero. En este caso la correlación es positiva o directa, y se dice que las variables* X *e* Y *están en dependencia aleatoria"* (pág. 443).

Se incluyen en el segundo apartado las que, a partir de ciertos valores del coeficiente de correlación, nos ofrecen información acerca de las rectas de regresión. Así, en el libro de texto 81A, se indica que: *"Si* $r = 1$ *se deduce que todos los puntos de la variable bidimensional* (X,Y) *se encuentran sobre la recta de regresión"* (pág. 443).

Finalmente, el apartado III) abarca aquellas propiedades que hacen alusión a características intrínsecas del coeficiente de correlación. Únicamente hemos encontrado una que tenga esta cualidad, estando en 8 de los 11 textos analizados. Por ejemplo, en el manual 86A se la expone de la siguiente manera: *" Si r es el coeficiente de correlación lineal, se verifica* $-1 \leq r \leq 1$*"* (pág. 261).

Es significativo mencionar que hay un sólo libro de texto, el 77A, que no muestra ninguna propiedad del coeficiente de correlación.

Otro aspecto importante es la interpretación del coeficiente de correlación lineal, según los valores que tome. En 10 de los libros se interpreta el coeficiente de correlación y en uno no.

Influencia de los valores atípicos en el coeficiente de correlación

Cuando se lleva a cabo el análisis estadístico de un conjunto de datos, uno de los aspectos que más se debe cuidar es la depuración de los mismos, tanto en su precisión como en su fiabilidad, lo cual nos permitiría obtener el máximo de información. Ciertas observaciones las denominaremos *outliers* si *"siendo atípica o/y errónea tiene un comportamiento muy diferente respecto al resto de los datos frente al análisis que se desea realizar sobre las observaciones experimentales"* (Muñoz y Pascual, 1.986, pág. 36). Es por todo ello que los valores atípicos (*outliers*) de una de las variables, o de ambas, tienen una gran influencia en el coeficiente de correlación, llegando, en casos extremos, a cambiarlo de signo según se tome en consideración dicho valor atípico o no (NCTM, 1.991; Estepa y Sánchez Cobo, 1.994). A pesar de ser un aspecto relevante, sobre todo en el

estudio de casos reales, ya que un valor atípico puede eclipsar el verdadero valor del coeficiente de correlación, ninguno de los 11 manuales lo analiza.

Correlación y causalidad

Como hemos señalado, el estudio de la correlación viene motivado, con frecuencia, por el interés en encontrar relaciones causales entre las variables. Recordamos que un fuerte valor del coeficiente de correlación no siempre se debe a la existencia de una relación causal entre las variables, como vimos al analizar los tipos de covariación.

Hemos observado que en 9 de los 11 libros no se clarifica este punto. En un libro de texto, el 77C, si se hace, señalándose que *"para definir el coeficiente de correlación no se exige una dependencia causal entre las variables"* (pág. 305). Pero en otro, el 77B, se confunde explícitamente correlación y causalidad, ya que define la correlación como *"la relación causal que pueda existir entre dos características (variables), se denomina correlación entre ambas variables"* (pág. 317).

Coeficiente de determinación y su interpretación

El cuadrado del coeficiente de correlación es el coeficiente de determinación, y nos indica la proporción de varianza explicada por la regresión. Nos da una interpretación alternativa del coeficiente de correlación, al indicar que un valor grande del mismo denota que una de las variables puede dar cuenta de un alto porcentaje de variabilidad de la otra. Consideramos que es éste un atributo destacado que debe ser tratado en este tema. Sin embargo, únicamente el manual 77D alude, a partir de una propiedad, a él: *"Por lo tanto*

$$r^2 = \frac{s_y^2 - s_r^2}{s_y^2}$$

es el tanto por uno de la mejora de la varianza. Nos mide en cuanto ha mejorado nuestra predicción al pasar de la simple predicción que a cada x_i *hacía corresponder* $\bar{y}$ *, a la predicción que establece que a cada* x_i *corresponde*

$$r \frac{s_y}{s_x} (x_i - \bar{x}) + \bar{y}$$

es decir, el valor de y *dado por la recta de regresión de* y *sobre* x*."* (pág. 300)

En cuanto a la descomposición de la varianza, el texto (77D) demuestra la relación $s_r^2 = s_y^2(1 - r^2)$.

V. ESTUDIO DE LA REGRESIÓN

Una vez que el estudio de la correlación indica la existencia de una relación suficientemente intensa, observamos que el diagrama de dispersión se distribuye alrededor de una línea ideal o tendencia. *"Un ajuste es la sustitución de un diagrama de dispersión por una línea que, sin que deba pasar por todos los puntos, se adapte lo mejor posible a ellos"* (Lóbez Urquía y Casa Aruta , 1.975, pág. 85). Surge así el problema de regresión: *"Para cualquier tipo de función de regresión que sea necesario ajustar a una cierta nube de puntos, el problema que se plantea es determinar los parámetros de la curva particular - perteneciente a una familia de funciones dada - que mejor se adapte a la muestra de datos particular"* (Batanero y cols., 1988, pág. 132).

Ya hemos visto que todos los libros utilizan diagramas de dispersión para estudiar este concepto. Un segundo aspecto a estudiar es si se hace referencia a la regresión lineal y a los distintos tipos de regresión no lineal. Sobre este aspecto 5, de los 11 libros estudiados, (77B), (77C), (81A), (88A) y (90A), citan los distintos tipos de regresión. Uno de ellos solamente estudia la regresión lineal sin hacer referencia a otros tipos de regresión, mientras que en los 5 restantes se deja entrever que, aunque en el tema presentado se analice la regresión lineal, existen otros tipos de regresión. Los tipos de regresión no lineal mencionados en los libros son: curvas de regresión, regresión parabólica, función afín, función cuadrática, función cúbica, función polinómica, función parabólica, función potencial, función exponencial y función hiperbólica.

En los manuales que se realiza la exposición de la regresión no lineal, se suele optar entre hacer una presentación genérica en la que se indique la existencia de otro modelo de regresión o exponer con más detalle otras posibles curvas de regresión. Así, un ejemplo de la primera opción es el libro de texto (88A) cuando nos dice: *"El determinar la línea apropiada dependerá de la forma que tenga la nube de puntos. La regresión es lineal cuando la línea es una recta, y no lineal en caso contrario. También puede ocurrir que tal línea no exista"* (pág. 161), mostrando a través de unos gráficos ejemplos de regresión no lineal. Como modelo de la segunda opción está, por ejemplo, el manual (81A) al expresarse de la

siguiente manera: *"A la hora de tener que realizar el ajuste de una línea de regresión a un diagrama de dispersión conviene tener en cuenta los siguientes puntos:*

1º. Elección de la línea de regresión

Se realizará de forma que la línea elegida sea la que mejor se ajuste al diagrama de dispersión. Algunos tipos sencillos de estas curvas son las siguientes:

1. Recta afín	$y = ax + b$
2. Parábola	$y = ax^2 + bx + c$
3. Cúbica	$y = ax^3 + bx^2 + cx + d$
4. Exponencial	$y = c\,a^x$
5. Hipérbola	$y = \frac{1}{a+bx}$ " (pág. 436)

Definición de regresión en los libros de texto

Cuando los autores definen la regresión se deciden por una doble alternativa, o bien se destaca que la regresión es la técnica matemática que ajusta una función a un conjunto de datos o resaltan el carácter predictivo de ella. El primer caso es el que se lleva a cabo de forma mayoritaria en los manuales examinados, (77B), (77C), (81A), (82A), (86A), (87A) y (90A).

Así, por ejemplo, en el libro (77C) se expone: *"El análisis de la regresión consiste en obtener una línea que se aproxime lo más posible a los puntos de la nube"* (pág. 298). Como ejemplo de los libros de texto que ponen el énfasis en la regresión como método de estimación está el (77D) que se expresa de la siguiente manera: *"Dada una variable bidimensional la teoría de la regresión tiene por objeto predecir o estimar los valores de una variable en función de la otra"* (pág. 296), aunque posteriormente apenas se destaca esta característica funcional de la regresión.

El texto (77A) es el único que no define la regresión, haciendo sólo referencia al origen histórico de dicho vocablo.

Método de los mínimos cuadrados

El objetivo de la regresión es encontrar una curva particular, dentro de una familia dada, que represente lo mejor posible la relación que hay entre las variables. La solución a este problema no es única, pues es posible ajustar a la distribución conjunta diferentes funciones matemáticas siguiendo distintos criterios de "bondad de ajuste" (Rius, Barón, Parras y Sánchez, 1.997).

Aunque existen otros criterios plausibles, el único procedimiento utilizado en los textos es el de ajuste de una recta por el método de los mínimos cuadrados. Por este criterio se trata de elegir, entre todas las rectas, aquella tal que la suma de los cuadrados de las desviaciones verticales de los puntos a la recta sea mínima. Esta es la recta de regresión de Y sobre X que minimiza el error cuadrático medio cometido al mantener en cada punto fija la variable X y sustituir el valor y por el correspondiente sobre la recta. Una razón para la popularidad de la regresión mediante el método de los mínimos cuadrados es que los coeficientes obtenidos para los parámetros de la recta tienen una interpretación sencilla en términos de estadísticos ya conocidos, como podemos ver en la ecuación de la recta de regresión de Y sobre X (Moore, 1.995):

$$\hat{y} = a + b_{yx}\, x$$

con pendiente $b_{yx} = r \frac{\sigma_y}{\sigma_x}$, y ordenada en el origen $a = y - b_{yx}\, x$.

El uso del método de mínimos cuadrados, como anteriormente expresábamos, es generalizado en todos los libros, siendo solamente dos textos, el (77A) y el (88A), los que citan otros posibles procedimientos para obtener la recta de regresión.

El manual (77A) además de mostrar otro método, que para los alumnos puede ser muy intuitivo, indica cuáles son sus dificultades. Este es un aspecto esencial cuando se expone cualquier procedimiento matemático, que es comúnmente ignorado. Únicamente, subrayando los aspectos ventajosos y los incovenientes es como los alumnos pueden integrar los métodos de forma pertinente, entendiendo el interés que los matemáticos tienen por buscar otros procedimientos alternativos para resolver un problema ya "resuelto". No olvidemos

que desde el momento en que se alcanza una solución de un problema, éste deja de ser el objetivo del quehacer matemático para serlo el método y su posible mejora.

Uno sólo de los manuales (88A) no lo menciona y da las ecuaciones normales sin deducirlas y 10 deducen las ecuaciones normales por este método.

El manual (77D) destaca la noción de derivada parcial, nuevo para los alumnos de este nivel - 3º de B.U.P. -, dedicándole un epígrafe específico. De esta manera, los autores intentan abordar la dificultad que se les plantea al tener que resolver un problema de extremos relativos de funciones de dos variables, cuando los alumnos a los que va destinado este libro de texto sólo conocen la derivabilidad de funciones reales de variable real y la determinación de sus máximos y mínimos. Con respecto a este punto, es el único texto que afronta esta situación con una breve introducción sobre este concepto matemático, pues los demás lo que hacen es indicar, sobre el problema concreto, cómo se hallaría la derivada parcial correspondiente.

Las rectas de regresión e interpretación de sus parámetros

Otra forma de expresar la recta de regresión de Y sobre X es mediante la ecuación

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

En esta ecuación advertimos otra propiedad importante: la recta de regresión pasa por el punto formado por las dos medias, o centro de gravedad de la distribución doble. Hemos observado que en 7 - (77B), (78A), (81A), (82A), (86A), (87A) y (88A) - de los 11 libros de texto se cita tal hecho estadístico, aunque hay uno, el (88A), que lo hace dentro de los ejercicios propuestos a los alumnos. Dos no lo hacen de manera expresa, mientras que otros dos no citan dicha propiedad a pesar de que luego presentan en el método de cálculo abreviado de las rectas de regresión aquella - $y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$ - que, precisamente, resalta dicha característica.

Una expresión estadística que escasamente aparece en los textos, únicamente en (77A) y (90A), es la ecuación de la recta de regresión reducida, indicándose en ambos casos las técnicas matemáticas que conducen a su obtención.

Además, 7 de los 11 manuales desarrollan el método de cálculo abreviado de las rectas de regresión. De los cuatro que no muestran dicho procedimiento, hay dos - (77A) y (90A) - que son, justamente, los dos que presentaban la ecuación de la recta de regresión reducida, por eso estimamos que puede que por este motivo no lo incluyan dentro de los contenidos del tema.

Hemos visto que la pendiente de la recta de regresión viene dada por la expresión $b_{yx} = r \frac{\sigma_y}{\sigma_x}$. De ella podemos deducir las siguientes propiedades:

1) El coeficiente de regresión b_{yx} tiene el mismo signo que el coeficiente de correlación.

2) El coeficiente de regresión es positivo si la dependencia es directa, negativo si es inversa y nulo en caso de independencia.

3) En caso de dependencia directa la recta de regresión de Y sobre X es creciente, decreciente cuando hay dependencia inversa y paralela al eje X en caso de independencia.

4) Dados tres de los coeficientes r, s_y , s_x y b_{yx} podemos determinar el cuarto.

Tres de los libros (77B), (82A) y (87A) interpretan la pendiente de la línea de regresión o coeficiente de regresión, induciendo a partir de ella el tipo de correlación (directa, inversa, independencia). Ninguno de los ocho libros restantes tienen en cuenta este aspecto. Entre los primeros podemos, por ejemplo, citar el texto (77B) que dice: *"En el diagrama de la figura 7.2 - número de choques con relación a la distancia - también podemos ajustar una línea recta, en las mismas condiciones de la anterior, pero vemos que la línea forma un ángulo obtuso con el sentido positivo del eje de abscisas. Sabemos, por lo estudiado anteriormente, que la pendiente de la recta, m', es negativa; por este motivo decimos que existe una correlación lineal negativa"* (pág. 318). En ninguno de estos tres casos se relaciona

el signo de la pendiente de la línea de regresión con el signo del coeficiente de correlación o de la covarianza. Es plausible que sea debido a que dichos parámetros se utilizan más como medida de la correlación que como indicadores del tipo de la misma. El resto de las propiedades mencionadas anteriormente no se contemplan en los manuales seleccionados.

Cuando intercambiamos el papel de las dos variables X e Y, tomando Y como variable independiente obtenemos la ecuación de la recta de regresión de X sobre Y, que viene dada por

$$(x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

Esta recta también pasa por $(\bar{x}, \bar{y})$, aunque su pendiente es diferente a la de la recta de regresión de Y sobre X. El criterio usado ahora para ajustar la recta es minimizar el error cuadrático medio obtenido al mantener fija la variable Y y sustituir la x por la correspondiente sobre la recta.

Diez de los libros estudiados tienen en cuenta las dos rectas de regresión, Y sobre X y X sobre Y. Un manual, el (77C), sólo estudia la recta de regresión de Y sobre X. Dentro de los diez textos que exponen las dos rectas de regresión, la mitad de ellos - (77B), (82A), (86A), (87A) y (90A) - indican de forma expresa que ambas rectas serán generalmente distintas, consideración que nos parece que debe de resaltarse a los alumnos. Algunos manuales utilizan como argumento de esta diferencia razones de tipo analítico. Por ejemplo el (87A) dice a este respecto: *"Observación: Puede sorprender que las dos rectas de regresión ($\underline{y}$ sobre $\underline{x}$ y $\underline{x}$ sobre $\underline{y}$) sean diferentes (aunque en algún caso coincidan como se explicará más adelante); pero así es, debido a que se obtienen minimizando dos sumas de cuadrados diferentes"* (pág. 355).

Sin embargo, nos parece más oportuna la reflexión que haga aflorar el carácter aleatorio de la dependencia, como indica el manual (90A): *"La razón de que ambas rectas no coincidan es precisamente que la dependencia es aleatoria y no funcional; en este último caso ambas rectas se confundirían"* (pág. 287). Asimismo, creemos conveniente destacar que un libro (77A) manifiesta que normalmente sólo nos interesa determinar una recta de regresión en un problema

concreto, cuestión que está íntimamente relacionada con los tipos de covariación existente entre las dos variables estadísticas analizadas.

En la Tabla 3.2.1.12 recogemos los aspectos más destacados que hemos observado sobre la regresión.

Tabla 3.2.1.12. Resumen sobre la regresión en los textos analizados

		77A	77B	77C	77D	78A	81A	82A	86A	87A	88A	90A
Definición de regresión	Ajuste		X	X			X	X	X	X		X
	Predictor				X	X					X	
Regresión	Lineal	X	X	X	X	X	X	X	X	X	X	X
	No lineal		X	X			X				X	X
Método mínimos cuadrados		X	X	X	X	X	X	X	X	X		X
Otros métodos		X									X	
Centro de gravedad			X			X	X	X	X	X	X	
Ecuación reducida de la regresión		X										X
Cálculo abreviado recta regresión				X	X	X	X	X	X	X	X	
Coeficiente regresión	Indica tipo correl.		X					X		X		
	No lo expresan	X		X	X	X	X		X		X	X
Bondad del ajuste		X										
Rectas regresión	Analíticamente		X						X	X		
(diferencias)	Depend. aleatoria							X				X

3.2.2. ANÁLISIS DE EJERCICIOS

Otra cuestión que hemos examinado son los ejercicios que se incluyen en la muestra de 11 manuales de Matemáticas de 3º de B.U.P., siendo pertinente su estudio dado que constituyen un poderoso test para juzgar al libro de texto en el cual se encuentran insertados (Cobb, 1.987). Los profesores, cuando realizan la planificación de esta materia, tienen que elegir la muestra de ejercicios que ellos consideran más conveniente para el afianzamiento de los conceptos matemáticos de sus alumnos. Los estudiantes deben el aprendizaje de una noción, en parte, a la

cantidad y diversidad de ejercicios que resuelven, adquiriendo con la práctica una experiencia acumulada que les ayudará a formar el conocimiento de tal concepto (González, 1.993).

Además, los ejercicios son un indicador de las representaciones que los autores de los manuales tienen sobre los temas a los que están referidos, ya que parece natural que trate de resaltar aquellos hechos, conceptos o procedimientos que considere básicos en la muestra que incluya de ellos . A un nivel más profundo son también un reflejo de las concepciones que los autores poseen de la enseñanza y aprendizaje de las matemáticas (Robert y Robinet, 1.989).

Los ejercicios cumplen una doble función. Son el hábitat natural donde el alumno pondrá en práctica los conocimientos adquiridos y también son un test que nos permite evaluar cuáles ha aprehendido y cuáles no. Muy a menudo, sirven como preparación del trabajo sobre un concepto o regla al que en un futuro, más bien inmediato, tendrán que enfrentarse los alumnos (Shuard y Rothery, 1.988).

También los ejercicios responden a unos objetivos que los autores desean ver cumplidos. Entre ellos podemos enumerar los siguientes:

"- adquirir una técnica o un tipo de razonamiento (movilizarlos)

- utilizar la noción en contextos diferentes

- dar sentido a una noción, a una técnica, a un teorema, hacerlos disponibles

- corregir errores persistentes

- retener lo que hay que retener sobre la noción

- mantener los conocimientos anteriores

- conjeturar, generalizar, aplicar, comprobar, aprender a plantearse cuestiones (pertinencia de tal o cual método, ventajas e inconvenientes -complementariedades-), descubrir los parámetros potenciales de una situación, aplicar la analogía" (Robert y Robinet, 1.989, pág. 21)

Las variables que hemos contemplado en el análisis de los ejercicios son las siguientes:

I. Contextos utilizados

II. Contenido matemático

III. Tipo de tarea

IV. Tipo de covariación

V. Tipo e intensidad de dependencia

I. CONTEXTOS UTILIZADOS

El contexto de los ejercicios fue uno de los puntos analizados en el trabajo de Navarro-Pelayo (1.991, 1.994). Al igual que esta autora, entendemos el contexto de un problema o ejercicio en el sentido de Kilpatrick (1.978), como campo de aplicación en el que se presenta el problema. Desde un punto de vista didáctico, las tareas que se proponen a los estudiantes deben estar contextualizadas en campos de interés del alumno (Jullien y Nin, 1.989). No olvidemos que, a partir de los trabajos de Tukey (1.977) se considera a la estadística como ciencia de los datos y un dato no es un número, sino información (numérica o no numérica) situada dentro de un contexto. Para Willet y Singer (1.992) cada conjunto de datos que proporcionemos al alumno debe ir acompañado de suficiente información sobre el contexto e incluso sobre los objetivos planteados al recoger estos datos. Asimismo, los datos deben ser relevantes e interesantes a los alumnos para permitirles un aprendizaje significativo. Hemos examinado los contextos en los que se proponen las tareas analizadas y hemos obtenido la Tabla 3.2.2.1. En esta tabla se puede observar que, prácticamente, la tercera parte de los ejercicios se dan descontextualizados, es decir, se ofrece al alumno unos pocos pares de números y se les pide que les apliquen alguna de las técnicas estudiadas sin indicar el contexto al que se refieren los datos.

Obviamente, al no haber contexto no hay interpretación de los resultados obtenidos. Sólo existe el mero cálculo de estadísticos, con el fin de llegar a dominarlo. Sin embargo, puesto que, en la actualidad, el problema del cálculo está resuelto con los ordenadores y las calculadoras científicas, son precisamente las actividades de interpretación y modelización las que debieran fomentarse en nuestros alumnos. En consecuencia, observamos aquí una tendencia de la

enseñanza hacia una actividad algorítmica, y que, debido a los avances infórmaticos actuales, no tiene ya sentido, ni desde el punto de vista conceptual ni desde el computacional (Botella, 1.996).

Los contextos más utilizados en el resto de los ejercicios son los que utilizan una ley físico-química, lo que produce unas correlaciones muy fuertes. No obstante, los fenómenos físico-químicos son usualmente descritos con modelos deterministas, siendo sólo el error aleatorio en la medición lo que hace que la dependencia no sea estrictamente funcional. Volveremos sobre este punto al hablar de la intensidad de las correlaciones en los ejercicios.

Tabla 3.2.2.1. Frecuencia y porcentaje de los ejercicios según su contexto

Contexto	Frecuencia	Porcentaje
Ley físico-química	38	13'1
Peso y altura	31	10'7
Talla padres-hijos	10	3'4
Otros contextos biológicos	14	4'8
Calificaciones	17	5'9
Economía	11	3'8
Experimentos aleatorios	18	6'2
Expresión matemática	18	6'2
Anuarios estadísticos	10	3'4
Otros	31	10'7
No hay	92	31'7
Total	290	100

El siguiente tipo de contexto, en cuanto a su frecuencia, es el referido a fenómenos biológicos, en cuyo campo han surgido los conceptos de regresión y correlación.

Las calificaciones en dos asignaturas (una de ellas, matemáticas) es otro de los contextos más empleados. El resto de contextos se refieren a: datos de carácter económico; datos propuesto para estudio que se han obtenido de algún Anuario estadístico (dinero en circulación e IPC según el año; alumnos matriculados y que terminaron sus estudios en la escuela de Topografía desde el curso 1961-62 al

curso 1970-71; aprobados y matriculas de honor en la Facultad de Ciencias en el curso 1970-71; tasa de natalidad en Francia durante el decenio 1946-1956,...). Este contexto es de singular interés, por ser los anuarios fuentes importantes para la recogida de datos en los estudios estadísticos.

También se ha planteado el uso de un experimento aleatorio para obtener la tabla de probabilidades, distribuciones marginales, rectas de regresión, coeficiente de correlación, etc...

El contexto que hemos denominado expresión matemática es aquél en que se ofrece al alumno algunos estadísticos o rectas de regresión y se le pide el cálculo de otros estadísticos. Este tipo de enunciado debería incluirse también en datos no contextualizados, lo que aumentaría el porcentaje total de éstos a un 37'9 por ciento. En cuanto a los contextos que hemos señalado con otros, y que serían aquéllos que no se pueden clasificar en las categorías anteriores, son: peso y antigüedad de monedas o el estudio de una prueba realizada a 10 estudiantes sobre la expresión oral y la habilidad manual.

En conjunto el espectro presentado de aplicaciones es amplio, aunque predominan los ejercicios descontextualizados.

II. CONTENIDO MATEMÁTICO

Un interrogante que consideramos pertinente plantearnos es cuáles de los contenidos teóricos que se presentan en los manuales se ponen en juego en los diversos ejercicios analizados. Los tipos de contenido que hemos considerado son los siguientes:

a) Construir la tabla de frecuencias de una distribución bidimensional. En este tipo de ejercicio se daría a los alumnos una serie de pares de valores de dos variables estadísticas, debiendo, a partir de ellos, construir una tabla de frecuencias bidimensional, agrupando o no alguna de las variables en intervalos de clase. El alumno debe recordar el concepto de frecuencia absoluta doble y realizar un recuento adecuado, disponiendo la tabla según los convenios aprendidos.

b) Hallar las distribuciones marginales o las condicionales a partir de una tabla de frecuencias bidimensionales. El alumno debe poner en juego los conceptos de frecuencias marginales y/o condicionales absolutas y relativas. Asimismo es preciso discriminar los dos tipos de frecuencias marginales y condicionales. Consideramos que esta actividad tiene un gran interés, ya que la línea de regresión se define precisamente como lugar geométrico de las medias de las distribuciones condicionales. Además, la comparación de distintas distribuciones condicionales de la variable Y variando X, es un primer medio de estudio de la asociación entre variables. Por último, las dificultades descritas sobre interpretación de frecuencias condicionales (Estepa, 1.994) y probabilidades condicionales (Falk, 1.986) aconsejan este tipo de ejercicios.

c) Cálculo de momentos, en general, y cálculo de medias, desviaciones típicas o covarianzas, en particular. Se trata de ejercitarse en el cálculo de los parámetros de la recta de regresión y del coeficiente de correlación. Este punto es, en la actualidad, menos necesario, debido a que los ordenadores e incluso las calculadoras realizan este cálculo en forma automática.

d) Cálculo o interpretación del coeficiente de correlación. Mientras que el cálculo es un ejercicio rutinario, nos parece fundamental que el alumno realice actividades interpretativas tanto referidas al signo como a la magnitud del coeficiente de correlación.

e) Representación gráfica del diagrama de dispersión a partir de un conjunto de datos.

f) Cálculo de los coeficientes de regresión, de una o de las dos rectas de regresión, o del ángulo que forman. Esta actividad es similar a la c) aunque ahora se refiere a las rectas de regresión.

En la Tabla 3.2.2.2 presentamos la clasificación de ejercicios según contenido. En el 81 por ciento de los ejercicios se pide alguna de las nociones básicas de la regresión y correlación lineal como: coeficiente de correlación, recta de regresión Y sobre X, diagrama de dispersión, dos rectas regresión, recta de regresión X sobre Y, covarianza. El contenido matemático más solicitado a los

alumnos ha sido el coeficiente de correlación, bien en su cálculo o en su interpretación. Con una frecuencia similar aparece la recta de regresión de Y sobre X.

Hay que destacar, además, que en 121 ejercicios (41'7 por ciento) se propone únicamente el cálculo de la recta de regresión, sin cálculo previo de la covarianza o del coeficiente de correlación y sin representar gráficamente los datos. Quiere decirse que no se hace un estudio inicial de la posible existencia de asociación entre las variables, sin la cual el ajuste de una recta no tiene el menor sentido. El estudio aislado del coeficiente de regresión aparece con una frecuencia muy baja.

Tabla 3.2.2.2. Frecuencia y porcentaje de ejercicios según su contenido matemático

Contenido matemático	Frecuencia	Porcentaje
Tablas de frecuencia	4	1'4
Distribuciones marginales	3	1'0
Momentos	4	1'4
Media	10	3'4
Desviación típica	16	5'5
Covarianza	17	5'9
Coeficiente de correlación	67	23'1
Estudiar el tipo de relación	9	3'1
Diagrama de dispersión	35	12'1
Coeficiente de regresión	7	2'4
Recta de regresión de Y sobre X	60	20'7
Recta de regresión de X sobre Y	30	10'3
Dos rectas de regresión	31	10'7
Ángulo rectas de regresión	1	0'3
Total	290	100

En general, el número de ejercicios interpretativos es muy bajo, lo mismo que los referidos a la elaboración de tablas de frecuencias, obtención de distribuciones

marginales o representación gráfica. Es decir, se presta poca atención a las actividades de interpretación cuya relevancia hemos comentado.

Si cruzamos estos contenidos con los analizados en la sección 3.2.1.2, podemos detectar que hay algunos que, aunque se incluyen dentro de la presentación teórica del tema, no se utilizan en los ejercicios propuestos. Así, con respecto a la correlación, las nociones de dependencia - funcional y aleatoria -, independencia aleatoria y covariación no se encuentran reflejados en los ejercicios. Tampoco el ajuste no lineal y el centro de gravedad tienen cabida dentro de los ofertados para la regresión. En todos los casos, son contenidos que no se hallan en la mayoría de los textos, pero en aquellos en los que se encuentran no se les presta mucha atención, a tenor del peso específico que se les concede en la propuesta de ejercicios que se exhibe.

III. TIPO DE TAREA

Otro aspecto importante en el estudio de los ejercicios propuestos a los alumnos en los libros de texto es el tipo de tarea solicitada en los mismos. Los tipos de tarea que hemos considerado son los siguientes:

Cálculo: cuando el ejercicio se limita al cálculo numérico de resúmenes estadísticos, como los momentos, medias y desviaciones típicas o de los coeficientes de correlación y regresión. La finalidad del ejercicio es puramente algorítmica. Como hemos comentado, esta es una capacidad que actualmente es innecesaria para la aplicación de la estadística a los problemas reales.

Interpretación: Si se pide al alumno interpretar los valores de estadísticos o coeficientes de asociación como la covarianza o el coeficientes de correlación para deducir, a partir de los mismos, el tipo de relación entre las variables.

Representación gráfica de la distribución bidimensional mediante diagramas de dispersión o histogramas bidimensionales: Este aspecto de la representación gráfica es, a nuestro juicio, importante, porque permite dar al alumno una imagen visual de lo que entendemos por asociación entre las variables. Además, es cada

vez más frecuente la aparición de este tipo de gráficos en los medios de comunicación. La familiarización del alumno con los mismos le permitirá una interpretación correcta de este tipo de gráficos y detectar posibles errores en su elaboración.

Predicción: Cuando se pide al alumno predecir valores de una de las variables conocida el valor de la otra, empleando los resultados sobre la regresión y la correlación. Ésta es, justamente, la principal finalidad del estudio de la regresión, especialmente cuando una de las variables es difícil o costosa de medir.

Comprobación de propiedades de los coeficientes o de los estadísticos: Se trataría de problemas de "probar" en la terminología de Butts (1.980). Mediante esta actividad los alumnos ejercitan su capacidad de reconstruir e incluso desarrollar por si mismos los pasos de demostraciones matemáticas sencillas. Usualmente estas demostraciones requieren desarrollos de tipo algebraico y aplicación de algunas conocidas propiedades de los conceptos en juego.

Comparar los grados de asociación o los valores de los coeficientes en diferentes conjuntos de datos.

Recoger y analizar datos: Cuando se le pide al alumno, por ejemplo, que averigüe la talla y el peso de sus compañeros de clase y realice un estudio estadístico con lo estudiado en el tema.

Los resultados se presentan en la Tabla 3.2.2.3. En primer lugar, destaca con casi el 50 por ciento el cálculo. Lo anterior nos hace suponer que el objetivo perseguido es que el alumno aprenda a *calcular*, dejándose en un segundo plano la interpretación y la representación gráfica, los cuales son dos de los principales fines de la educación estadística.

Los 8 ejercicios de "recoger y analizar datos" de la Tabla 3.2.2.3 corresponden al único enunciado de todos los libros en que se pide a los alumnos que realicen un estudio estadístico completo de la variable estadística bidimensional (talla, peso) de todos los alumnos de la clase. Es la única actividad de todos los libros estudiados en que se pide a los alumnos que recojan datos y los estudien. Abogamos por un mayor peso de este tipo de tareas, no sólo por su

relevancia dentro de la educación estadística, sino porque harían aflorar las dificultades que los alumnos muestran en el análisis de datos (Konold y cols., 1.996), en especial la que tienen con la idea de asociación.

Tabla 3.2.2.3. Frecuencia y porcentaje de ejercicios según el tipo de tarea

Tipo de tarea	Frecuencia	Porcentaje
Cálculo	143	49'3
Interpretación	53	18'3
Representación gráfica	40	13'8
Predicción	30	10'3
Comprobación	12	4'1
Comparación	4	1'4
Recoger y analizar datos	8	2'8
Total	290	100

IV. TIPO DE COVARIACIÓN

En el presente trabajo se ha estudiado el tipo de covariación presentado en los ejercicios, respecto a los citados por Barbancho (1.973), obteniéndose la Tabla 3.2.2.4. En ella se puede observar, en primer lugar, que la máxima frecuencia corresponde a la categoría "No hay contexto". Ello es debido a que en muchos ejercicios no se expresa el contexto y no se puede encontrar el tipo de covariación que se presenta. Además, éste tipo de ejercicios sin contexto, ocultan la parte más motivadora e interesante de la estadística la interpretación de los resultados obtenidos y su aplicación a casos reales. Se crea, en consecuencia, un estereotipo del trabajo estadístico consistente en que debemos realizar largos y tediosos cálculos para encontrar una solución numérica.

Cuando se ha podido observar, respecto al tipo de covariación presentado en el ejercicio, hemos encontrado que el 93 por ciento de estos ejercicios (32 por ciento del total) tienen un tipo de asociación de interdependencia o dependencia causal unilateral. Esto pudiera influir en que algunos alumnos mantengan una concepción causal de la asociación estadística (Estepa, 1994).

Tabla 3.2.2.4. Frecuencia y porcentaje de los ejercicios según tipo de covariación

Tipo de covariación	Frecuencia	Porcentaje
Interdependencia	90	31'0
Dependencia causal unilateral	79	27'2
Dependencia indirecta	12	4'1
Concordancia	1	0'3
No hay contexto	108	37'2
Total	290	100

V. TIPO E INTENSIDAD DE DEPENDENCIA

En la Tabla 3.2.2.5, se presenta el tipo de dependencia mostrados en los ejercicios. Se puede observar la desproporción existente entre los tres tipos de dependencia presentados en los ejercicios. Creemos que esto puede contribuir a que el alumno mantenga una noción de correlación errónea, incitándole a creer que la mayoría de las veces que resuelve problemas de correlación debe existir correlación directa, haciendo que la correlación inversa y la independencia parezcan casos excepcionales. En consecuencia, creemos que debería darse un mayor énfasis al estudio de la correlación negativa que la abreviada que ofrecen los libros analizados.

Tabla 3.2.2.5. Frecuencia y porcentaje de ejercicios según tipo de dependencia

Tipo de dependencia	Frecuencia	Porcentaje
Directa	181	62'4
Inversa	62	21'4
Independencia*	44	15'2
No hay datos	3	1'0
Total	290	100

* Se ha tomado independencia si el valor absoluto del coeficiente de correlación es inferior a 0.3

El valor absoluto del coeficiente de correlación lineal nos indica la intensidad de la correlación presentada en los ejercicios. En nuestro estudio sólo el 18'3 por ciento de los ejercicios propuestos tienen un valor inferior a 0'5, el 52'7 por ciento de los ejercicios tienen un coeficiente de correlación superior a 0'9 en valor absoluto, prácticamente el 40 por ciento de los ejercicios tienen un valor para el coeficiente de correlación superior a 0'95 en valor absoluto y el 23'5 por ciento de los ejercicios tienen un valor del coeficiente de correlación igual o superior a 0'99 en valor absoluto.

Todo esto nos indica una fuerte desviación en la muestra de ejercicios estudiados hacia las correlaciones altas y muy altas, trabajando en muy escasa proporción las correlaciones de tipo medio y bajas, las cuales también se presentan en los casos reales. Esto puede inducir al alumno a pensar que para que exista asociación estadística el valor absoluto del coeficiente de correlación debería ser muy alto, situación poco frecuente cuando se trabaja con datos reales.

3.3. CONTENIDOS DEL CURSO DE INICIACIÓN A LA ESTADÍSTICA EN LA UNIVERSIDAD

Al realizar el análisis de la programación del profesor que imparte esta asignatura, hemos querido tener una primera aproximación a la enseñanza que sobre las nociones de la correlación y de la regresión, presumiblemente, han recibido los alumnos, los cuáles, con posterioridad, conformaron la muestra a las que se les ha pasado el cuestionario. Para ello, hemos solicitado al profesor de los grupos de alumnos que participaron en la investigación una copia de los apuntes de clase, la cual hemos incluído en el Anexo III. En esta sección analizamos este material. El estudio efectuado es similar al realizado con los libros de texto en la investigación de Sánchez Cobo (1.996), y que hemos resumido en la sección 3.2, aunque teniendo siempre presente las diferencias entre ambos materiales curriculares.

3.3.1. METODOLOGÍA DE LA PRESENTACIÓN DEL TEMA

Entre los aspectos de carácter general que hemos analizado está si el autor incluye o no objetivos. A pesar de que se trata de unos apuntes de clase, no obstante, el autor los incluye, aunque de una forma implícita, al inicio del tema. Así, indica:

> *"Al comenzar el estudio de las variables estadísticas bidimensionales ya comentamos que nuestro objetivo era conocer si habrá algún tipo de dependencia entre ambas variables unidimensionales"* (pág. 13)

No hemos detectado ninguna referencia histórica.

En cuanto a los contenidos estadísticos exhibidos están la inmensa mayoría de los estudiados en la investigación de Sánchez Cobo (1.996), a saber:

<u>Correlación</u>

1. Dependencia funcional y aleatoria
2. Independencia aleatoria
3. Diagrama de dispersión
4. Correlación: concepto
5. Correlación (tipos):
    - 5.1. Directa
    - 5.2. Inversa
    - 5.3. Independencia
6. Correlación (medidas)
    - 6.1. Covarianza
    - 6.2. Coeficiente de correlación
    - 6.3. Otras
7. Propiedades del coeficiente de correlación

<u>Regresión</u>

1. Ajuste lineal
2. Regresión: concepto

3. Derivadas parciales
4. Rectas de regresión
    4.1. Recta de regresión de Y sobre X
    4.2. Recta de regresión de X sobre Y
5. Coeficiente de regresión

No se mencionan los distintos tipos de covariación (Barbancho, 1.973). También en los textos analizados en la investigación de Sánchez Cobo (1.996) hemos encontrado un hecho análogo, pues sólo uno de los 11 manuales incluía dicha clasificación, desarrollándola brevemente e insertando algunos ejemplos.

Por el contrario, respecto a los textos de bachillerato, se lleva a cabo una profundización en la noción de correlación, al introducir los conceptos de coeficiente de determinación y razón de correlación de Pearson, así como cuando se estudia el efecto de un cambio de origen y de escala sobre el coeficiente de correlación.

En cuanto a la regresión, nuevamente, como por otra parte es natural, la casi totalidad de los contenidos tratados en los textos de bachillerato estan presentes aquí. El único que, de forma explícita, no lo está es el centro de gravedad. No obstante hemos observado que el autor lo muestra en los diagramas, como, por ejemplo, en las Figuras 9 y 10 del Anexo III (pág. 21).

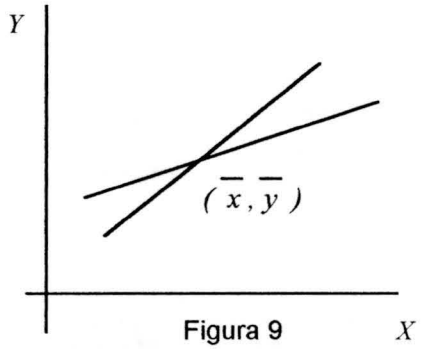


Figura 9

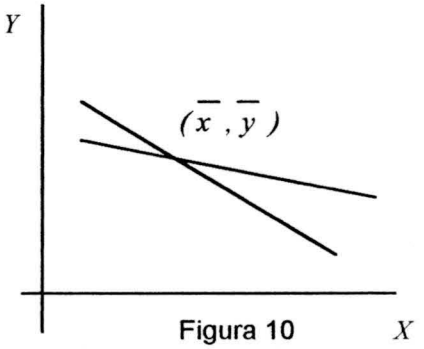


Figura 10

En este subtópico el autor dedica un mayor esfuerzo de profundización, pues, además de inserir la regresión de tipo I, despliega y analiza prolijamente diversas clases de ajustes no lineales: regresión hiperbólica, regresión potencial, regresión exponencial, etc. Finalmente, el orden de presentación elegido por el

autor ha sido primero la regresión y después la correlación, lo cual coincide con la mayor parte de los textos de bachillerato.

A diferencia de los manuales de 3º de B.U.P., la mayoría de las definiciones son del tipo relacional o instrumento-relacional, lo cual permite una mejor integración de dicho concepto. Por ejemplo, cuando expresa lo que es la regresión indica: *"La regresión consiste en la búsqueda de una función que exprese lo mejor posible la relación existente entre dos o más variables (en nuestro caso dos variables)"* (Anexo III, pág. 13).

Sigue un esquema de teoría-práctica, incluyendo un total de 10 ejemplos que muestran las técnicas procedimentales estadísticas que el autor considera más importantes en este tema:

- Cálculo de las rectas de regresión de Y sobre X y de X sobre Y;
- Regresión hiperbólica;
- Regresión potencial;
- Regresión exponencial;
- Coeficiente de correlación;
- Ajuste exponencial;
- Ajuste parabólico;
- Estudio de la bondad del ajuste.

Como ocurrió con los libros de texto de bachillerato, también aquí el autor utiliza representaciónes gráficas como ejemplos, presentando hasta un total de 18. Los conceptos así ejemplificados son: Dependencia aleatoria (parabólica y lineal), diagrama de dispersión, correlación (fuerte y débil), correlación (positiva y negativa), independencia, rectas de regresión, regresión no lineal, métodos para obtener la recta que mejor se ajusta (mínimos cuadrados y otros), regresión hiperbólica y centro de gravedad. Este último, como indicábamos con anterioridad, no se define ni se menciona en ningún lugar del tema, pero se encuentra de forma implícita en ellos.

No se incluye ninguna demostración en el texto.

3.3.2. ESTUDIO DE LA CORRELACIÓN

El autor expresa de forma tácita la idea de asociación o covariación, así indica que: *"Grado de asociación ... o lo que es lo mismo el grado de dependencia mutua"* (Anexo III, pág. 27), o cuando manifiesta que: *"Si* $\sigma_{XY} > 0$, *ya dijimos que ambas variables variaban con el mismo sentido"* (Anexo III, pág. 31). En cambio, si menciona explícitamente la dependencia aleatoria cuando expone que: *"Pero existen otros fenómenos en donde las variables presentan alguna relación pero en las que es imposible definir sobre ellas una función matemática que verifiquen exactamente. Este tipo de dependencia se le llama dependencia estadística"* (Anexo III, pág. 13).

Con respecto a los diferentes tipos de covariación - dependencia causal unilateral, interdependencia, dependencia indirecta, concordancia, covariación causal (Barbancho, 1.973) - que pueden darse, no se hace referencia alguna a los mismos.

La noción de correlación se definía en los libros de texto de bachillerato bajo una doble alternativa: como dependencia o relación entre dos variables o como la medida de la intensidad de dicha relación. El autor opta por esta última, aunque añade un matiz nuevo como es el de la relación entre correlación y bondad del ajuste, y así expresa que: *"Una vez que hemos determinado la forma en que se relacionan las variables se plantea el problema de medir el grado de asociación de las mismas, o lo que es lo mismo el grado de dependencia mutua. Además podremos comprobar si el ajuste realizado es bueno o no. Esto es lo que se llama correlación: el estudio de la bondad del ajuste de una curva"* (Anexo III, pág. 27).

Igualmente, el coeficiente de correlación admitía en los manuales no universitarios un enfoque diverso, que también son asumidos por el autor. Así, hemos encontrado que emplea una doble opción:

- Definición de tipo instrumental, cuando indica que: *"Definimos el coeficiente de correlación lineal* $r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ *"* (Anexo III, pág. 30)

- A través de su relación con los coeficientes de regresión lineal: *"Observemos que dicho coeficiente se puede obtener a partir de los coeficientes de regresión lineal de las rectas X/Y , Y/X: $b \cdot b' = r^2$ "* (Anexo III, pág. 30)

Además, significa que éste es el parámetro estadístico que muestra el grado de asociación existente entre dos variables: *"El coeficiente de correlación lineal sirve para medir el grado de asociación lineal entre dos variables, pero no de una forma cuantitativa sino cualitativa"* (Anexo III, pág. 30).

Sí pone de manifiesto la relación entre el signo del coeficiente de correlación lineal *r* y el de la covarianza σ_{XY} (*"El signo de r depende del signo de la covarianza"* [Anexo III, pág. 30]), pero por el contrario no menciona la relación del signo del coeficiente de correlación lineal *r* con el signo de las pendientes ni con $\sum x_i y_i$.

En cuanto a las propiedades del coeficiente de correlación *r*, no muestra ninguna que haga referencia a las características de la correlación, si a las que hacen referencia a las características de la regresión, por ejemplo cuando dice: *"Si r = 0, implica que $\sigma_{XY} = 0$, y las rectas de regresión son $y = \bar{y}$, $x = \bar{x}$, paralelas a los ejes de coordenadas"* (Anexo III, pág. 31) y a las que hacen alusión a las propiedades intrínsecas del propio coeficiente de correlación lineal *r*, como son el campo de existencia de él y que es invariante ante cambios de origen y de escala.

No se estudia la influencia de los valores atípicos, ni la relación entre correlación y causalidad.

Entre los conceptos que no se presentan en los manuales de bachillerato y si se hace en el nivel universitario está el de coeficiente de determinación. Aquí el coeficiente de determinación es definido como *"la proporción de la varianza total de Y que aparece explicada por la regresión*

$$R^2 = \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = 1 - \frac{\sigma^2_e}{\sigma^2_Y}$$ " (Anexo III, pág. 29)

También se incide en que el coeficiente de determinación es adimensional, lo cual es importante para poder efectuar comparaciones, exponiéndose a continuación el campo de existencia y la interpretación de este parámetro

estadístico. Además, se pone de manifiesto que es la herramienta que se debe utilizar para determinar la bondad del ajuste:

"En definitiva, el coeficiente de determinación mide el grado de ajuste de la recta de regresión" (Anexo III, pág. 29).

Otros aspectos novedosos para los alumnos son la bondad del ajuste, que se desarrolla con cierto detenimiento, y la razón de correlación de Pearson.

Se finaliza, dada la modalidad de titulación en la que se imparte este curso de introducción a la Estadística, con unas aplicaciones de las técnicas de la correlación y de la regresión a la economía.

3.3.3. ESTUDIO DE LA REGRESIÓN

Es en este apartado donde mayor profundización se lleva a cabo con respecto al nivel de bachillerato. No sólo se recoge la regresión lineal, sino que se amplía con la regresión de tipo I y la regresión no lineal, la cual se desarrolla con notable exhaustividad.

En la definición de regresión se tiene en cuenta tanto su carácter matemático de aproximación de una función como su condición de útil empleado en la predicción. De ahí que se indique que:

"La forma en que estudiaremos esta relación es mediante la regresión. La regresión consiste en la búsqueda de una función que exprese lo mejor posible la relación existente entre dos o más variables (en nuestro caso dos variables).

La principal aplicación que puede tener el estudio de la regresión es la de predecir: conociendo el valor de una de las variables estimar el valor que presentará la otra variable, de forma que ese valor estimado sea lo más exacto posible" (Anexo III, págs. 13-14)

Se detalla y se ejemplifica el método de los mínimos cuadrados, significándose cuáles son las carencias de otros procedimientos.

En cuanto a las rectas de regresión y la interpretación de sus parámetros, únicamente se da la siguiente expresión de la recta de regresión

$$y - \bar{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{x})$$

no mencionándose, como ya aludíamos anteriormente, a la propiedad de que el centro de gravedad pertenece a dicha recta.

De los parámetros analizados hay que destacar el coeficiente de regresión, definiéndose ambos, siendo, por ejemplo, la definición del coeficiente de regresión de Y sobre X:

"Vamos a llamar $b = \frac{\sigma_{XY}}{\sigma_Y^2}$ *coeficiente de regresión de* Y/X*"* (Anexo III, pág. 20)

Se mencionan una serie de propiedades de dicho coeficiente de regresión *b*:

- b es la pendiente de la recta
- b mide la tasa de incremento de Y para variaciones unitarias de X
- Los signos de b y b' dependen del signo de la covarianza
- b y b' son invariantes ante un cambio de origen y de escala

En cambio no se contemplan otras como:

- El coeficiente de regresión b tiene el mismo signo que el coeficiente de correlación
- El coeficiente de regresión es positivo si la dependencia es directa, negativo si es inversa y nulo en caso de independencia
- En caso de dependencia directa la recta de regresión de Y sobre X es creciente, decreciente cuando hay dependencia inversa y paralela al eje X en caso de independencia
- Dados tres de los coeficientes r, σ_Y , σ_X y b podemos determinar el cuarto

que se encuentran expresadas en la investigación de Sánchez Cobo (1.996).

No se alude a que, normalmente, para nuestro problema carece de interés determinar ambas rectas de regresión, lo cual está conectado con los diversos tipos de covariación (Barbancho, 1.973).

En la introducción a la regresión no lineal se estudian la regresión hiperbólica, la regresión potencial y la regresión exponencial, entre aquéllas a las que se les puede aplicar un proceso de linealización, y la regresión parabólica, entre las cuales no es posible desarrollar tal transformación.

3.4. ANÁLISIS DE LOS APUNTES DE LOS ALUMNOS

Como complemento del análisis de la enseñanza recibida por los alumnos de la muestra hemos examinado, además, los apuntes, correspondientes a este tópico, que han tomado dos de las alumnas pertenecientes al grupo participante, en el cual imparte su docencia el profesor, cuyos apuntes estudiábamos en el epígrafe anterior. El fin primordial de este análisis es observar qué contenidos son los que, realmente, se han desarrollado en clase y que, consideramos, el profesor estima son los más importantes de este tema, puesto que son primados a través de su exposición en el aula. Estos apuntes aparecen fotocopiados en el Anexo IV.

Hemos encontrado que los apuntes de ambas alumnas presentan una gran homogeneidad. En general, hemos detectado que existe una notable adecuación entre lo expuesto en clase y los apuntes elaborados previamente, básicamente en lo referido a los núcleos centrales de las nociones de correlación y de regresión: dependencia aleatoria y funcional, independencia aleatoria, diagrama de dispersión, covarianza, correlación, tipos de correlación, coeficiente de correlación, propiedades del coeficiente de correlación, ajuste lineal, regresión, rectas de regresión de Y sobre X y de X sobre Y, coeficiente de determinación, bondad del ajuste, etc.

Entre aquellas cuestiones que, encontrándose en la progamación del profesor, no han sido transmitidas a los alumnos en clase o lo han sido pero con menor profundidad están la regresión no lineal, bondad del ajuste para otras funciones y la razón de correlación de Pearson. Tampoco se indica que b y b' son los coeficientes de regresión de Y sobre X y de X sobre Y, respectivamente, a pesar de que estos conceptos se encuentran en los apuntes del profesor. Por el contrario,

se les menciona que las ecuaciones obtenidas en el ajuste mínimo-cuadrático son las ecuaciones normales, término que no se halla en la programación.

El número de ejemplos propuestos en clase han sido 9, fundamentalmente dedicados a las nociones de variable estadística bidimensional, distribuciones marginales y condicionadas, independencia estadística y representación gráfica. Es de subrayar que, prácticamente, no se ejemplifican la mayoría de los conceptos implicados en el tópico de la correlación y de la regresión.

El número de gráficas que se les ha mostrado a los alumnos ha sido de 12, siendo, al contrario del caso de los ejemplos, dedicados casi exclusivamente a la correlación y la regresión. Puede considerarse que, como se comprobó en la investigación de Sánchez Cobo (1.996), las gráficas se han utilizado como ejemplo para estas nociones, como *ideogramas* (Lacasta, 1.995).

3.5. CONCLUSIONES SOBRE LA ENSEÑANZA DE LA CORRELACIÓN Y REGRESIÓN

En este capítulo hemos presentado una aproximación al estudio de la forma en que la población de alumnos, de la cual proviene la muestra, podría haber sido introducida al estudio de la correlación y regresión. Es decir, nos hemos interesado por especificar el significado de la correlación y regresión en los cursos introductorios de estadística descriptiva en bachillerato y primeros cursos de universidad. Este estudio se ha basado en el análisis de los siguientes materiales:

* Una muestra representativa de libros de texto de tercer curso de bachillerato, que, en los cuestionarios vigentes de la época, era el nivel donde se introducían la correlación y regresión cuando los alumnos de la muestra cursaban la enseñanza secundaria.

* Los apuntes del profesor que ha impartido el tema a los alumnos de la muestra, así como los apuntes tomados en clase por dos alumnas participantes.

A continuación exponemos las conclusiones obtenidas, haciendo notar que el estudio tan sólo constituye una aproximación al análisis de la enseñanza, puesto que nos faltan datos sobre las actividades y tiempo dedicado a la enseñanza durante el bachillerato. Sí tenemos esta información para la enseñanza recibida por los alumnos el curso en el que fueron evaluados los conocimientos de éstos. Por otro lado, queremos hacer notar que el análisis de la enseñanza ha tenido otro objetivo importante, cual es el servir de base para la construcción de los instrumentos de evaluación usados en nuestro trabajo.

Metodología de la presentación del tema

A pesar de la diferencia de nivel (bachillerato y universidad), hemos encontrado bastantes similitudes en la metodología de presentación del tema a los alumnos. Dentro de las conclusiones de tipo general, tanto de los libros de texto de Matemáticas de 3º de B.U.P. como de la programación del profesor, queremos subrayar las siguientes:

Existe una presentación fuertemente deslizada hacia el esquema teoría-práctica, que, además, se encuentra fortalecida por la ubicación de los ejemplos en relación al concepto que ejemplifican. Consideramos que esto pudiera ser producto de las creencias que los autores de los manuales y los profesores poseen respecto al proceso de enseñanza-aprendizaje de nuestra disciplina. Esta presentación, tan formalista, impide que se proporcionen oportunidades para que se debatan ideas matemáticas, lo que posibilitaría al profesor detectar los conocimientos de sus alumnos, según se recomienda en N.C.T.M. (1.987). En la programación del profesor hemos observado que se mantiene, asimismo, este esquema.

En general, se ofrece una muestra insuficiente de ejemplos, tanto por la cantidad, como por el espectro de contenidos que ilustran, que permita a los alumnos comprender la noción presentada. En el caso de la programación del profesor, además, hemos detectado un reducido número de ejemplos, que, a su

vez, ha sido restringido cuando se ha llevado a cabo la exposición del tema, como se muestra en el análisis de los apuntes de las alumnas. Ello es explicable, debido a las limitaciones del tiempo disponible para la enseñanza y al número de propiedades y conceptos diferentes que se incluyen en el tema, lo cual hace difícil el poder dedicar el tiempo suficiente a mostrar ejemplos de cada uno de ellos.

Para resolver este problema, frecuentemente se emplean como ejemplos representaciones gráficas (diagramas de dispersión o rectas de regresión) insertadas en el texto. Hemos detectado un fuerte sesgo en las nubes de puntos presentadas, pues más de los dos tercios ejemplifican casos de asociación directa, siendo poco significativo tanto los dedicados a la asociación inversa como los que representan casos de incorrelación. Por el contrario, la programación del profesor comprende una gama equilibrada de diagramas que exhiben todas las posibilidades de tipos de correlación (positiva, negativa, independencia). Creemos que los autores de libros de texto y profesores deberían pensar con cuidado los ejemplos a utilizar, para de esta forma poner de manifiesto al alumno la riqueza de los conceptos y ayudarle a construir concepciones adecuadas sobre las nociones matemáticas.

Otro elemento indicador de las concepciones de los autores de los manuales sobre los procesos de enseñanza-aprendizaje de las matemáticas son los ejercicios. Respecto a ellos, se posee un margen de maniobra más amplio para su selección, que para la de los contenidos, que vienen impuestos por la autoridad educativa. En los textos analizados, prácticamente la mitad de los ejercicios responden a tareas de cálculo, ignorándose casi de forma total otras destrezas, como las de interpretación, recogida y análisis de datos o predicción, que son esenciales para la formación estadística de todo ciudadano.

Una carencia que hemos detectado en todos los libros de texto es la ausencia generalizada de actividades que abarquen las diferencias individuales de los estudiantes (N.C.T.M., 1.987), ofreciéndose ejercicios prototípicos para alumnos prototípicos.

Los libros de texto suelen incluir las demostraciones de las fórmulas de cálculo de la recta de regresión. En la programación del profesor y su posterior

exposición en el aula no se aborda, en general, ninguna demostración, posiblemente debido a la limitación del tiempo disponible para la enseñanza.

Contenidos incluídos

Del análisis de los contenidos de la correlación y de la regresión, tanto de los manuales de Matemáticas de 3º de B.U.P. como de la programación del profesor, debemos destacar las siguientes conclusiones:

Existen diversos modos de abordar el tema que son la presentación de las distribuciones dobles a partir de dos variables estadísticas unidimensionales, de una variable estadística bidimensional o de una variable aleatoria bidimensional. En la programación del profesor se realiza una aproximación a esta cuestión mediante la presentación de las distribuciones dobles mediante una variable estadística bidimensional.

La influencia de los valores atípicos (Nurhonen y Puntanen, 1.992) o el coeficiente de determinación y su interpretación son muy escasamente tratados en los libros analizados, a pesar de ser básicos y constitutivos de fuente de diversos errores de los alumnos. La confusión de correlación y causalidad (Estepa, 1.994) es otro punto poco iluminado en la muestra escogida, siendo presentados, en algún caso, como términos similares. De todo lo anterior, en la programación del profesor solamente se efectúa una aproximación el coeficiente de determinación y su interpretación, ignorándose las demás cuestiones.

Una de las características de la regresión es su valor predictivo, esto es, la posibilidad de estimar el valor de una variable a partir del valor de la otra variable (Martínez, 1.991). Sin embargo, este aspecto esencial se encuentra un tanto minusvalorado en la muestra de textos analizada, mientras que por el contrario se destaca su vertiente de ajuste de una función a un conjunto de datos, aunque se evite pronunciarse sobre la bondad de dicho ajuste. Además, no hay manual alguno que se plantee si dicha predicción es válida en un proceso de interpolación, extrapolación o en ambos. En la programación del profesor se aborda el valor

predictivo de la regresión, aunque en los ejemplos insertados no se haga referencia a esta cuestión.

El método de los mínimos cuadrados es universal dentro de la muestra de libros de texto estudiada. Es pertinente resaltar una dificultad inherente a la obtención de la recta de regresión por el método de los mínimos cuadrados, cual es la utilización de las derivadas parciales, puesto que, como reseñábamos antes, esta noción no ha sido trabajada con anterioridad por los alumnos de 3º de B.U.P (Romero y López, 1.998). A pesar de ello, únicamente se complementa este método de regresión lineal con una aproximación intuitiva mediante la utilización de nubes de puntos, pero no se potencia una vía manipulativa con el uso, por ejemplo, de artilugios analógicos como el de Hawley (Dewdney, 1.985), aspecto éste que consideramos básico. En la programación del profesor además del método de los mínimos cuadrados también se exponen otros, describiéndose sus ventajas e inconvenientes.

Consideramos que debería explicitarse más los distintos tipos de covariación existentes (Barbancho, 1.973), y relacionarse con la determinación de las dos rectas de regresión, dado que la mayoría de las ocasiones más que interdependencia lo que existe es una dependencia causal unilateral. Tampoco se mencionan los distintos tipos de covariación en la programación del profesor.

Ejercicios propuestos

Respecto al análisis de los ejercicios, realizado exclusivamente en los libros de texto de Matemáticas de 3º de B.U.P., consideramos pertinente señalar las siguientes conclusiones:

Hay un sesgo en el alto número de ejercicios que implican interdepencia o dependencia causal unilateral, lo que puede colaborar a que aflore en los alumnos una concepción causalista (Estepa, 1.994), al confundir correlación con causalidad (Estepa y Sánchez Cobo, 1.996b).

A pesar de ser esencial la contextualización de los ejercicios en campos de interés de los alumnos (Jullien y Nin, 1.989), hemos observado que se da una notable descontextualización de los mismos, lo que muestra a las matemáticas como abstractas y no relacionadas con la realidad (Buxton, 1.981), y que impiden que el alumno se ejercite en actividades como la interpretación o la predicción, que actuarían de catalizadores de los aprendizajes significativos de los estudiantes.

Otro sesgo que hemos descubierto es el producido por el tipo e intensidad de la dependencia de los ejercicios planteados. Prácticamente los dos tercios de los ejercicios propuestos plantean situaciones de dependencia directa. Asimismo, la intensidad del coeficiente de correlación está sesgada hacia valores muy altos, siendo muy inferior los casos de dependencia débil, lo cual pudiera crear el estereotipo de que cuando hay asociación ésta tiene que ser alta. En general, ocurre todo lo contrario cuando se trabaja con actividades del entorno real (Sánchez Cobo y Estepa, 1.996).

De los tipos de tarea que hemos considerado, cabe subrayar, que casi la mitad de los ejercicios tienen, únicamente, una finalidad algorítmica, pues sólo solicitan el cálculo de parámetros estadísticos. Posiblemente, los autores de los textos consideren que esta es la tarea fundamental que se debe potenciar en los alumnos, y que una experiencia con ejercicios de este tipo reforzaría su aprendizaje, a pesar de que no existe investigación que confirme que una metodología tan convencional sea una forma exitosa de aprender a resolver problemas (González, 1.993). Otros tipos de tareas fundamentales para la educación estadística como la interpretación, representación gráfica, predicción, comparación y recogida y análisis de datos apenas son tratadas.

Capítulo 4

Construcción del cuestionario

4.1. INTRODUCCIÓN

En este capítulo describimos el proceso seguido en la construcción del cuestionario que ha servido para la recogida de datos, analizando los items que lo componen, así como las variables de tarea de los mismos. Como hemos indicado en la sección dedicada a describir la metodología, el método de recogida de datos de nuestra investigación se encuadra en la medición, puesto que su finalidad es obtener datos a nivel profundo sobre los conocimientos de los alumnos. En este tipo de método planteamos a los sujetos unas tareas a nivel consciente para generar una medida de su conocimiento o destreza, que es un constructo no directamente observable, sino que debe ser inferido a partir del desarrollo de las tareas planteadas (Dane, 1.990).

Entre los tipos de técnicas utilizadas dentro del método de medición, nosotros hemos utilizado en la presente investigación la primera de ellas, el cuestionario, ya que por sus características lo hemos considerado especialmente adecuado a los fines de nuestro trabajo. El cuestionario es indicado para ser aplicado y valorado de forma uniforme en los distintos sujetos, incluso cuando se pase en días diferentes, siendo, según Scott (1.989), una técnica objetiva de

recogida de datos. En las siguientes secciones analizamos los objetivos, contenido y proceso de construcción del cuestionario, así como los items, las tareas y problemas que lo componen.

4.2. OBJETIVOS Y PROCESO DE ELABORACIÓN DEL CUESTIONARIO

Nuestro objetivo principal no es determinar el grado de maestría que los sujetos muestran sobre un dominio, en cuyo caso consideraríamos que el cuestionario elaborado para esta investigación sería una *prueba con referencia a criterio* (Thorndike, 1.989). Nosotros estamos interesados, principalmente, por el aspecto cualitativo - qué significado atribuyen a los conceptos que componen el dominio -, aunque también analizaremos, indirectamente, algunas variables cuantitativas y los errores que sobre las nociones de correlación y regresión se detecten.

De los objetivos explicitados para la investigación en el Capítulo 1, se deduce que prestaremos notable atención al significado que para los alumnos reviste la noción de la asociación estadística. Usaremos como indicadores empíricos (Carmines y Zeller, 1.979) para la medición de estos constructos inobservables (Babbie, 1.989) los juicios sobre la correlación y la regresión, los procedimientos aportados por las contestaciones de los sujetos y los razonamientos exhibidos por éstos, identificándose las estrategias utilizadas por los alumnos en la resolución de los problemas. Estamos igualmente interesados en identificar los errores de los alumnos, ya que pueden contribuir al proceso de aprendizaje. Puesto que los errores aparecen ligados a marcos conceptuales existentes en el alumno (Rico y Castro, 1.994), la instrucción debería tenerlos en cuenta y anticiparlos en la planificación de la enseñanza.

La muestra elegida se ha tomado entre los alumnos de primeros cursos universitarios. En concreto, se optó por una muestra seleccionada de las siguientes titulaciones:

- 1º de la Diplomatura de Ciencias Empresariales

- 1º de la Diplomatura en Enfermería

Naturalmente, tanto la asignatura de Introducción a la Estadística de la Diplomatura en Ciencias Empresariales como la de Bioestadística de la Diplomatura en Enfermería contienen entre sus descriptores el de Correlación y Regresión.

Fases en la construcción del cuestionario

En la construcción del cuestionario hemos procedido, siguiendo a Scott (1.989), a cumplimentar las siguientes fases:

a) Definición del propósito

b) Delimitación del contenido

c) Redacción de los items

d) Diseño del formato

e) Prueba piloto del cuestionario

A continuación, efectuaremos la descripción de estas fases, aunque en las dos secciones siguientes volveremos, con más detalle, sobre dos puntos importantes que son el b) y el c).

Propósito del cuestionario

En la sección 1.2. del Capítulo 1, ya se hacía mención a los objetivos que impulsan esta investigación. Todos y cada uno de ellos son referencia obligada a la hora de la elaboración del cuestionario, particularmente, el que alude al estudio evaluativo que, con los instrumentos construidos, se efectuará sobre una muestra de alumnos universitarios de los primeros cursos para caracterizar el significado que este tipo de estudiantes atribuye a las nociones de correlación y regresión al

finalizar un curso introductorio de estadística. Dicha evaluación tendrá como ejes los siguientes puntos:

1. Sesgos asociados a elementos de significado intensional del coeficiente de correlación, covarianza, rectas de regresión, tipos de covariación y relaciones entre correlación y causalidad. Es decir, queremos evaluar los ajustes o desajustes entre el significado de estos conceptos mostrado en el análisis que hemos realizado de los libros de texto y de los apuntes de clase y los construidos por los alumnos, puesto de manifiesto en los errores conceptuales que revelan en sus respuestas a los items del cuestionario.

2. Prácticas que realizan los alumnos para resolver problemas relacionados con estos conceptos. Entre ellas destacamos los siguientes:

2.1. Estimación de la correlación que hacen los alumnos a partir de diversas representaciones (verbal, gráfica y numérica). Esta estimación es para nosotros muy importante, pues de ella depende la toma de decisiones en las situaciones donde la covariación entre dos variables afecta a las consecuencias. La estimación adecuada de la correlación será un primer paso para resolver problemas relacionados con este concepto.

2.2. Interpretación que hacen los alumnos de valores numéricos del coeficiente de correlación, en particular, la construcción que realizan de situaciones asociadas a valores específicos del coeficiente de correlación, representadas en forma verbal y gráfica.

2.3. Estrategias de los alumnos en el ajuste de una recta a un conjunto de datos, cálculo del coeficiente de correlación y estimación de valores de las variables.

2.4. Interpretación del coeficiente de correlación y de las rectas de regresión en una situación problemática semejante a las que han resuelto durante su período de aprendizaje. Esta interpretación podrá compararse con la considerada en el apartado 2.2. en situaciones más abiertas.

La construcción del cuestionario se ha basado en el examen de la literatura pertinente de las cuatro últimas décadas sobre las investigaciones llevadas a cabo con respecto a las nociones de la correlación y de la regresión, tanto desde el enfoque de la psicología como desde el de la educación matemática. El correspondiente análisis se encuentra expuesto en el Capítulo 2. De igual modo, se ha apoyado en el análisis de la enseñanza que se muestra en el Capítulo 3.

Como ya indicábamos en el Capítulo 1, en la sección 1.4, una investigación que hemos tenido como referente para nuestro trabajo es la de Janvier (1.978, 1.987). Dado que la dependencia aleatoria es una generalización de la dependencia funcional, hemos tratado de ver si es posible extender el trabajo de Janvier para el estudio de las traducciones entre los distintos lenguajes de representación de la asociación estadística. Esta asociación puede venir dada o evocada verbalmente, por medio del conjunto de datos (pares de valores), gráficamente, en un diagrama de dispersión representando a estos datos en un sistema de ejes coordenados, y numéricamente, por medio del coeficiente de correlación que es un resumen estadístico del conjunto de datos. Cada una de estas formas de expresión remiten a la asociación - o falta de asociación - entre el par de variables dadas, no obstante cada una de ellas tiene una especificidad y una precisión determinada.

Por ejemplo, en la descripción verbal de las variables las ideas previas sobre la asociación pueden ser determinantes, pues no hay un apoyo de los datos concretos. La representación de éstos mediante el diagrama de dispersión permite visualizar la forma (lineal o no) y dirección (inversa o directa) de la asociación. Pero la mayor precisión para valorar no sólo el signo sino la intensidad la da el coeficiente de correlación. Es pertinente por ello, tratar de analizar como pasan los alumnos de uno a otro tipo de representación.

Puesto que la asociación estadística es un marco más amplio, hemos tenido que reformular el trabajo de Janvier a las características singulares de este concepto. Por ejemplo, en la asociación no tiene sentido hablar de fórmula, puesto que la fórmula del coeficiente de correlación es siempre única, mientras que su valor particular depende del conjunto de datos.

Es por todo esto, que hemos establecido las siguientes representaciones para la correlación: i) Descripción verbal, ii) tabla, iii) diagrama de dispersión, y, iv) coeficiente de correlación.

Otra diferencia relevante que podemos destacar es que en la dependencia funcional los procesos son biunívocos, o sea, por ejemplo a una ecuación corresponde una sola tabla, mientras que en la dependencia aleatoria existe una multiplicidad de posibilidades: a) Hay una función biunívoca entre tabla y diagrama de dispersión, ya que si se tiene la tabla se tiene el diagrama de dispersión y viceversa; por tanto cuando traducimos hacia el diagrama de dispersión, tenemos en realidad la tabla de valores; b) existe una función sobreyectiva entre tabla y coeficiente de correlación, y entre diagrama de dispersión y coeficiente de correlación, pues a un coeficiente de correlación se le pueden asociar distintos diagramas de dispersión, y, c) no hay función cuando se traduce de descripción verbal a tabla o diagrama de dispersión. Véase la Figura 1.4.1. Esto enriquece y dificulta las actividades de interpretación y predicción.

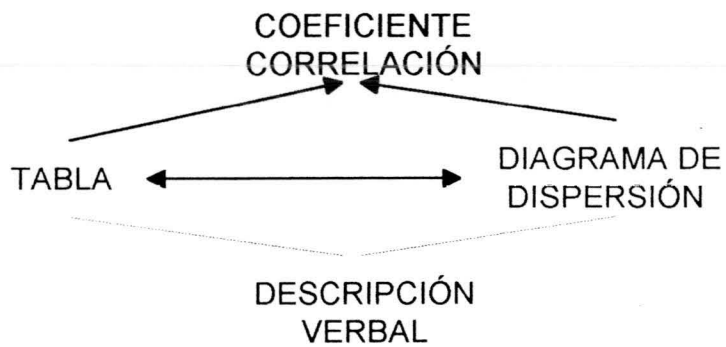


Figura 1.4.1

Tomando como referencia la Tabla 1.4.1, dada en el Capítulo 1 - sección 1.4 -, hemos diseñado una matriz - Tabla 4.2.1 - que ponga de manifiesto los diversos procesos de traducción entre las representaciones de la correlación, los cuales, como vemos, difieren de los correspondientes al caso de las funciones. Como juzgamos que muchas de las actividades de cambio de marco son bastante elementales para el alumnado universitario al que va dirigido este instrumento de

evaluación, consideraremos en la presente investigación, únicamente, las resaltadas en negrita en la Tabla 4.2.1.

Otra característica a destacar es que, a pesar de que algunas investigaciones - como por ejemplo Jennings, Amabile y Ross (1.982) - ha estudiado cierto tipo particular de traducción, ninguna de ellas, que conozcamos, ha elaborado un esquema general en el que se recoja por completo todas las posibilidades de traducción tal como el que hemos diseñado y que a continuación se ofrece.

Como en el caso de las funciones, también podemos distinguir entre traducciones directas e indirectas, considerándose, a priori, más complejas las segundas que las primeras, ya que podemos estatuir una ordenación no total en cuanto al tipo de información que proporcionan las cuatro representaciones de la correlación, dado que a partir de algunas categorías podemos obtener las otras -gráfico => tabla, coeficiente <= tabla => gráfico-, lo que implica que si pedimos una transformación, por ejemplo, de gráfico a coeficiente pudiera esta implícita las transformaciones intermedias y, por tanto, forzosamente debe ser más complicada.

Tabla 4.2.1. Traducciones de las representaciones de la correlación

Hacia \ Desde	Descripción verbal	Tabla	Diagrama de dispersión	Coeficiente de correlación
Descripción verbal		Tipo 1 Estimación	**Tipo 2 Estimación**	**Tipo 3 Estimación**
Tabla	Tipo 4 Lectura		Tipo 5 Lectura y trazado	**Tipo 6 Lectura y estimación**
Diagrama de dispersión	Tipo 7 Interpretación	Tipo 8 Lectura		**Tipo 9 Lectura y estimación**
Coeficiente de correlación	**Tipo 10 Interpretación**	Tipo 11 Interpretación y estimación	**Tipo 12 Interpretación, estimación y trazado**	

Para la regresión no parece factible determinar unas representaciones equivalentes a las anteriores y, menos aún, establecer entre ellas procesos de traducción. No obstante, podemos efectuarlos entre algunas representaciones de la correlación y ciertos modelos de la regresión.

Como modelos de actividades relacionadas con la regresión se han considerado los siguientes: i) Estimar y/o predecir un valor de Y; ii) dibujar la curva de regresión, y, iii) obtener la ecuación de la recta de regresión.

Las diversas transformaciones que podemos realizar, y que se encuentran plasmadas en la Tabla 4.2.2, se verán reflejadas en los dos problemas que hay en el cuestionario.

Tabla. 4.2.2. Ajuste de modelos de regresión a partir de diversas representaciones de la correlación

Hacia \ Desde	Estimar / predecir un valor de Y	Dibujar la curva de regresión	Obtener la ecuación de la recta de regresión
Tabla	**Tipo 13 Estimación y cálculo**	**Tipo 14 Estimación y trazado**	**Tipo 15 Cálculo**
Diagrama de dispersión	**Tipo 16 Estimación**	**Tipo 17 Estimación y trazado**	No añade nada al Tipo 15

Delimitación del contenido

Una vez definidos los objetivos y tomada la decisión de contemplar los tipos de tareas que acabamos de describir, se decidió completar el cuestionario de modo que quedase cubierta una parte importante de las propiedades y problemas relacionados con la correlación y regresión. Además de todo lo expuesto, hemos aplicado los resultados de la investigación de Sánchez Cobo (1.996), de la cual se presenta un resumen en la sección 3.2, para describir el significado institucional de la asociación estadística en un curso usual de Estadística descriptiva a nivel de bachillerato.

Como trabajo complementario hemos examinado, también, la programación del profesor y los apuntes de los alumnos del curso a los que se les administrará el cuestionario, y, de esta manera, poder observar qué elementos de significado de los detectados se desarrollan en la enseñanza universitaria, adecuándose la prueba a ellos. Tanto un resumen de la investigación de Sánchez Cobo (1.996), como los análisis de la programación del profesor y de los apuntes de dos alumnas se insertan en el Capítulo 3. Exponemos, con mayor profundidad, en el apartado 4.3 los contenidos que hemos considerado.

Redacción de los items

Una de las fases más importantes en la elaboración del cuestionario es la redacción de los items. Para ello, hemos tenido en consideración las recomendaciones indicadas por Fox (1.987) sobre las características que deben reunir, que son:

"1. Claridad de lenguaje. Esta característica significa que la intención de la pregunta y la naturaleza de la información que se busca están claras para el encuestado.

2. Concreción del contenido.

3. Unicidad de propósito. Con esta característica se intenta asegurar que cada pregunta busque un elemento o fragmento de información, y sólo uno.

4. Independencia de otros supuestos. Con esta característica se intenta conseguir la seguridad de que, para contestar a una pregunta, el encuestado no tiene que hacerlo a otra anterior que no se haya formulado.

5. Independencia de sugerencias. En la formulación de la pregunta no debe haber nada que, de alguna manera, sugiera al sujeto que se esperan obtener ciertas respuestas, o que algunas son más deseables o aceptables que otras.

6. Completitud lingüística y coherencia gramatical." (págs. 590 a 594)

Además de seguir estas recomendaciones, el cuestionario está estructurado en cuatro bloques:

1. Preguntas preliminares: Son aquéllas que permitirán analizar la muestra estudiada, y en las que se solicitan datos diversos: El nombre y apellidos de los alumnos, su edad, el curso y la titulación que están estudiando, la fecha de realización de la prueba. Otros aspectos considerados son el nivel de procedencia (B.U.P., F.P., etc.) y si habían estudiado en cursos anteriores nociones de estadística. Como Morris (1.997), deseamos conocer si los alumnos consideran que la asignatura de Estadística, y en particular el tópico de la Correlación y Regresión, tiene interés para ellos, pues en algunas encuestas realizadas los alumnos consideran que la estadística es una disciplina difícil e incluso inútil (Garfield y Ahlgren, 1.988; Shute y Gawlick-Grendell, 1.993).

2. Items de verdadero/falso: Son 12 items, que presentan entre 3 y 5 subitems cada uno de ellos. Para reducir la posibilidad de acierto aleatorio, los items tienen, en general, varias respuestas correctas. Tal salvedad fue informada a los alumnos instantes antes de administrarles la prueba.

3. Tareas de traducción: Son 6 tareas con 5 opciones cada una. Se tratarían, en terminología de Sax (1.989), de pruebas de ensayo, teniendo, de acuerdo con este autor, las siguientes ventajas: Libertad del sujeto para responder, pudiendo incluir sus argumentaciones de por qué da esa respuesta; evita, en un alto porcentaje, el azar en las respuestas; son favorecedoras del pensamiento divergente, y ofrece la posibilidad de encontrar respuestas que el investigador no ha considerado.

4. Problemas: Son 2 problemas, uno de ellos con 5 cuestiones.

En cuanto a las respuestas, el cuestionario admite tanto la forma abierta como cerrada. En aquellas pruebas en las cuales estimulamos a los alumnos a que argumenten su contestación, como por ejemplo las tareas de traducción y los problemas, éstas se formulan para permitir una respuesta abierta. En cambio, hay

otras pruebas de respuesta cerrada, como por ejemplo los items de opciones múltiples.

Una menor incidencia en el cuestionario tienen los reactivos con respuesta de escala, por ejemplo escala de Likert, con respuestas categóricas, por ejemplo con un "sí" o "no", o de elección de la mejor respuesta.

Asimismo, se ha tenido en cuenta no formular items de carácter negativo, evitando así que algún sujeto, por omisión, respondiera como si estuviera redactado en forma positiva. Hemos intentado que los items sean lo más cortos posibles, para facilitar su rápida lectura y una comprensión cómoda. Hemos redactado los items para que no impliquen sesgos que puedan desvirtuar o sugerir la respuesta (Scott, 1.989).

En algunos estudios, los juicios sobre correlación se presentan a través de variables descontextualizadas, tales como pares de números (Erlick y Mills, 1.967; Jennings, Amabile y Ross, 1.982). Mediante tal presentación, se trata de evitar que las expectativas previas de los sujetos influyan en sus juicios sobre la asociación existente entre las variables. Por otra parte, una persona puede creer que dos variables familiares y significativas no muestren correlación. En tal caso, se pueden generar escenarios que representarían una asociación positiva entre ellas (Bower y Masling, 1.978). Siempre que ha sido factible los reactivos se han presentado contextualizados. No olvidemos, que una sugerencia de la educación estadística es que debemos transmitir a los alumnos que la estadística trata no con números, sino con números en contexto (Moore, 1.990). Además, dado que el contenido significativo de las variables tiene una influencia notable sobre los juicios (Trolier y Hamilton, 1.986), estamos interesados en estudiar ese efecto.

Por otro lado, las tareas de traducción combinan las siguientes variables tarea y valores que describimos a continuación: i) Intensidad; ii) tipo de ajuste; iii) tipo de covariación; iv) teorías previas, y, v) signo de la correlación. La intensidad, en valor absoluto, se ha dividido en los intervalos [0,0'1), [0'1,0'35), [0'35,0'65), [0'65,0'9) y [0'9,1]. El tipo de ajuste tiene en cuenta tanto la lineal como la no lineal. En cuanto a los tipos de covariación se abarcan todos los mostrados en Barbancho (1.973): a) Dependencia causal unilateral; b) interdependencia; c) dependencia indirecta; d) concordancia, y, e) covariación casual.

Tabla 4.2.3. Diseño de las tareas

Tarea		Intensidad (en valor absoluto)					Tipo ajuste		Tipo de covariación					Teorías previas		Signo	
Tipo	nº	0 0'1	0'1 0'35	0'35 0'65	0'65 0'9	0'9 1	lineal	no lineal	depe. c.uni.	inter-depe.	depe. indi.	con-cord.	cov. casu.	coin-cide	no coin.	+	-
1	1a				-0'85			X		X							-
	1b					0'98		X	X							+	
	1c			0'6			X					X				+	
	1d		-0'3				X				X						-
	1e	0'09						X					X			no se aplica	
2	2a					0'92	X			X						+	
	2b		-0'12				X					X					-
	2c	0'01						X					X			+	
	2d			0'52				X	X							+	
	2e				-0'87		X				X						-
3	3a					-0'98		X	X					X			-
	3b	-0'09						X		X					X		-
	3c			0'42			X						X		X	+	
	3d				0'84		X				X			X		+	
	3e		0'32				X					X		X		+	
4	4a	0'1					X			X					X	+	
	4b			0'53				X				X		X		+	
	4c					0'93	X		X						X	+	
	4d		-0'22				X						X	X			-
	4e				-0'69			X			X			X			-
5	5a					1										+	
	5b		-0'3														-
	5c	0'05														+	
	5d				-0'8												-
	5e			0'5												+	
6	6a		0'25													+	
	6b					-1											-
	6c	-0'01															-
	6d				0'7											+	
	6e			-0'4													-

Se muestran tanto situaciones que coinciden con las teorías previas como otras en las que no hay esa coincidencia. También se presentan correlaciones de ambos signos. Todo ello, y como pone de manifiesto la Tabla 4.2.3, se ha conjugado en el diseño de las actividades para que haya de todos los tipos de variables de tarea en ellas.

Diseño del formato

Si deseamos que los sujetos se impliquen y contesten con corrección el cuestionario, debemos dotarlo de un diseño que lo haga atractivo. Con tal fin se ha utilizado un tipo de letra que sea fácilmente legible, así como una impresión lo más nítida posible. Se ha buscado una organización que sea facilitadora para las respuestas, donde se han agrupado los distintos items en los bloques, claramente identificables, que anteriormente reseñábamos: i) Preguntas preliminares, ii) items de opciones múltiples, iii) tareas de traducción, y, iv) problemas.

En el momento de la recogida de los datos se utilizó la interacción mixta (Fox, 1.987), dándoseles a los alumnos una serie de orientaciones para la correcta resolución de la prueba. Asimismo, al inicio de cada uno de estos bloques se presentaban las instrucciones, claras y breves, para que el sujeto pudiera contestar con garantías. Por ejemplo, se indicaba si estaban permitidos los cálculo numéricos o no.

Se ha cuidado también que las preguntas tengan espacio suficiente para ser contestadas, así como que no se encuentren en dos páginas distintas, es decir, al final de una página y al principio de la siguiente, lo cual dificultaría su lectura. Además, el cuestionario se ha entregado grapado, lo que evita las posibles pérdidas de respuestas.

Puesta a prueba del cuestionario

Una vez terminada la construcción del instrumento se ha utilizado una muestra piloto, compuesta por 6 alumnos de la Diplomatura en Estadística, con la cual se pretendía obtener información sobre la comprensión y legibilidad de la prueba. Para ello, cuando los estudiantes completaron el cuestionario, se les realizó una entrevista, en la que se les pidió que indicaran aquellos términos que no estuvieran claros o que consideraran que eran ambiguos. Además, se evaluó el tiempo medio que era necesario para completar el cuestionario, con el objetivo de obtener una estimación de cuál debía ser el tiempo del que dispondrían los sujetos de la muestra para responder. Agradecemos aquí la colaboración a todos los alumnos, tanto de la muestra piloto como de la definitiva, por el interés que pusieron en colaborar con nuestro trabajo.

4.3. CONTENIDOS INCLUIDOS

Como consecuencia del trabajo sobre los libros de texto llevado a cabo en la Memoria de Tercer Ciclo (Sánchez Cobo, 1.996) y del estudio de los apuntes, tanto del profesor como de los tomados en clase por un par de alumnas, llegamos a la identificación del contenido preciso de la enseñanza efectuada a estos alumnos sobre el tema. En particular, obtuvimos la enumeración de elementos del significado de los conceptos de asociación y regresión que ofrecemos a continuación:

1. Dependencia funcional y dependencia aleatoria
    1.1. Asociación o covariación
    1.2. Dependencia funcional y aleatoria
    1.3. Independencia aleatoria
    1.4. Tipos de covariación (Barbancho, 1.973):
        1.4.1. Dependencia causal unilateral

1.4.2. Interdependencia

1.4.3. Dependencia indirecta

1.4.4. Concordancia

1.4.5. Covariación casual

2. Distribuciones dobles

2.1. Variables estadísticas unidimensionales

2.2. Variables estadísticas bidimensionales

2.3. Distribución conjunta

2.4. Diagrama de dispersión:

2.4.1. Lectura de los puntos de un diagrama de dispersión

2.4.2. Significado de un punto en un diagrama de dispersión

2.5. Tablas de frecuencias dobles o de contingencia:

2.5.1. Construcción de una tabla de contingencia

2.5.2. Lectura de una tabla de contingencia:

- Frecuencias absolutas
- Frecuencias marginales absolutas y relativas
- Frecuencias relativas en cada casilla
- Frecuencias condicionales por filas y columnas

2.5.3. Comparación de tablas de contingencia para detectar asociación perfecta, parcial e independencia de variables

2.6. Momentos dobles: Ordinarios y centrales

2.7. Distribuciones marginales

2.8. Momentos marginales

2.9. Distribuciones condicionales

2.10. Momentos condicionales

3. Covarianza

3.1. Covarianza o momento mixto de segundo orden

3.2. Cálculo en distribución estadística conjunta:

$$\text{cov}\,(X,Y) = \frac{1}{N}\sum_{i}\sum_{j}(x_i - \bar{x})(y_j - \bar{y})f_{ij}$$

3.3. Cálculo a partir de dos variables estadísticas unidimensionales:

$$\text{cov}(X,Y) = \frac{1}{N}\sum (x_i - \bar{x})(y_j - \bar{y})$$

3.4. Expresión simplificada:

$$\text{cov}(X,Y) = \frac{1}{N}\sum x_i\ y_j\ n_i - \overline{xy}$$

3.5. Simbología: $\text{cov}(X,Y) = E(x, y) = \sigma_{xy}$

3.6. El signo de la covarianza indica el tipo de asociación (directa o inversa)

3.7. La $\text{cov}(X,Y) = 0 \Rightarrow X$ e Y son independientes

3.8. La $\text{cov}(X,Y)$ no es independiente de las unidades de medida

4. Correlación

4.1. Concepto de correlación:

4.1.1. Estudio de la dependencia o relación entre variables

4.1.2. Medida de la intensidad de la relación

4.2. Definiciones del coeficiente de correlación

4.2.1. $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

4.2.2. $r = \sqrt{b_{xy} b_{yx}}$, donde b_{xy} y b_{yx} son las pendientes de las rectas de regresión

4.2.3. $r = \frac{\sum_{i=1}^{n} x_i\ y_j}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{j=1}^{n} y_j^2}}$ donde $x_i = x_i - \bar{x}$, $y_i = y_i - \bar{y}$

4.2.4. $r = \frac{\sum x_i\ y_j - N\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - N\bar{x}^2}\ \sqrt{\sum y_j^2 - N\bar{y}^2}}$

4.3. El coeficiente de correlación *r* es adimensional

4.4. Relación existente entre el signo de *r* y la pendiente de la recta de regresión

4.5. Relación existente entre el signo de *r* y el signo de la $\text{cov}(X,Y)$

4.6. Signo del coeficiente de correlación a partir del diagrama de dispersión

4.7. Variación de las variables a partir del signo del coeficiente de correlación

4.8. Relación entre la dispersión de datos y la correlación

5. Propiedades e interpretación del coeficiente de correlación

5.1. Aquéllas que hacen referencia a alguna característica de la correlación:

5.1.1. Si $0 < r < 1$ la correlación es positiva o directa

5.1.2. Si $-1 < r < 0$ la correlación es negativa o inversa

5.1.3. Si $0 < |r| < 1$ las variables X e Y están en dependencia aleatoria

5.1.4. Si $0 < |r| < 1$ la correlación es más intensa (débil) cuando r tiende a 1 (r tiende a 0)

5.1.5. Si $r = 0$ no hay ningún tipo de dependencia entre X e Y

5.1.6. Si los puntos del diagrama de dispersión se encuentran más densamente distribuidos en los cuadrantes 1º y 3º, la relación es positiva

5.2. Aquéllas que hacen referencia a alguna característica de la regresión:

5.2.1. Si $r = 1$ los puntos de la variable estadística bidimensional (X,Y) están sobre la recta de regresión

5.2.2. Si $r = 0$ las rectas de regresión son perpendiculares

5.2.3. El grado de correlación se corresponde con el ángulo que forman ambas rectas de regresión, siendo tanto mayor cuanto menor sea este ángulo

5.3. Aquéllas que aluden a alguna cualidad del coeficiente de correlación:

5.3.1. El valor del coeficiente de correlación r verifica: $-1 \leq r \leq 1$

5.4. Influencia de los valores atípicos en el coeficiente de correlación

5.5. Relación entre el coeficiente de correlación de una población y el coeficiente de correlación de sus submuestras

5.6. El coeficiente de correlación mide la relación lineal de dos variables

6. Correlación y causalidad

6.1. La correlación entre las variables es simétrica

6.2. Distinción entre correlación y causalidad

7. Coeficiente de determinación y su interpretación

7.1. El coeficiente de determinación nos indica la proporción de varianza explicada por la regresión

7.2. Relación entre la varianza de Y, la varianza residual y el coeficiente de determinación

8. Regresión

8.1. Definición del concepto de regresión

8.1.1. Técnica matemática que ajusta una función a un conjunto de datos

8.1.2. Sirve para predecir o estimar los valores de una variable en función de la otra

8.2. Regresión lineal

9. Método de los mínimos cuadrados

9.1. Este criterio trata de elegir entre todas las rectas aquélla tal que la suma de los cuadrados de las desviaciones verticales de los puntos a la recta sea mínima

9.2. Definición de la recta de regresión: La recta de regresión de Y sobre X minimiza el error cuadrático medio cometido al mantener en cada punto fija la variable X y sustituir el valor y por el correspondiente sobre la recta

9.3. Interpretación de los coeficientes a y b_{yx}

$$\hat{y} = a + b_{yx}\, x,\ a = y - b_{yx}\, x,\ b_{yx} = r\,\frac{\sigma_y}{\sigma_x}$$

9.4. Bondad del ajuste

9.5. Otros métodos

9.6. Utilización de la ecuación de regresión lineal para realizar predicciones

10. Rectas de regresión e interpretación de sus parámetros

10.1. Las rectas de regresión pasa por el centro de gravedad $(\bar{x}, \bar{y})$

10.2. Ecuación reducida de la recta de regresión

10.3. Cálculo abreviado de la recta de regresión

10.4. Coeficiente de regresión: $b_{yx} = r\,\frac{\sigma_y}{\sigma_x}$

10.4.1. b_{yx} tiene el mismo signo que r

10.4.2. La dependencia directa implica que $b_{yx} > 0$. La dependencia inversa implica que $b_{yx} < 0$. La independencia implica que $b_{yx} = 0$

10.4.3. Dados tres de los coeficientes r, b_{yx}, σ_y, σ_x podemos determinar el cuarto

10.5. Relación entre el valor absoluto del coeficiente de correlación y la predicción de Y a partir de X

11. Rectas de regresión

11.1. Recta de regresión de X sobre Y

11.2 Distinción entre ambas rectas de regresión

La enumeración anterior muestra la gran riqueza de conceptos y de propiedades asociadas a la correlación y la regresión. En la presente investigación no estamos interesados en todos los contenidos anteriormente expuestos.

Tabla 4.3.1. Contenidos sobre la regresión

	ÍTEM	TAREA	PROBLEMA
Concepto de regresión			P1 / P2
Regresión lineal y no lineal	3,6,7		
Método de los mínimos cuadrados			P1 / P2
Interpretación de los coeficientes a y b_{yx}	3		P1
Predicciones a partir ecuación recta regresión	9		P2
Distinción entre var. depend. e independiente	10		P1 / P2
Recta regresión pasa por centro gravedad			P1
Relación r y predicción de Y a partir de X			P2
Coeficiente de regresión b_{yx}			P1 / P2
Recta de regresión de X sobre Y	12		P2
Distinción entre ambas rectas regresión	12		P2
Normalmente sólo nos interesa una recta reg			P2
Relación grado relación y ángulo rectas reg.	6,11,12		

No hemos tomado en consideración: La variable estadística unidimensional; distribución conjunta; frecuencias dobles; tablas de frecuencias dobles o de contingencia y frecuencias asociadas; momentos; simbología para la covarianza; la

covarianza no es independiente de las unidades de medida; relación existente entre el signo de r y el de $\sum_{i=1}^{n} x_i y_i$; influencia de los valores atípicos en el coeficiente de correlación; relación entre el coeficiente de correlación de una población y el coeficiente de correlación de sus submuestras; relación del coeficiente de correlación con el número de casos N; otras propiedades del coeficiente de correlación y determinación.

Tabla 4.3.2. Contenidos sobre la correlación

	ÍTEM	TAREA	PROBLEMA
Dependencia funcional y aleatoria		1,2,3,4,5,6	P2
Independencia aleatoria		1,2,3,4,5,6	
Tipos de covariación		1,2,3,4,5,6	
Concepto de covarianza	1		
Cálculo de la covarianza			P2
Signo cov (X,Y) indica tipo asociación	1		
Relación signo cov(X,Y) y pendiente recta regresión	1		
r = 0 implica independencia variables	7		
Concepto de correlación		1,2,3,4,5,6	P2
Dispersión datos y correlación	3,7		
Cálculo del coeficiente de correlación			P2
r es adimensional	2	1,2,3,4,5,6	
Relación signo r y pendiente recta regresión	3,11	1,2,3,4,5,6	
Relación signo r y signo cov (X,Y)	1,7		
Relación diagrama de dispersión y r		1,2,3,4,5,6	
A partir signo de r variación variables	4	1,2,3,4,5,6	
A partir tabla estimar correlación existente		1,2,3,4,5,6	
Propiedades r referidas a característica correlación	5,8	1,2,3,4,5,6	P2
Propiedades r referidos a característica regresión	7	1,2,3,4,5,6	
Propiedades r referidos cualidad coef. correlación		1,2,3,4,5,6	P2
Significado tamaño r		1,2,3,4,5,6	P2
Magnitud en que se mide r	2		
Correlación entre variables es simétrica		1,2,3,4,5,6	
Distinción entre correlación y causalidad	8,9,11		

Con el fin de evitar que algún concepto de los presentados en el cuestionario pudiera no tener reflejo en la enseñanza universitaria, hemos cruzado los contenidos que se han considerado en la presente investigación con los incluidos en la progamación del profesor, habiéndose observado que todos ellos se encuentran comprendidos en la misma.

De igual modo, se ha verificado que el cuestionario recogiera la totalidad de los contenidos que conforman el núcleo de este trabajo. Las Tablas 4.3.1 y 4.3.2 muestran los contenidos evaluados, así como en qué bloque y prueba del cuestionario se ven reflejadas dichas nociones. A estos contenidos puede remitirse la validez del cuestionario. Creemos que otras investigaciones pueden continuar nuestro trabajo, analizando el resto de los contenidos que acabamos de enumerar.

4.4. ANÁLISIS DE LOS ITEMS DE OPCIONES MÚLTIPLES

En esta sección realizamos un análisis en profundidad tanto de los doce items de respuesta múltiple insertados en el cuestionario, como de los distractores de cada uno de ellos y los errores plausibles que, de forma apriorística, suponemos pueden aparecer en las respuestas de los sujetos de la muestra.

Ítem 1: Signo de la covarianza y tipo de asociación

	1. Si la covarianza de las variables X e Y es mayor que 0, las variables X e Y presentan
X	a. Correlación positiva
X	b. La regresión podría ser no lineal
	c. Las variables podrían estar no correlacionadas
X	d. La pendiente de la recta de regresión tiene un signo positivo
X	e. El coeficiente de correlación es positivo

Este ítem, con el presente enunciado, no lo hemos encontrado en la literatura de investigación, dado que no conocemos ningún estudio que se haya efectuado sobre la comprensión de la noción de covarianza. En él, fundamentalmente, se alude a la covarianza y relación del signo de ésta con el tipo

de correlación, pendiente de la recta de regresión y con el signo del coeficiente de correlación *r*.

En la investigación de Sánchez Cobo (1.996) se puso de manifiesto que, a pesar de ser un concepto importante dentro de este tópico, quedaba relegado a un papel de mera noción propedéutica de otra, el coeficiente de correlación *r*, que se considera de más relevancia. Además, ninguno de los textos analizados que incluyen el estudio de la covarianza relaciona el signo de ésta con la dependencia existente.

Queremos evaluar si, después de la enseñanza que han recibido los sujetos de la muestra, los alumnos relacionan el signo de la covarianza con el signo del coeficiente de correlación y si son sensibles a que, aunque obtengamos un valor para la covarianza, la regresión puede no ser lineal.

Ítem 2: El coeficiente de correlación es adimensional

2. Juan correlaciona alturas y pesos de los estudiantes varones de 1º de la Diplomatura en Empresariales utilizando como unidades el metro y el kilogramo. Ángela registra las alturas y los pesos empleando centímetros y gramos como medida. Ambos calculan la correlación entre sus dos conjuntos de medidas.

a. La correlación de Ángela será mayor que la de Juan

X b. Los dos coeficientes de correlación serán aproximadamente iguales

c. El coeficiente de Juan tenderá a ser mayor que el de Ángela

X d. El valor del coeficiente depende de la dispersión de los datos

El ítem 2 lo hemos elaborado a partir de la transformación pertinente de uno de Cruise, Dudley y Thayer (1.984, pág. 261). Un hecho bien conocido es que la covarianza no es independiente de las unidades de medida, lo que comporta la búsqueda de otro parámetro estadístico que, sirviendo para indicarnos la relación existente entre las características de una variable estadística bidimensional, goce de esta propiedad. Otro aspecto que deseamos estimar es si los estudiantes son conscientes de la influencia de la dispersión de los datos sobre el valor del coeficiente de correlación, cuestión ésta poco destacada en la literatura de investigación.

Ítem 3: Relación entre la intensidad y el diagrama de dispersión

	3. Cuando la intensidad de la relación entre dos variables decrece
	a. La pendiente de la recta de regresión de Y sobre X crece
	b. La pendiente de la recta de regresión de X sobre Y crece
X	c. Hay mayor dispersión en la nube de puntos
	d. La covarianza aumenta de valor absoluto

Este ítem se ha obtenido modificando, de forma conveniente, uno de Cruise, Dudley y Thayer (1.984, pág. 303). La respuesta que es correcta es la c), porque cuando disminuye la intensidad de la relación aumenta la dispersión de la nube de puntos. La respuesta a) podría ser cierta sólo si la relación entre las variables fuera inversa, lo mismo ocurre con la respuesta b).

Ítem 4: Variación de las variables a partir del signo de *r*

	4. Si las dos variables están correlacionadas positivamente
X	a. Cuando una aumenta, la otra también aumenta
	b. Cuando una disminuye, la otra aumenta
X	c. Cuando una disminuye, la otra disminuye
	d. La relación entre las dos variables es de tipo lineal

También este ítem se ha elaborado a partir de uno de Cruise, Dudley y Thayer (1.984, pág. 258). Se persigue observar si los estudiantes conocen un criterio para determinar si dos variables correlacionan en forma directa. Todavía más, del conocimiento de como se comportan dos variables que muestran una correlación positiva podemos deducir una estrategia de estimación del signo de la dependencia y que podrá utilizar en el bloque del cuestionario que está dedicado a las tareas de interpretación. Las respuestas correctas son la a) y la c), no obstante algunos alumnos podrían considerar sólo el crecimiento conjunto de las variables y no el decrecimiento conjunto - c) -. El distractor d) está incluido para evaluar si los alumnos son conscientes de que un coeficiente de correlación positivo podría implicar una relación de tipo no lineal.

El ítem 5 se ha tomado de la investigación de Morris (1.997). Con anterioridad, la investigación de Estepa (1.994) puso de manifiesto que algunos alumnos juzgan que un coeficiente de correlación cuyo valor sea de -0'1 indica una relación más intensa que otro que valga -0'9, ya que -0'1 es mayor que -0'9, dado que se produce una inversión del orden (González y cols., 1.990). Esto puede estar motivado porque los estudiantes exhiben una concepción unidireccional (Estepa, 1.994), confundiendo la dependencia inversa y la no dependencia.

Ítem 5: Propiedades e interpretación de *r* que hacen referencia a alguna característica de la correlación

5. Ordene los siguientes valores según expresen mayor correlación entre dos variables: 0'5, - 0'8, 0'2, - 0'4,0.

_____ mayor valor de la correlación

_____

_____

_____

_____ no existe correlación

Así, un sujeto de la investigación realizada por Morris, después de responder de forma correcta de que el mayor valor de la correlación corresponde a -0'8, indica: *"No comprendo"* (pág. 20). Por lo tanto, parece que ciertos alumnos *"tienen dificultades con el concepto de correlación. En particular, algunos estudiantes no tienen una idea clara de la correlación inversa y de la no dependencia"* (Morris, 1.997, pág. 20). De acuerdo con Morris (1.997), los errores que consideramos, a priori, que serían susceptibles de aflorar son: i) El de efectuar la clasificación de los valores del coeficiente de correlación según el orden de los números reales, es decir, 0'5, 0'2, 0, -0'4, -0'8, y, ii) considerar que el valor 0 indica que no existe correlación, y clasificar el resto de los valores según el orden de los números reales, es decir, 0'5, 0'2, -0'4, -0'8, 0.

Al igual que el ítem 1, no hemos encontrado otro similar al ítem 6 en la literatura de investigación. Trata de determinar si los estudiantes tienen aprehendido que ciertos valores del coeficiente de correlación nos ofrecen

información acerca de las rectas de regresión, así como la relación existente entre el grado de dependencia que hay entre las variables y el ángulo que forman ambas rectas de regresión. Esto es especialmente significativo si tenemos en cuenta que es una estrategia básica para estimar el coeficiente de correlación a partir de un diagrama de dispersión, actividad esta a la que se enfrentarán en la tarea 4. Por ello, sería conveniente tener en cuenta los resultados obtenidos en este ítem a la hora de analizar dicha tarea. Además, se incluyen la conexión entre la covarianza y el coeficiente de correlación *r* y se hace hincapié, una vez más, en la posibilidad de regresión no lineal.

Ítem 6: Propiedades e interpretación de *r* que hacen referencia a alguna característica de la regresión

	6. Si el coeficiente de correlación entre dos variables es nulo
	a. Ambas rectas de regresión de Y sobre X y de X sobre Y son paralelas
X	b. La covarianza también es nula
	c. Ambas rectas de regresión de Y sobre X y de X sobre Y coinciden
X	d. Las variables pueden tener una relación no lineal
X	e. Ambas rectas de regresión de Y sobre X y de X sobre Y son perpendiculares

El ítem 7 es una modificación del correspondiente de Cruise, Dudley y Thayer (1.984, pág. 260). Los elementos tratados en este ítem son la relación del coeficiente de correlación y el tipo de asociación y la interpretación de *r* como un porcentaje de la varianza.

Ítem 7: Algunas propiedades de *r*

	7. Si r es el coeficiente de correlación de dos variables, indique qué afirmaciones son correctas
X	a. $r = 0$ indica que las variables son independientes
	b. Si $r = 0'6$ la correlación entre las variables X e Y es doble que cuando $r = 0'3$
X	c. Una relación funcional entre variables se corresponde con un valor de r de +1 ó -1
X	d. El coeficiente de correlación puede interpretarse como un porcentaje de la varianza

Es evidente que si entre dos características de una variable bidimensional hay una relación de causa efecto esto supone que, también, haya correlación entre ambas. Sin embargo, la implicación contraria no es, en general, cierta, pues *"aunque la correlación es una característica necesaria de una relación causal, no es suficiente para probar que tal relación existe"* (Phillips, 1.992, pág. 143). La investigación de Estepa (1.994) puso de manifiesto que ciertos alumnos confundían la correlación con la causalidad. A esta concepción errónea la denominó concepción causalista.

Ítem 8: Distinción entre correlación y causalidad

	8. Al estudiar las superficies sembradas de trigo en miles de hectáreas y las cosechas obtenidas en millones de quintales métricos, en cinco años consecutivos, el coeficiente de correlación obtenido fue 0'91. Luego
X	a. Podría haber otros factores que hagan variar los resultados
	b. Deberíamos tomar una muestra más grande para poder expresar la relación entre la superficie plantada y la cosecha obtenida
X	c. La cosecha obtenida presenta una alta correlación con la superficie plantada
	d. Si plantamos doble superficie, obtendremos con seguridad doble cosecha

La investigación de Morris (1.997) confirma este extremo, indicando que, al preguntarles a los estudiantes sobre si de la relación entre dos variables podría concluirse siempre una relación causal, un 25 % de los encuestados responden afirmativamente. Todo lo anterior induce a pensar que los alumnos tienen una tendencia hacia el pensamiento determinista, lo cual refleja su ausencia de consciencia o comprensión de la variación (Pfannkuch y Brown, 1.996). Un error que es factible que surja es que, dado que la correlación presentada es fuerte (el coeficiente de correlación es de 0'91), los estudiantes juzguen que la superficie plantada es la causa de la cosecha obtenida, no tomando en consideración que puede haber otras variables que influyan en este proceso (Phillips, 1.992), lo que indicaría que poseen un esquema muy restringido de tipos de covariación, dado que, únicamente, prestan atención a la dependencia causal unilateral, pero no a la dependencia indirecta (Barbancho, 1.973).

Ítem 9: Tipos de covariación

9. ¿En qué predicción tendría más confianza?

a. La predicción de la estatura de un hombre a partir de su peso

b. La predicción del peso de un hombre a partir de su estatura

X c. Las dos me dan la misma confianza

Este ítem está tomado de Tversky y Kahneman (1.982a). Una diferencia significativa entre correlación y causalidad es que mientras en la correlación la relación entre las variables es simétrica, ya que el coeficiente de correlación es invariante ante una permutación de las variables, en la causalidad esta relación es asimétrica, existe una variable causa y una variable efecto. A la hora de realizar una inferencia es más fácil llevarla a cabo si el esquema es causa-efecto que si, por el contrario, es efecto-causa. De esta manera, se considera más aceptable explicar la estatura de un niño a partir de la de su padre que no al revés.

Esto podría ocurrir, a pesar de que ambas variables no deban verse como causa una de la otra, *"siempre que la primera variable parezca explicar a la segunda mejor que la explicación que la segunda ofrece de la primera"* (Tversky y Kahneman, 1.982a, pág. 120). Por consiguiente, una concepción errónea que puede surgir es que, aunque el peso y la altura de una persona no sean causa una de la otra, los sujetos estimen más fiable explicar el peso de una persona a partir de su estatura.

Ítem 10: Distinción entre variable dependiente e independiente

10. Las rentas se usan para predecir los ahorros, ambos medidos en miles de pesetas. Para la ecuación de regresión y = 1000 + 0'1x, ¿cuál de las siguientes afirmaciones es verdadera?

a. Y es la renta, X es el ahorro, la renta es la variable independiente

b. Y es la renta, X es el ahorro, el ahorro es la variable independiente

c. Y es el ahorro, X es la renta, el ahorro es la variable independiente

X d. Y es el ahorro, X es la renta, la renta es la variable independiente

Este ítem está tomado de Cruise, Dudley y Thayer (1.984, pág. 285). La distinción entre la variable independiente y la variable dependiente es esencial para

el concepto de regresión y para la distinción entre ambas rectas de regresión. Por ello, deberemos cruzar los resultados de este ítem con los del Problema 2 en sus apartados c), d) y e), especialmente con este último.

Estimamos que el sesgo notable que las actividades matemáticas, de marcado carácter funcional, que el alumno ha realizado a lo largo de su dilatada experiencia escolar, puede inducirlo a considerar sólo una alternativa: x variable independiente, y variable dependiente. Además, este ítem estimula el que emerjan los distintos tipos de covariación (Barbancho, 1.973) que los alumnos poseen.

Ítem 11: Relación entre *r* y el ángulo de las dos rectas de regresión

11. ¿Qué valor ha de tener *r* si las dos rectas de regresión tienen una pendiente idéntica?

a. 0
X b. 1
X c. -1
d. 0'5

Este ítem es una modificación de otro de Cruise, Dudley y Thayer (1.984, pág. 304) y de Morris (1.997, pág. 17). En él se trata de hacer evidente tanto la distinción entre ambas rectas de regresión, X sobre Y e Y sobre X, como la relación existente entre el ángulo formado por las dos rectas de regresión y la fuerza de la dependencia que hay entre las variables. Este ítem es una profundización de la cuestión 7 planteada por Morris (1.997), donde únicamente se les solicitaba a los alumnos por el valor que indica una correlación positiva perfecta. Una concepción errónea típica, que posiblemente surja con este ítem 11, sería la de ofrecer como única respuesta el valor 1, no tomando en consideración el valor -1. Esto puede ser debido a una concepción unidireccional (Estepa, 1.994), confundiendo la dependencia inversa con la independencia (Morris, 1.997).

Ítem 12: Relación entre el ángulo de las dos rectas de regresión y el grado de relación

12. Si la correlación entre X e Y es perfecta, el ángulo que forman las rectas de regresión es de

a. 120º

b. 90º

c. 45º

X d. 0º

Este ítem es una modificación del correspondiente de Cruise, Dudley y Thayer (1.984, pág. 304). En este ítem intervienen elementos importantes como son la distinción entre las dos rectas de regresión y la relación entre el ángulo formado por las dos rectas de regresión y el coeficiente de correlación *r*.

Por otra parte, dado que no existe proporcionalidad entre el ángulo formado por las rectas y el coeficiente de correlación *r*, esto equivale a que el coeficiente de correlación *r* no es proporcional. Los ángulos que se han utilizado como distractores son de amplitudes habituales para los estudiantes.

4.5. TAREAS DE TRADUCCIÓN

En esta sección analizaremos las seis tareas de traducción que contiene el instrumento de evaluación y cuya finalidad es la examinar las concepciones que los sujetos muestran en las actividades de estimación, predicción e interpretación, y sobre las cuales, como mostraba la investigación de Sánchez Cobo (1.996), existía un notable déficit en los ejercicios incluidos en los libros de texto de secundaria, y, teniendo en cuenta que los profesores consideran a los manuales *"el paradigma del conocimiento que hay que trasmitir"* (Rico, 1.990, p.22), creemos que dicho sesgo también caracterizaría la enseñanza recibida por los estudiantes a su paso por este nivel.

Por otro lado, este tipo de tareas apenas se contemplan en la enseñanza universitaria, pues se supone una habilidad ya adquirida por el alumno. A continuación realizamos el análisis.

La Tarea 1 solicita al estudiante que traduzca de una parejas de variables (descripción verbal) a un diagrama de dispersión que responda de forma coherente a las mismas. Esta prueba no la hemos encontrado, con el enunciado presente, en la literatura de investigación. El alumno debería hacer la Tarea 1 -véase Tabla 4.2.1-, es decir, traducir de descripción verbal a tabla, y realizar a continuación la representación gráfica.

Otra posibilidad es hacer un diagrama de dispersión aproximado sin una verdadera estimación de los valores particulares de los puntos. Juzgamos que un potencial heurístico de los estudiantes será valorar, en primer lugar, el tipo de correlación existente y con posterioridad dibujar un diagrama de dispersión aproximado. En tal caso, podemos obtener información sobre las estrategias que ellos usan para asociar una determinada nube de puntos a la dependencia intuida.

Es significativo comparar los resultados obtenidos en esta prueba con los de la Tarea 2. Se puede esperar que los alumnos se vean influidos por sus teorías previas - por ejemplo, sí las poseen con respecto a la pareja de variables número de pintores y tiempo que dura el trabajo, pero no la tienen o poseen poca información sobre longitud de una lista de números y tiempo en memorizarla -, siendo probable que se produzca una situación de correlación ilusoria (Chapman y Chapman, 1.969; Tversky y Kahneman, 1.982b; Murphy y Medin, 1.985; Jong, Merckelbach y Arntz, 1.995; Tomarken, Sutton y Mineka, 1.995). Se abarcan todos los intervalos de intensidad, en especial el caso de correlación alta corresponde a una dependencia funcional - apartado b) -, siendo la situación que hemos utilizado conocida en la literatura científica como *curva de aprendizaje* (Gimeno, 1.985; Goldstein, Lay y Schneider, 1.990; Hoffmann y Bradley, 1.994).

Tarea 1. Dadas las siguientes parejas de variables, dibuje un diagrama de dispersión que contenga 10 puntos y que muestre razonablemente su variación conjunta.

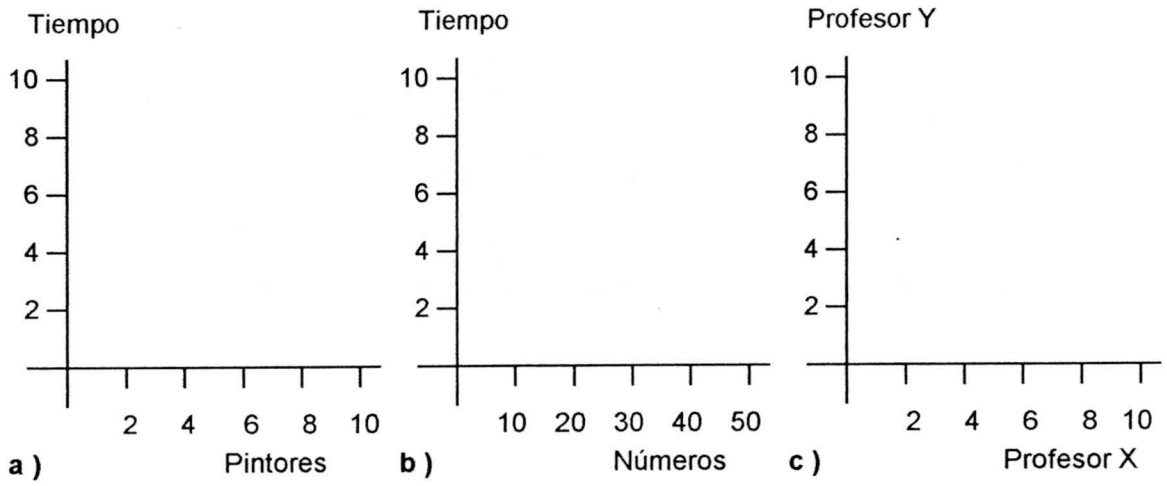
a) Número de pintores pintando una habitación y tiempo en horas para acabar el trabajo.

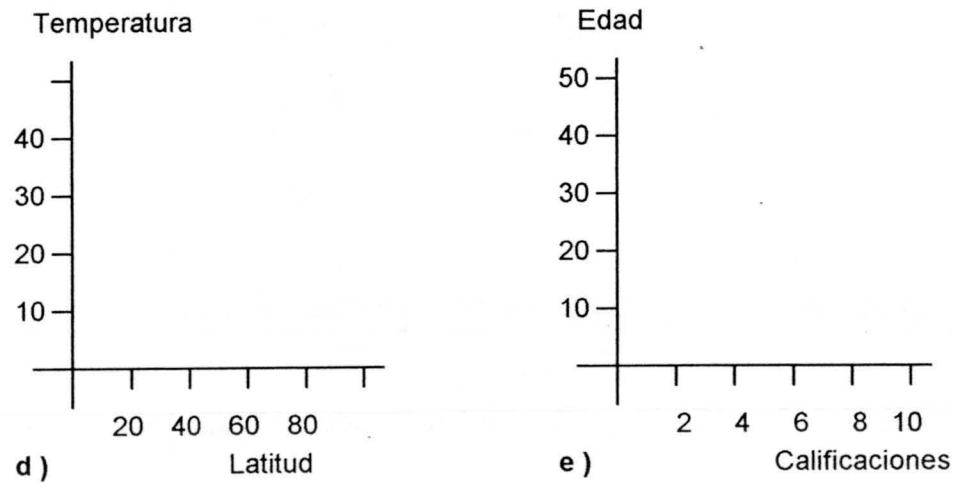
b) Longitud de una lista de números y tiempo empleado por una persona en memorizarla.

c) Calificaciones de un mismo examen por dos profesores de un tribunal de oposiciones.

d) La latitud de capitales europeas y la temperatura que hace en ellas un día determinado.

e) Las calificaciones en Estadística y la edad del alumno.





En la Tabla 4.5.1 se presentan las diferentes variables contempladas en la Tarea 1. Los valores de la intensidad en todas las tareas han sido estimadas a partir de conjuntos de datos reales.

Tabla 4.5.1. Valores de las variables en los apartados de la Tarea 1

VARIABLES	Tarea 1: Descripción verbal ¬> Estimación ¬> Gráfica				
	a)	b)	c)	d)	e)
1.1.Intensidad	-0'85	0'98	0'6	-0'3	0'09
1.2.Tipo ajuste	no lineal	no lineal	lineal	lineal	no lineal
1.3.Tipo covariación	interdep	dep.c.uni	concord	dep.ind	cov.casual
1.4.Teorías previas					
1.5.Signo	-	+	+	-	no se aplica

La Tarea 2 es de la misma clase que la usada por Jennings, Amabile y Ross (1.982), quienes la utilizaron para contrastar la estimación que se hace del coeficiente de correlación a partir de la descripción de pares de variables y a partir de datos representados en forma numérica (pares de puntos). Estos autores suponían que la descripción verbal activa las teorías previas y esto puede influir en la estimación de la correlación. Esta prueba es una modificación obtenida a partir de algunas situaciones de las empleadas por los investigadores antes citados, junto a otros distractores de contextos más próximos al alumnado español (Vizmanos y Anzola (1.988) y otro de elaboración propia). Puede servir para valorar la fuerza relativa de las teorías previas y los datos empíricos sobre la estimación del coeficiente de correlación.

A diferencia del trabajo de Jennings, Amabile y Ross (1.982), en la presente investigación se le pide el valor numérico del coeficiente de correlación, mientras en dicho estudio se les solicitaba que lo marcaran en una escala lineal de 0 a 100. Estimamos que este procedimiento es más complejo para el resolutor, pues lleva aparejado el plus de dificultad de la utilización de números reales y su representación en la recta real, y para el investigador es de difícil evaluación.

Tarea 2. Dadas las siguientes parejas de variables, indique un valor razonable para el coeficiente de correlación que expresaría el tipo de relación entre las variables (directa, inversa o independencia) y su intensidad (fuerte o débil).

a) Altura y envergadura - distancia entre los extremos de los brazos puestos en cruz - de los estudiantes de la Diplomatura en Empresariales.
r =

b) Ordenación por grado de timidez de los estudiantes de la Universidad de Jaén y ordenación por número de ciudades diferentes que han visitado.
r =

c) Grado de ambición y estatura de los estudiantes de la Universidad de Jaén.
r =

d) Tiempo semanal dedicado por los estudiantes a actividades atléticas y la posición que obtienen en una prueba de rendimiento físico.
r =

e) Número de días de lluvia y número de horas de sol registradas durante un año por un observatorio de las diferentes comunidades autónomas.
r =

En la Tabla 4.5.2 se presentan las diferentes variables contempladas en la Tarea 2. Los valores correspondientes del coeficiente de correlación se han obtenido de conjuntos de datos reales sobre las variables presentadas: Datos tomados sobre alumnos de Bachillerato - a), d) -, datos aportados en la investigación de Jennings y cols. (1.982), aunque cambiando el contexto - b), c) - y datos de anuarios estadísticos - e) -.

Tabla 4.5.2. Valores de las variables en los apartados de la Tarea 2

VARIABLES	Tarea 2: Descripción verbal¬>Estimación¬>C.correlación				
	a)	b)	c)	d)	e)
2.1.Intensidad	0'92	-0'12	0'01	0'52	-0'87
2.2.Tipo ajuste	lineal	lineal	no lineal	no lineal	lineal
2.3.Tipo covariación	interdep	concord	cov.casual	dep.c.uni.	dep.ind.
2.4.Teorías previas					
2.5.Signo	+	-	+	+	-

La Tarea 3 es una modificación de la utilizada en la investigación de Estepa (1.994), aunque tareas semejantes fueron estudiadas en la investigación de Jennings, Amabile y Ross (1.982). Los contextos de las situaciones se han

ampliado, buscando aquellos que presentan situaciones más cotidianas para los sujetos (Vizmanos y Anzola, 1.988; Guzmán, Colera y Salvador, 1.988).

Tarea 3. Para cada una de las siguientes tablas de datos, estimar un valor razonable del coeficiente de correlación que muestre el tipo de relación entre las variables (directa, inversa o independencia) y su intensidad (fuerte o débil). **[No efectuar cálculos numéricos]**

a) Tiempo en meses desde que se prepara un medicamento y porcentaje de efectividad para una cierta enfermedad.

r =

Tiempo en meses	1	2	3	4	5
% de efectividad	90	75	42	30	21

b) Calificaciones de 10 alumnos de COU en los exámenes de Matemáticas y Física.

r =

Matemáticas	1	2	2	3	4	4	5	6	7	7
Física	3	7	6	2	2	7	4	5	3	5

c) Calificaciones de 10 alumnos de COU en los exámenes de Matemáticas y Educación Física.

r =

Matemáticas	2	3	4	4	5	6	7	8	9	10
Educación Física	2	5	7	8	5	4	6	5	5	9

d) Valores de la presión sanguínea antes y después de haber efectuado un cierto tratamiento médico a un grupo de 10 mujeres.

r =

	Presión sanguínea en cada mujer									
Mujer	Sra A	Sra B	Sra C	Sra D	Sra E	Sra F	Sra G	Sra H	Sra I	Sra J
Antes tratamiento	115	112	107	119	115	138	126	105	104	115
Después tratamiento	128	115	106	128	122	145	132	109	102	117

e) Ordenación dada por dos entrenadores a 10 atletas según su estado físico.

r =

	A	B	C	D	E	F	G	H	I	J
Entrenador A	1	2	3	4	5	6	7	8	9	10
Entrenador B	2	8	1	3	9	10	4	5	7	6

Como en esta situación se parte de un marco numérico, analizaremos si los estudiantes tienen en cuenta la tendencia de la variación conjunta para determinar

el signo de la correlación, para ello se cruzará con la Tarea 4, y qué métodos emplean para dar una estimación de la intensidad de la misma. Nuevamente deseamos estudiar el peso que las teorías previas tienen en los juicios de asociación. Es por ello que, en concreto en los apartados b) y c), se han planteado escenarios en los cuales se da correlación ilusoria (Chapman y Chapman, 1.969; Tversky y Kahneman, 1.982b; Murphy y Medin, 1.985; Jong, Merckelbach y Arntz, 1.995; Tomarken, Sutton y Mineka, 1.995), pues los datos contradicen las creencias de los alumnos.

En la Tabla 4.5.3 se presentan las diferentes variables contempladas en la Tarea 3.

Tabla 4.5.3. Valores de las variables en los apartados de la Tarea 3

VARIABLES	Tarea 3: Tabla ¬> Lectura y estimación ¬> C.correlación				
	a)	b)	c)	d)	e)
3.1.Intensidad	-0'98	-0'09	0'42	0'84	0'32
3.2.Tipo ajuste	no lineal	no lineal	lineal	lineal	lineal
3.3.Tipo covariación	dep.c.uni.	interdep	cov.casual	dep.ind.	concord
3.4.Teorías previas	coinciden	no coinciden	no coinciden	coinciden	coinciden
3.5.Signo	-	-	+	+	+

La Tarea 4 es una modificación del correspondiente de la investigación de Estepa (1.994). En él se pretende analizar qué estrategias utilizan los estudiantes al estimar el valor numérico del coeficiente de correlación que mejor se adapta a la gráfica ofrecida. Consideramos a priori que los métodos empleados abarcarían un espectro que va desde el puramente intuitivo, pasando por la valoración de la forma del diagrama de dispersión y llegando al ajuste más o menos intuitivo de las rectas de regresión.

Tarea 4. Dadas las siguientes gráficas que representan la variación conjunta de variables, estimar el valor de su coeficiente de correlación teniendo en cuenta el tipo de relación (directa, inversa o independencia) y la intensidad de la misma. **[No efectuar cálculos numéricos]**

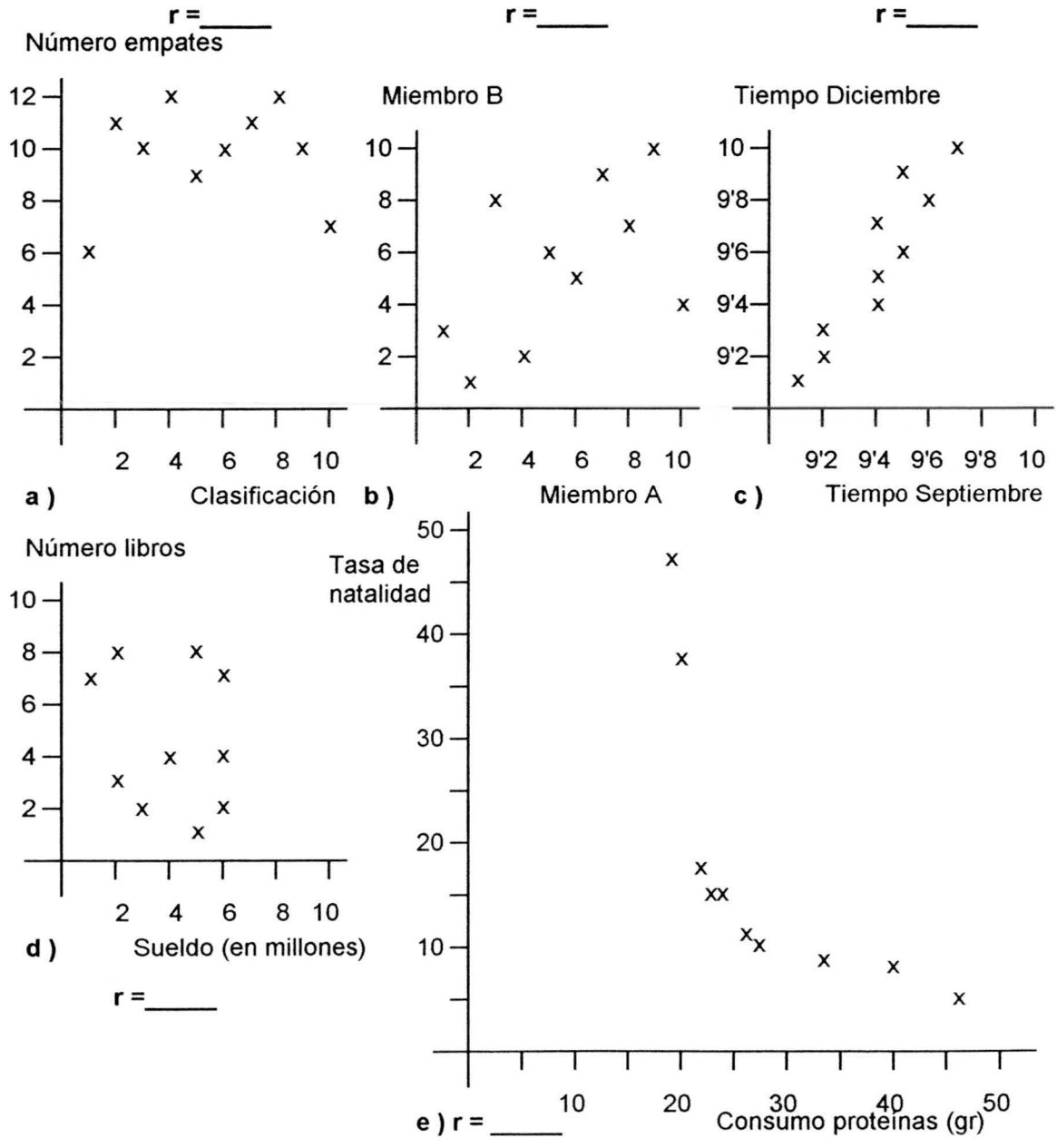
a) Puesto ocupado por los 10 primeros equipos de 1ª División de la liga de fútbol en la temporada 1.987-88 y los partidos empatados.

b) Puntuaciones otorgadas por los miembros A y B de un tribunal a 10 proyectos presentados.

c) Tiempo, en segundos, de 10 atletas en correr 100 m lisos en septiembre y diciembre.

d) Sueldo, en millones de pesetas, de los empleados de una empresa y número de libros que leen al cabo de un año.

e) Tasa de natalidad y consumo diario de proteínas animales en 10 países.



En la Tabla 4.5.4 se presentan las diferentes variables contempladas en la Tarea 4.

Tabla 4.5.4. Valores de las variables en los apartados de la Tarea 4

VARIABLES	Tarea 4: Gráfica¬>Lectura y estimación¬>Coeficiente correlación				
	a)	b)	c)	d)	e)
4.1.Intensidad	0'1	0'53	0'93	-0'22	-0'69
4.2.Tipo ajuste	lineal	no lineal	lineal	lineal	no lineal
4.3.Tipo covariación	interdep	concord	dep.c.uni.	cov.casual	dep.ind.
4.4.Teorías previas	no coinciden	coinciden	no coinciden	coinciden	coinciden
4.5.Signo	+	+	+	-	-

Tarea 5. Dados los siguientes valores del coeficiente de correlación lineal, describir dos variables para las cuales fuese razonable obtener este coeficiente de correlación en función del tipo de dependencia entre las variables (directa o inversa) y la intensidad de la misma.

a) r = 1

b) r = - 0'3

c) r = 0'05

d) r = - 0'8

e) r = 0'5

La Tarea 5, con el enunciado actual, no la hemos encontrado en la literatura de investigación. Con ella se intenta explorar como los alumnos a partir del valor numérico del coeficiente de correlación indican una pareja de variables cuya covariación conjunta se adapte, convenientemente, al mismo. Es una prueba de interpretación. Una cuestión que se estudiará es si el signo de la dependencia ofrecida por las variables que el alumno asocia coincide o no con el del coeficiente de correlación dado. Consideramos que una de las concepciones erróneas que puede emerger en esta tarea será la correspondiente al concepto de variable estadística bidimensional, pues la pareja de variables debe referirse a una unidad estadística. En la Tabla 4.5.5 se presentan las diferentes variables contempladas en la Tarea 5.

Tabla 4.5.5. Valores de las variables en los apartados de la Tarea 5

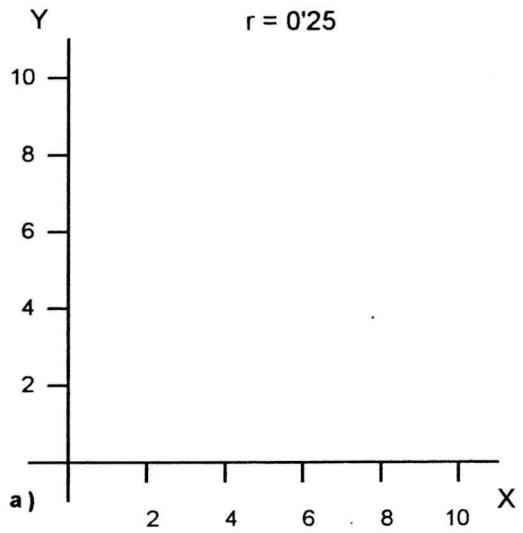
VARIABLES	Tarea 5: C.correlación¬>Interpretación¬>Descripc. verbal				
	a)	b)	c)	d)	e)
5.1.Intensidad	1	-0'3	0'05	-0'8	0'5
5.2.Tipo ajuste					
5.3.Tipo covariación					
5.4.Teorías previas					
5.5.Signo	+	-	+	-	+

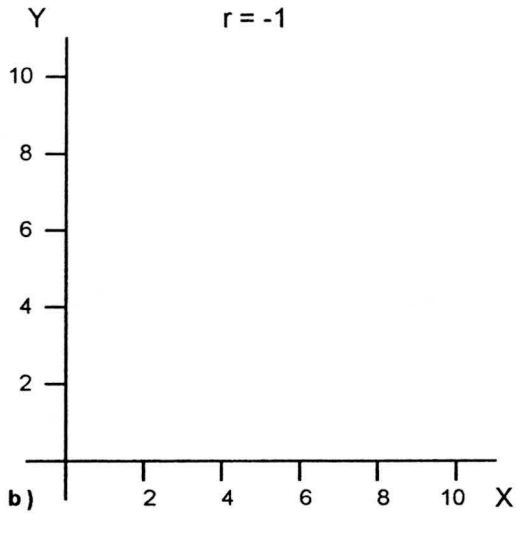
Tampoco hemos encontrado dentro de la literatura de investigación la Tarea 6. En ella se solicita a los alumnos que, a partir del valor numérico del coeficiente de correlación, dibuje un diagrama de dispersión que se ajuste, razonablemente, a dicho parámetro estadístico. Es una prueba bastante completa pues incluye tareas de interpretación, estimación y trazado. Deseamos conocer qué sentido tiene para los estudiantes el ángulo formado por las dos rectas de regresión y qué aplicación hacen ellos de la relación entre este ángulo y el coeficiente de correlación. Estimamos, también, que esta prueba debe aportar luz sobre las estrategias intuitivas que los alumnos ponen en juego cuando tienen que ajustar una recta a un diagrama de dispersión. En la Tabla 4.5.6 se presentan las diferentes variables contempladas en la Tarea 6.

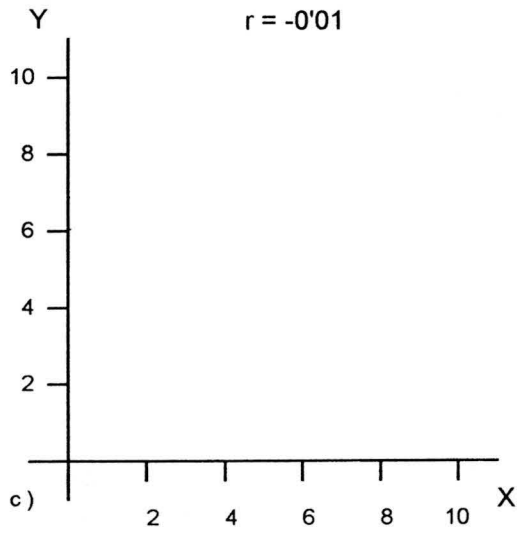
Tabla 4.5.6. Valores de las variables en los apartados de la Tarea 6

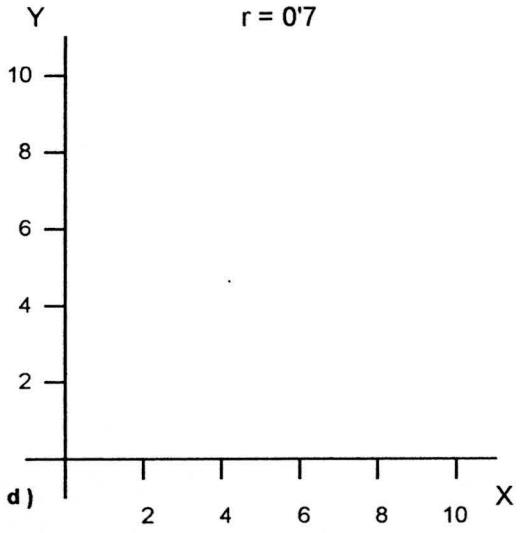
VARIABLES	Tarea 6: C.correlación¬>Interpretación¬>Gráfica				
	a)	b)	c)	d)	e)
6.1.Intensidad	0'25	-1	-0'01	0'7	-0'4
6.2.Tipo ajuste					
6.3.Tipo covariación					
6.4.Teorías previas					
6.5.Signo	+	-	-	+	-

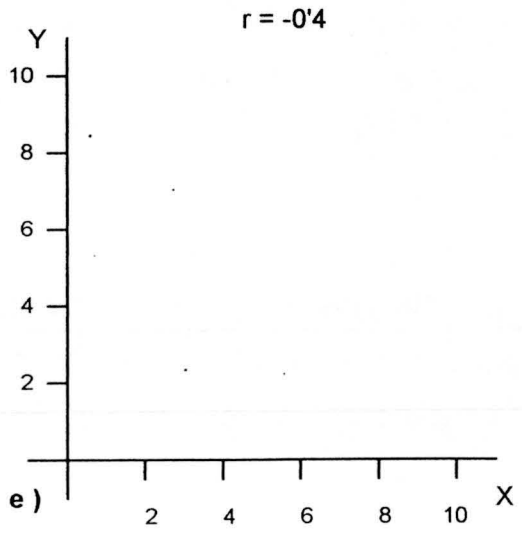
Tarea 6. Dados los siguientes valores del coeficiente de correlación entre dos variables X e Y, dibujar un diagrama de dispersión, con 10 puntos, que se adapte razonablemente a ellos.


Y
r = 0'25
10
8
6
4
2
a)
2
4
6
8
10
X


Y
r = -1
10
8
6
4
2
b)
2
4
6
8
10
X


Y
r = -0'01
10
8
6
4
2
c)
2
4
6
8
10
X


Y
r = 0'7
10
8
6
4
2
d)
2
4
6
8
10
X


r = -0'4
Y
10
8
6
4
2
e)
2
4
6
8
10
X

4.6. ANÁLISIS DE LOS PROBLEMAS PROPUESTOS

El cuestionario se completaba con dos problemas, esencialmente dirigidos a explorar las destrezas que los estudiantes poseían sobre procedimientos fundamentales y el dominio de ciertos hechos estadísticos sobre la noción de regresión. Son pruebas abiertas, donde no sólo se les solicita que expliciten los cálculos numéricos que estimen oportunos sino, de igual modo, que argumenten las estrategias utilizadas. Pensamos que en su etapa no universitaria los estudiantes podrían haber recibido una enseñanza deficitaria en actividades de interpretación y predicción, dado que la investigación de Sánchez Cobo (1.996) puso de manifiesto la carencia que los manuales de secundaria presentaban en este tipo de ejercicios. Deseamos evaluar que dominio han alcanzado los alumnos en este tipo de tareas después de un curso introductorio de Estadística.

Problema 1. Una recta de regresión tiene una pendiente de 16 y corta al eje de ordenadas en el punto y = 4. Si la media de la variable independiente es 8, ¿cuál es la media de la variable dependiente?

Este problema está tomado de Cruise, Dudley y Thayer (1.984, pág. 288). En la investigación de Sánchez Cobo (1.996) se destaca el escaso eco que recibía el centro de gravedad $(\bar{x},\bar{y})$ en los manuales de secundaria. Sin embargo, como subrayábamos con anterioridad, es una noción que juega un papel destacado en tareas de ajuste intuitivo de una recta de regresión a un diagrama de dispersión. Asimismo, es primordial en la obtención de la ecuación de dicha recta de regresión. Este problema pone de relieve la correspondencia entre el contexto geométrico y el estadístico de esta situación, pues el estudiante debe saber interpretar los conceptos de pendiente de una recta y de ordenada en el origen (Truran, 1.997). Precisamente, aquí consideramos que radicará una de las posibles fuentes de errores. También, se hace referencia a una de las características esenciales de la regresión: La predicción.

En un análisis a priori de este problema, consideramos que los alumnos podrían optar entre alguna de las potenciales estrategias de resolución siguientes:

A. Los alumnos para resolver el problema podrían apoyarse en alguna de los conceptos incluidos en este problema:

- Tratarían de utilizar la definición de media aritmética: $\bar{x}, \bar{y}$
- Tratarían de usar la propiedad del centro de gravedad $(\bar{x}, \bar{y})$
- Tratarían de utilizar la definición de covarianza: cov (X,Y)

B. Los alumnos para resolver el problema se apoyarían en un modelo lineal

1. El modelo lineal empleado sería una recta vectorial $y = kx$
2. El modelo lineal utilizado sería una recta de regresión:

a) Usarían las dos rectas de regresión: $R_{Y/X}$ y $R_{X/Y}$

- Determinarían la media de la variable dependiente en cada caso
- A partir de ella buscarían el punto de intersección: El centro de gravedad

b) Usarían una única recta de regresión:

- Aplicarían la recta de regresión de X sobre Y:
  - La expresión de la ecuación de la recta utilizada es $x = m'y + n'$
  - La expresión de la ecuación de la recta utilizada es

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

- Aplicarían la recta de regresión de Y sobre X:
  - La expresión de la ecuación de la recta utilizada es $y = m\,x + n$
  - La expresión de la ecuación de la recta utilizada es

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

  - Utilizarían ambas expresiones de la ecuación de la recta

Los conceptos de asociación estadística y de regresión lineal son fundamentales para el pensamiento estadístico, y en el caso de la correlación es un tópico central para muchos métodos estadísticos (Falk y Well, 1.997). La experiencia indica que las ideas básicas sobre estas nociones comportan, a veces, dificultades (Franklin, 1.988; Tamura, 1.994), lo que ha motivado algunos trabajos

interesantes en la enseñanza de la correlación y de la regresión (Goode y Gold, 1.987; Laviolette, 1.994).

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

a) Calcule el coeficiente de correlación lineal

b) Decir qué tipo de relación (directa, inversa o independencia) existe entre ambas variables

c) Determine la recta de regresión de y, número de bacterias por unidad de volumen, sobre x, número de horas

d) ¿Qué número de bacterias cabe esperar que habrá, transcurridas 2'5 horas?¿Y cuando pasen 6 horas?

e) ¿Qué tiempo deberá pasar para que el número de bacterias del cultivo sea de 27?

Este problema es una modificación del correspondiente de Vizmanos y Anzola (1.988, pág. 372). Con los apartados a), b) y c) pretendemos, aparte de examinar qué dominio y comprensión muestran los alumnos de los hechos estadísticos correspondientes -coeficiente de correlación (Truran, 1.995), tipos de dependencia, obtención de la ecuación de la recta de regresión, etc-, determinar si los estudiantes experimentan *"dificultades al realizar determinados procedimientos estadísticos"* (Morris, 1.997, pág. 6). El apartado d) tiene la intención de analizar si los sujetos de la muestra tienen interiorizado uno de los objetivos principales de la regresión, la predicción, y si comprenden la naturaleza estocástica de ella (Truran, 1.997).

Un error que consideramos que puede aflorar es al tratar de estimar el número de bacterias que cabe esperar que habrá, transcurridas 2'5 horas, los estudiantes, en vez de utilizar la recta de regresión de Y sobre X, empleen una estrategia interpolatoria. En este caso, será interesante analizar si los estudiantes

que utilicen dicha estrategia comparan el resultado obtenido mediante este procedimiento con el hallado en el apartado e), que también es 27.

Una función esencial de las matemáticas es la de proporcionar modelos que resuelvan un conjunto de problemas. Pero no se debe olvidar que la solución ofrecida por el método matemático aplicado es la correspondiente al modelo en cuestión, siendo imprescindible adecuarla al contexto de la situación problemática. Este es el caso del Problema 2 en su apartado d). Estamos interesados en observar si los alumnos tienen en cuenta el escenario dado y ajustan al mismo el resultado obtenido mediante la recta de regresión.

Otra cuestión relevante que subyace en este apartado es averiguar si los alumnos son sensibles a los peligros de la extrapolación, manifestando alguna limitación que sea inherente con su utilización en este caso.

Por último, el apartado e) aborda la distinción entre la recta de regresión de Y sobre X de la X sobre Y. En particular, juzgamos que podría surgir la concepción errónea de utilizar la misma recta de regresión Y sobre X para averiguar el tiempo que deberá pasar para que el número de bacterias del cultivo sea de 27, acaso producto de una visión funcional de la regresión. Naturalmente, un objetivo común a todos los apartados es analizar las estrategias y concepciones que los sujetos manifiesten en ellos.

Capítulo 5

Resultados de los items de opciones múltiples

5.1. INTRODUCCIÓN

Como se mostraba en el Capítulo 4, el cuestionario empleado en la presente investigación estaba constituído por cuatro bloques: i) Cuatro preguntas preliminares, ii) doce items con varios apartados, iii) seis tareas sobre actividades de traducción de una representación a otra, y, iv) dos problemas sobre regresión. En este capítulo analizaremos los resultados obtenidos en los 12 items del cuestionario, en los cuales se evaluaba la comprensión de determinados elementos de significado que no se consideraban en los otros dos bloques.

Para evitar que los alumnos acertasen por azar, cada uno de los items se compone de varios subitems, por lo que hay más de una respuesta correcta. Todos los items cumplen la exigencia de haber sido respondidos por más del 90% de la muestra (López Feal, 1986). En la Sección 4.3 se han relacionado los elementos de significado tenidos en cuenta en la fase previa a la construcción del cuestionario. Es por ello, que el análisis no se efectuó a partir de los items, sino atendiendo a esa clasificación, exponiéndose, a continuación, los resultados encontrados.

Las tablas con las diversas respuestas a los items se encuentran en el Anexo VI.

5.2. COVARIANZA, DEPENDENCIA E INDEPENDENCIA

Según Wild y Pfannkuch (1.998) uno de los elementos que fundamentan el pensamiento estadístico es reconocer el papel de la variación. En esta sección analizaremos los items cuyos contenidos versan sobre las nociones de covarianza, dependencia e independencia y las implicaciones que esto supone para las de correlación y regresión. Estos aspectos se encuentran recogidos en los items 1, 3, 6 y 7, que transcribimos a continuación, junto con la frecuencia de respuestas afirmativas obtenidas en cada subítem:

	Ítem 1 Si la covarianza de las variables X e Y es mayor que 0, las variables X e Y presentan	Nº respuestas afirmativas
X	a) Correlación positiva	127
X	b) La correlación podría ser no lineal	22
	c) Las variables podrían estar no correlacionadas	17
X	d) La pendiente de la recta de regresión tiene un signo positivo	87
X	e) El coeficiente de correlación es positivo	115

	Ítem 3 Cuando la intensidad de la relación entre dos variables decrece	Nº respuestas afirmativas
	a) La pendiente de la recta de regresión de Y sobre X crece	23
	b) La pendiente de la recta de regresión de X sobre Y crece	32
X	c) Hay mayor dispersión en la nube de puntos	124
	d) La covarianza aumenta de valor absoluto	33

	Ítem 6 Si el coeficiente de correlación entre dos variables es nulo	Nº respuestas afirmativas
	a) Ambas rectas de regresión de Y sobre X y de X sobre Y son paralelas	28
X	b) La covarianza también es nula	84
	c) Ambas rectas de regresión de Y sobre X y de X sobre Y coinciden	19
X	d) Las variables pueden tener una relación no lineal	64
X	e) Ambas rectas de regresión de Y sobre X y de X sobre Y son perpendiculares	96

	Ítem 7 Si r es el coeficiente de correlación de dos variables, indique qué afirmaciones son correctas	Nº respuestas afirmativas
X	a) $r = 0$ indica que las variables son independientes	138
	b) Si $r = 0'6$ la correlación entre las variables X e Y es doble que cuando $r = 0'3$	44
X	c) Una relación funcional entre variables se corresponde con un valor de r de +1 ó -1	112
	d) El coeficiente de correlación puede interpretarse como un porcentaje de la varianza	20

Para poder sintetizar las implicaciones que estas respuestas (o ausencia de respuestas) afirmativas tienen sobre la comprensión de los elementos de significado relacionados con la idea de covarianza y su conexión con el tipo de dependencia presentamos la Tabla 5.2.1. En la Tabla 5.2.1 se muestran las frecuencias y porcentajes, respecto al total de alumnos, de las respuestas afirmativas a los subitems de los items 1, 3, 6 y 7, relacionados con la covarianza y la dependencia de variables.

Los resultados son, en general, buenos, como podemos colegir del alto número de repuestas correctas. En la enseñanza recibida se ha definido la covarianza por su expresión algebraica clásica, $\sigma_{xy} = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i - \bar{x})(y_j - \bar{y}) f_{ij}$, y se ha argumentado su signo de manera expresiva, atendiendo a los signos de los dos primeros factores que integran la expresión anterior. En el epígrafe 4.4, ya exponíamos nuestro interés por determinar si los estudiantes tomaban en consideración la relación existente entre el signo de la covarianza y el del coeficiente de correlación. De los resultados obtenidos en los items podemos observar que una covarianza mayor que cero la relacionan con una correlación positiva el 65'8 % de los alumnos, mientras que cuando se trata de un coeficiente de correlación positivo el porcentaje de aciertos disminuye en más de un 6 %. Incluso, se nota que no han tomado en consideración la fórmula para el cálculo del coeficiente de Pearson, $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, de la que, obviamente, se deduce que de una covarianza mayor que cero se obtiene una correlación positiva y un coeficiente de correlación, asimismo, mayor que cero.

Tabla 5.2.1. Frecuencia y (porcentaje) de las respuestas referidas a la covarianza y la dependencia

	$\sigma_{XY}>0$	$\sigma_{XY}=0$	Intensidad crece	Relación no lineal	Independencia	Doble relación	Relación funcional	Rectas paralelas	Rectas coincidentes	Rectas perpendiculares
$r = 1, -1$							112 (58) ítem **7c**			
$r > 0$	115 (59'6) ítem**1e**									
$r = 0$		84 (43'5) ítem**6b**		64 (33'2) ítem **6d**	138 (71'5) ítem**7a**			28 (14'5) ítem **6a**	19 (9'8) ítem **6c**	96 (49'7) ítem **6e**
r doble						44 (22'8) ítem**7b**				
correlación positiva	127 (65'8) ítem**1a**									
correlación no lineal	22 (11'4) ítem**1b**									
variables no correladas	17 (8'8) ítem**1c**									
aumenta cov (X,Y)			33 (17'1) ítem **3d**							
aumenta la dispersión			124 (64'2) ítem **3c**							
pendiente positiva	87 (45'1) ítem**1d**									
pendiente Y sobre X crece			23 (11'9) ítem **3a**							
pendiente X sobre Y crece			32 (16'6) ítem **3b**							

También de la fórmula anterior, se deduce que cuando la covarianza decrece, también decrece la intensidad de la dependencia, lo que parece que no han tenido en cuenta un 17'1 % de los alumnos.

Como expresábamos en la sección 4.4, estamos interesados en conocer si los alumnos son conscientes de que una covariación positiva no lleva aparejada una regresión lineal. Aunque en la enseñanza se ha prestado interés tanto a la relación lineal como a algunos tipos de relación no lineales, no obstante, el uso casi exclusivo del coeficiente de correlación lineal y de la recta de regresión provoca, a nuestro juicio, la escasez de respuestas a la existencia de correlación no lineal cuando la covarianza es positiva, 11'4 %.

El hecho de que un coeficiente de correlación nulo implica la independencia de las variables, ha sido manifestado por un 71'5 % de los sujetos de la muestra.

5.3. CORRELACIÓN

En este epígrafe vamos a desarrollar el análisis de aquellos items que tengan conexión con la noción de correlación y que presentamos en los siguientes apartados:

i) El coeficiente de correlación es adimensional

ii) Correlación positiva y sentido de la covariación

iii) Correlación y dependencia lineal

iv) Intensidad del coeficiente de correlación

v) Relación entre la intensidad de la dependencia y la dispersión de la nube de puntos

vi) Correlación y proporcionalidad

vii) Confusión entre coeficiente de correlación y coeficiente de determinación

viii) Correlación y causalidad

El coeficiente de correlación es adimensional

Una característica importante del coeficiente de correlación, que ya destacábamos en el Capítulo 3, es la de ser un parámetro estadístico adimensional. Esta propiedad no es verificada por la covarianza, lo que implica que su utilización sea ardua como medida absoluta de la dependencia entre dos variables y notablemente complicado averiguar, mediante la simple inspección de una tabla de valores, la magnitud de la misma (Mendenhall, Scheaffer y Wackerly, 1.986). Este aspecto se encuentra recogido en el ítem 2, que transcribimos a continuación:

	Ítem 2 Juan correlaciona alturas y pesos de los estudiantes varones de 1º de la Diplomatura en Empresariales utilizando como unidades el metro y el kilogramo. Ángela registra las alturas y los pesos empleando centímetros y gramos como medida. Ambos calculan la correlación entre sus dos conjuntos de medidas.	Nº respuestas afirmativas
	a) La correlación de Ángela será mayor que la de Juan	23
X	b) Los dos coeficientes de correlación serán aproximadamente iguales	107
	c) El coeficiente de Juan tenderá a ser mayor que el de Ángela	17
X	d) El valor del coeficiente de correlación depende de la dispersión de los datos	102

La propiedad, anteriormente reseñada, es aceptada por el 55'4 % de los estudiantes de la muestra, que son los que eligen en el ítem 2 la opción b, *"Los dos coeficientes de correlación serán aproximadamente iguales"*. Sin embargo, el 11'9 % de los estudiantes consideran que el coeficiente de correlación será mayor si los datos son mayores en su expresión numérica, lo que equivale a decir que son dados en una unidad más pequeña, mientras que el 8'8 % juzgan que el coeficiente de correlación será mayor si la expresión numérica de los datos es menor, es decir, vienen dados en una unidad mayor. De lo anterior se infiere que, aproximadamente, un quinto de la muestra no es sensible a esta característica del coeficiente de correlación, estimando que el valor de r variará en función de las unidades de medida en que se exprese dicho parámetro.

Correlación positiva y sentido de la covariación

Una propiedad significativa de la dependencia directa es que las características de la variable estadística bidimensional covarían en la misma dirección, o sea, si la variable explicativa aumenta (disminuye) la variable explicada aumenta (disminuye). Este hecho estadístico es de notable influencia en los procedimientos que emplean los estudiantes para determinar el signo de la relación, como veremos en los capítulos 6 y 7, al analizar las tareas de traducción, en especial las actividades 1 y 2, sección 6.3, y el Problema 2 apartado b), sección 7.3.1.

Esta cuestión es planteada en el ítem 4, que presentamos a continuación:

	Ítem 4 Si las dos variables están correlacionadas positivamente	Nº resuestas afirmativas
X	a) Cuando una aumenta, la otra también aumenta	152
	b) Cuando una disminuye, la otra aumenta	8
X	c) Cuando una disminuye, la otra disminuye	90
	d) La relación entre las dos variables es de tipo lineal	91

El 78'7 % de los alumnos señalan que si las variables están correlacionadas positivamente, cuando una aumenta la otra también aumenta (opción a del ítem 4). No obstante, y como manifestábamos en la sección 4.4, su equivalente en el mismo ítem, cuando una variable disminuye la otra también disminuye, solamente es elegida por el 46'6 % de los alumnos (opción c del ítem 4). Un hecho similar se encuentra en las investigaciones psicológicas sobre las estrategias utilizadas para encontrar la asociación en una tabla de contingencia 2 x 2, donde en diversos estudios se ha encontrado que los sujetos utilizan con preferencia la casilla [a] (presente - presente) para realizar el juicio de asociación (Smedlund, 1963; Shaklee y Tucker, 1980; Beyth-Marom, 1982; Crocker, 1982; Shaklee y Mims, 1982; Yates y Curley, 1986; Estepa, 1994).

En concreto, Crocker (1982) encontró que el 77 % de los sujetos consideraban que el conocimiento de la casilla (presente - presente) era necesario

y suficiente para realizar un juicio de asociación, mientras que solamente el 52% percibían la importancia que tiene conocer la casilla (ausente-ausente). Asimismo, es posible que los alumnos se vean influídos por los inconvenientes suscitados por un proceso de *asimetría en el aprendizaje* (Martinón y Sauret, 1.990), o sea, entre el trabajo con secuencias crecientes, que son las habituales, no sólo en este nivel, sino desde las primeras aproximaciones que tienen los estudiantes a las matemáticas, y con secuencias decrecientes, que sería el proceso asimétrico, siendo éstos más difíciles para los alumnos, como han encontrado los autores anteriormente reseñados.

Solamente 8 alumnos (4'1 %) dan la respuesta incorrecta, afirmando que cuando las dos variables están correlacionadas positivamente, cuando una disminuye la otra aumenta (opción b del ítem 4).

Una cuestión que es interesante señalar es que esta propiedad de la correlación, crea dificultades a los alumnos cuando tienen que estimar el signo de la covariación basándose en una tabla de valores, como, por ejemplo, en la Tarea 3, pues, en general, las secuencias de aumento o disminución de las variables no se presentan de forma totalmente uniforme, sino, más bien, por rachas.

Correlación y dependencia lineal

Dada una variable estadística bidimensional hay dos problemas importantes, que tienen implicaciones mutuas, que son: a) Estudiar la vinculación o grado de relación que pueda existir entre las componentes de dicha variable estadística bidimensional, y, b) evaluar el tipo de ajuste que nos permitiría efectuar una mejor predicción de la variable explicada a partir de la variable explicativa (López Cachero, 1.990). Por tanto, primeramente estamos interesados en decidir si hay o no dependencia entre las variables y, una vez encontrado que dicha relación es lo suficiente intensa, buscar el tipo de ajuste, lineal o no, que explica de forma más conveniente la variación de una componente a partir de la otra. Para determinar cuál es el más eficiente tendremos que estudiar la bondad del ajuste. Evidentemente, de todo lo expuesto anteriormente, estos procesos son

independientes, significando con esto que la existencia de asociación no implica que la regresión sea de un tipo determinado, ni al contrario.

Esta propiedad no es asumida por 91 estudiantes (47'2 %) - opción d del ítem 4 -, los cuales declaran que si dos variables están correlacionadas positivamente, la relación entre las variables es de tipo lineal. Por el contrario, en el ítem 6 opción d, el 33'2 % de los estudiantes asumen que, si el coeficiente de correlación es nulo, las variables pueden tener una relación no lineal. Aunque en la enseñanza se ha estudiado de forma explícita que la relación entre las variables puede ser diferente a la lineal, en concreto se han estudiado diversos tipos de ajustes no lineales, en realidad, se produce un sesgo al hacer más hincapié en .los ajustes lineales y en el coeficiente de correlación lineal, lo cual puede constituirse en un obstáculo de tipo didáctico y conducir a los alumnos a apreciar que, en caso de que el coeficiente de correlación lineal sea distinto de cero, la asociación sería lineal.

Intensidad del coeficiente de correlación

Una actividad importante es la ordenación de varios valores de un coeficiente de correlación, dado que lleva implícito el conocimiento de los diversos tipos de dependencia y su conexión con la intensidad de la misma. Esta actividad se ha recogido en el ítem 5 que transcribimos a continuación:

Ítem 5 Ordene los siguientes valores según expresen mayor correlación entre dos variables: 0'5, -0'8, 0'2, -0'4, 0

_____ mayor correlación

_____

_____

_____

_____ no existe correlación

En la tabla 5.3.1 se dan las frecuencias y porcentajes del modo en que los alumnos ordenan cinco valores dados del coeficiente de correlación según la

intensidad, decreciente, de los mismos. Morris (1.997) proponía el mismo ítem y obtuvo resultados bastante similares. Para la respuesta [II] alcanzó un 15 % y para la [IV] un 25 %, pero si se tiene en cuenta que en la citada investigación no se contempla la que hemos denotado con [III] y si sumamos la [III] y [IV], que únicamente se diferencian en el orden del -0'4 y -0'8, conseguimos un 18'2 % no muy lejano del 25 % de Morris. Esta investigadora explica estos errores como un indicio de las dificultades que tienen los alumnos con el concepto de correlación.

Estimamos que la explicación de las respuestas erróneas - de la [II] en adelante, en la Tabla 5.3.1 -, se puede llevar a cabo de una manera más completa. En primer lugar, en la respuesta [II], se puede observar que los alumnos clasifican los valores aplicando la ordenación usual en **R**, es decir, sólo tienen en cuenta su significado numérico y no valoran su significado estadístico, o sea, que son los valores de un coeficiente de correlación.

Tabla 5.3.1. Frecuencia y porcentaje de las ordenaciones de las intensidades del coeficiente de correlación

RESPUESTA[1]	FRECUENCIA	PORCENTAJE
[I] -0'8, 0'5, -0'4, 0'2, 0*	89	46'1
[II] 0'5, 0'2, 0, -0'4, -0'8	33	17'1
[III] 0'5, 0'2, -0'4, -0'8, 0	26	13'5
[IV] 0'5, 0'2, -0'8, -0'4, 0	9	4'7
[V] 0, 0'2, -0'4, 0'5, -0'8	3	1'6
[VI] -0'8, -0'4, 0'5, 0'2, 0	3	1'6
[VII] Otras**	23	11'9
No responde	7	3'6
Total	193	100

[1] Ordenación de mayor a menor intensidad de los coeficientes de correlación.
* Respuesta correcta
**Otras ordenaciones con frecuencia absoluta 2 ó 1

En Estepa (1.994), Batanero y cols. (1.997), Batanero, Godino y Estepa (1.998), Batanero, Estepa y Godino (1.998) en el análisis del proceso de aprendizaje de una pareja de alumnos, ya se observó que los alumnos, al comparar

dos coeficientes de correlación negativos, sentían cierto rechazo a considerar como mayor correlación el de mayor valor absoluto, ya que al ordenar estos valores negativos como números es mayor el de menor valor absoluto. El conocimiento del orden de los números negativos se constituye en obstáculo para ordenar la intensidad de la correlación (González y cols., 1.990).

Por otra parte, como se verá en el análisis de las tareas, Capítulo 6, y en el estudio del Problema 2, Capítulo 7, hemos detectado que algunos alumnos interpretan el coeficiente de correlación como un parámetro matemático, cuyo campo de existencia sería el conjunto de los números reales **R**, haciendo corresponder los diversos tipos de dependencia con un subconjunto de **R** de la siguiente manera:

- Si existe asociación, el coeficiente de correlación *r* pertenecerá al intervalo [-1,1], alcanzándose la máxima intensidad cuando *r* es nulo, disminuyendo la fuerza de la relación cuanto más nos acercamos a 1 y -1, a cuyos valores asocian una dependencia débil.
- Si hay independencia, el valor del coeficiente de correlación *r* pertenecerá al conjunto complementario del intervalo [-1,1], o sea, $(-\infty, -1) \cup (1, \infty)$.

En consecuencia, cuando se pide que realicen la estimación del coeficiente de correlación de una variable bidimensional cuyas componentes consideran que están incorreladas, dan un valor fuera del intervalo [-1,1], como por ejemplo $r = 3$ o bien $r = -2$, según las juzguen "más o menos independiente". Así, por ejemplo, el sujeto 183 responde a la Tarea 2 apartado b) de la siguiente manera: *"r = 2. No hay relación"*. Estos alumnos, que no son muchos, habrán dado una ordenación del tipo [V] de la tabla 5.3.1.

Es interesante mencionar que en la investigación de Morris (1.997, pág. 19), se plantea a los estudiantes la siguiente pregunta: *"¿Cuál es probablemente el coeficiente de correlación que puedes obtener que exprese la no existencia de relación entre dos variables? (Por ejemplo entre la autocalificación de la propia ambición y la altura de los estudiantes)"*. El veinte por ciento de las respuestas de los estudiantes, según esta autora, muestran una falta de conocimiento para

responder a esta pregunta, o la respuesta es confusa, dando como ejemplo de respuesta la siguiente: *"probablemente menos que 3'8 o menos que 2' algo - esto indica no correlación"*.

Según esta autora, el quince por ciento de los estudiantes piensan que una correlación perfecta negativa indicaría la no relación de las variables estudiadas, dando como ejemplo el siguiente: *"muy cercano a -1. Probablemente -0.95"*. Estamos de acuerdo que la concepción unidireccional (Estepa, 1,994) aparece en muchos alumnos, pero también interpretamos estos ejemplos de Morris (1.997) según lo expuesto al principio de este párrafo, es decir, existen alumnos que creen que para dos variables independientes el coeficiente de correlación cae fuera del intervalo [-1,1] y si la dependencia es débil, próxima a la independencia, el coeficiente de correlación estará próximo a +1 ó -1. Volveremos a este punto en los Capítulos 6 y 7, secciones 6.3 y 7.2.2 y 7.3.1, mostrando otros ejemplos de respuestas de sujetos de la muestra que presenten esta concepción errónea.

En las restantes respuestas consideramos que se ponen de manifiesto dos características relevantes. En primer lugar, los alumnos juzgan que el cero indica que no existe correlación, esto es, todos los que dan el cero en último lugar, lo cual ha sido contestado por 48 estudiantes. Si, además, contamos con los que realizan la ordenación [I] tenemos que el número total de alumnos de la muestra que tiene presente esta propiedad sería de 137, lo que equivale a un 71 %, lo que es corroborado en la opción a del ítem 7 donde 138 alumnos (71'5 %) responden que "r = 0 *indica que las variables son independientes*". En segundo lugar, los estudiantes que utilizan las ordenaciones [III], [IV] y [VI], clasifican el resto de los valores, o sea los no nulos, siguiendo el orden de los números reales, [III] y [IV], o siguiendo una ordenación de carácter topológica, los negativos a la izquierda y los positivos a la derecha siguiendo el modelo de la recta real. En el caso [III] y [IV] los alumnos escriben primero los positivos y después los negativos, clasificando los valores de cada subconjunto siguiendo el orden numérico usual. A este respecto, deseamos subrayar que en [IV] los alumnos anteceden -0'8 a -0'4, lo cual muestra que estos alumnos no dominan la ordenación en $\mathbf{R}^-$. En el caso [VI] los alumnos ordenan los valores no nulos del coeficiente de correlación de manera isomorfa a la

que tienen en la recta real, presentando las mismas dificultades que hemos comentado anteriormente.

Relación entre la intensidad de la dependencia y la dispersión de la nube de puntos

Un concepto básico en estadística, que aparece en diferentes contextos, es el de dispersión. En la sección 4.4, ya indicábamos nuestra preocupación por determinar si los alumnos son conscientes de la influencia de la dispersión sobre la intensidad de la asociación y el coeficiente de correlación. En particular, en el tópico de la correlación el coeficiente de correlación nos expresa "cuán pegados" o condensados están los puntos del diagrama de dispersión alrededor de la recta de regresión, siendo ésta la más próxima, en el sentido de los mínimos cuadrados, a los puntos de la distribución (Calot, 1.982). El 52'9 % de los alumnos, al aceptar que el valor del coeficiente de correlación está en función de la dispersión de los datos, confirman que dominan este conocimiento, por el cual nos hemos interesado en la opción d del ítem 2: *"El valor del coeficiente depende de la dispersión de los datos"*.

Mientras, el 64'3 % de los alumnos aceptan que si la intensidad de la relación disminuye, hay mayor dispersión de los puntos del diagrama de dispersión, cuestión por la que nos hemos interesado en la opción c del ítem 3 - *"Cuando la intensidad de la relación entre dos variables decrece hay mayor dispersión en la nube de puntos"* -. Observemos que esto implicaría que los alumnos consideran que si la intensidad decrece, la dependencia entre las componentes de la variable estadística bidimensional se torna débil y, por consiguiente, los puntos de la nube estarán muy dispersos.

Estos extremos se relacionarán con el estudio de las tareas, realizado en el Capítulo 6, secciones 6.2 y 6.3, para analizar mejor este conocimiento.

Correlación y proporcionalidad

La noción de proporcionalidad es fundamental y básica dentro de las matemáticas. Además, presenta múltiples conexiones con otras disciplinas como, por ejemplo, la geometría, la técnica, la economía, el arte, la ecología, ... (Fiol y Fortuny, 1.990). Un punto significativo que se aborda desde las etapas iniciales es el de la clasificación de la proporcionalidad en directa e inversa.

En la enseñanza y aprendizaje de la proporcionalidad la discriminación entre proporcionalidad directa e inversa se transforma en una actividad primordial, procurándose dotar a los alumnos de herramientas que les permitan, de forma eficiente y cómoda, establecer tal distinción. Así, si consideramos, por ejemplo, la proporcionalidad directa y lo hacemos desde una dimensión funcional, ésta se puede caracterizar como una *función lineal* $f(x) = kx$, dado que verifica las siguientes propiedades: i) $f(x + y) = f(x) + f(y)$, y, ii) $f(\lambda x) = \lambda f(x)$. En la enseñanza es frecuente utilizar esta segunda propiedad como procedimiento para determinar si la proporcionalidad es directa. Con tal fin, se le suele redefinir basándose en el hecho de que a un aumento (disminución) de una variable, o sea, $x \rightarrow \lambda x$, le corresponde un aumento (disminución) de la otra, o sea, $f(x) \rightarrow \lambda f(x)$, en igual proporción.

Asimismo, en la enseñanza de la correlación, para distinguir el tipo de dependencia, a los alumnos se les ofrece proposiciones similares a la anteriores. De esta manera, si la relación es directa se cumplirá, respecto de la variable predictora y de la variable respuesta (Cook y Weisberg, 1.994), que un aumento o disminución de la primera conlleva un aumento o disminución de la segunda, respectivamente.

Analogías como la anterior o como, por ejemplo, la referida a la terminología (correlación directa e inversa / proporcionalidad directa e inversa), estimamos que favorecen, en algunos estudiantes, la confusión entre proporcionalidad y asociación. No es de extrañar, por tanto, que ciertos alumnos interpreten al coeficiente de correlación r como si tuviera un papel análogo al de la constante de proporcionalidad. Esto es cierto, únicamente, en el caso de que las variables estén

tipificadas, donde la ecuación de regresión se reduce a una función lineal homogénea y *r* es la constante de proporcionalidad que relaciona el valor X con la media de las distribuciones Y condicionadas con X. El 22'8 % de los alumnos responden que *"si r = 0'6 la correlación entre las variables X e Y es doble que cuando r = 0'3"*, opción b del ítem 7, donde pudiera ser que estos alumnos extiendan la proporcionalidad al caso general para el cual esta característica no se verifica. Además, lo anterior queda reforzado por las respuestas correctas en la elección de las opciones a y c del ítem 4.

Confusión entre *r* y r^2

El coeficiente de determinación pone de manifiesto el grado de aproximación de los puntos del diagrama de dispersión a una relación lineal. Asimismo, nos expresa cómo se ha reducido la varianza al emplear para predecir el valor de y la recta de regresión de Y sobre X en lugar de la recta $y = \bar{y}$ (Ríos, 1.976; Hermoso y Hernández, 1.989).

Parece ser que algunos alumnos confunden al coeficiente de correlación *r* con el coeficiente de determinación r^2 , pues juzgan (10'4 %, opción d del ítem 7) que el primero puede interpretarse como un porcentaje de la varianza.

Correlación y causalidad

Muchas de las relaciones existentes entre dos variables son de tipo causal, es decir, presentada la causa podemos, a priori, predecir el efecto con una cierta probabilidad. Aunque la asociación está interesada, también, por la relación entre dos variables, la existencia de un valor significativo del coeficiente de correlación puede ser debida a otros tipos de relación entre las variables. Este punto ya ha sido discutido cuando hablábamos de los diversos tipos de covariación. Así la correlación puede ser explicada por la causalidad, dependencia indirecta, concordancia, interdependencia, efecto de una variable oculta o ser simplemente

espúrea. Sin embargo, es muy frecuente que los alumnos consideren semejantes ambas nociones.

Así, en la investigación de Morris (1.997), el 25 % de los sujetos de la muestra indican que es posible inferir de la correlación alguna forma de causalidad. Por ejemplo, al preguntárseles sobre si de la relación entre dos variables podría inferirse siempre una relación causal, un alumno responde: *"Sí, una variable puede ser la causa de que la otra ocurra"* (pág. 16). Por el contrario, el 55 % de los estudiantes responden negativamente a esta cuestión.

En el cuestionario el ítem 8 recoge esta característica y lo transcribimos a continuación:

	Ítem 8 Al estudiar las superficies sembradas de trigo en miles de hectáreas y las cosechas obtenidas en millones de quintales métricos, en cinco años consecutivos, el coeficiente de correlación obtenido fue de 0'91. Luego	Nº respuestas afirmativas
X	a) Podría haber otros factores que hagan variar los resultados	56
	b) Deberíamos tomar una muestra más grande para poder expresar la relación entre la superficie plantada y la cosecha obtenida	6
X	c) La cosecha obtenida presenta una alta correlación con la superficie plantada	172
	d) Si plantamos doble superficie, obtendremos con seguridad doble cosecha	43

La concepción causal, puesta de manifiesto en la tesis de Estepa (1994), Estepa y Batanero (1.995) y Batanero y cols. (1.997), aparece de nuevo aquí, cuando el 22'3 % de los alumnos (opción d del ítem 8) afirman que si el coeficiente de correlación es de 0'91, al plantar doble superficie obtendremos con seguridad el doble de cosecha. Esto nos induce a considerar que estos alumnos piensan que la relación entre las variables es funcional y que un alto valor de la correlación implicaría una relación causa-efecto determinista entre ellas. Por otra parte, estos alumnos exhiben un espectro de tipos de covariación (Barbancho, 1.973) muy reducido, pues no toman en consideración más que la dependencia causal unilateral, ignorando las restantes, en especial y en relación con este ítem, la dependencia indirecta, así como el carácter aleatorio de la dependencia.

Treintiséis alumnos, 18'7 % del total, han elegido simultáneamente las respuestas c y d de dicho ítem y el 89'1 % de los alumnos dan una interpretación correcta del coeficiente de correlación al expresar que, si el coeficiente de correlación es de 0'91, la cosecha obtenida presenta una alta correlación con la superficie plantada. Siete alumnos, 3'7 % del total, han elegido las respuestas a y d simultáneamente y el 29 % de los alumnos admiten la dependencia aleatoria al afirmar que, si el coeficiente de correlación es de 0'91, podría haber otros factores que hagan variar los resultados, lo que muestra una concepción aleatoria correcta y no causalista de la asociación.

Solamente el 3'1 % de los alumnos (opción b del ítem 8) opinan que se debería tomar una muestra más grande para poder expresar la relación entre la superficie plantada y la cosecha obtenida.

5.4. REGRESIÓN

En esta sección abordamos el análisis de los items que se centran sobre hechos estadísticos que están conectados con la regresión. Lo vamos a desarrollar en dos subsecciones, que son las siguientes: i) Interpretación de la bondad del ajuste, y, ii) distinción entre la variable explicativa y la variable explicada.

Interpretación de la bondad del ajuste

El ítem 9 plantea una situación de interdependencia entre las dos variables, en la que no se distingue entre la variable dependiente y la independiente. Transcribimos a continuación el ítem 9:

	Ítem 9 ¿En qué predicción tendría más confianza?	Nº respuestas afirmativas
	a) La predicción de la estatura de un hombre a partir de su peso	13
	b) La predicción del peso de un hombre a partir de su estatura	67
X	c) Las dos me dan la misma confianza	114

Puesto que el coeficiente de correlación r es simétrico y su cuadrado expresa la bondad del ajuste, la opción c es la respuesta correcta, habiendo sido elegida en solitario por el 58'5 % de los estudiantes, es decir, la contestan 113 de los 193 sujetos de la muestra.

Es de notar que esta propiedad no ha sido comprendida por quienes eligen en solitario las opciones b (34'2 %, 66 de los 193 sujetos de la muestra) y a (6'2 %, 12 de 193 sujetos de la muestra), siendo necesario manifestar que hay un sujeto de la muestra que señala todas las respuestas. Estos resultados están en concordancia con los obtenidos por Tversky y Kahneman (1.982a), quienes encontraron, igualmente, que los sujetos consideraban que la variable estatura es la que mejor explica para ellos el peso, y no al revés.

Distinción entre la variable explicativa y la variable explicada

Como indicábamos en el Capítulo 4 la distinción entre variable explicativa y explicada es de notable importancia para la regresión y para poder discriminar entre las dos rectas de regresión (Porkess, 1.996). Con este objetivo hemos planteado el ítem 10, que transcribimos a continuación:

	Ítem 10 Las rentas se usan para predecir los ahorros, ambos medidos en miles de pesetas. Para la ecuación de regresión $y = 1000 + 0'1x$, ¿cuál de las siguientes afirmaciones es verdadera?	Nº respuestas afirmativas
	a) Y es la renta, X es el ahorro, la renta es la variable independiente	26
	b) Y es la renta, X es el ahorro, el ahorro es la variable independiente	66
	c) Y es el ahorro, X es la renta, el ahorro es la variable independiente	56
X	d) Y es el ahorro, X es la renta, la renta es la variable independiente	70

En el ítem 10, dada la ecuación de la recta de regresión y la situación que describe, se pide a los alumnos que identifiquen en ella las variables y señalen la variable explicativa (independiente). Dentro de los que señalan entre sus respuestas la opción d, el 36'3 % del total, únicamente el 25'9 % responden en solitario la opción d -respuesta correcta- y el 10'4 % responden las opciones b y d

(b no es correcta). Es de destacar que el 36'8 % de los sujetos de la muestra confunden la recta de regresión dada con la X sobre Y, o sea contestan los apartados a y/o c, hecho que está en consonancia con los resultados obtenidos en el Problema 2 apartado e), sección 7.3.3.

Esta dificultad fue identificada en los trabajos de Estepa (1.994) y Batanero y cols. (1.997), Batanero, Godino y Estepa (1.998).

5.5. CORRELACIÓN Y REGRESIÓN

En este epígrafe estudiamos los items que tienen como objetivo fundamental hechos estadísticos que conectan características de la correlación y de la regresión.

Este estudio se ha realizado en cuatro apartados: i) Relación de la intensidad de la dependencia entre dos variables y las rectas de regresión, ii) relación entre el valor del coeficiente de correlación y la pendiente de las rectas de regresión, iii) dependencia funcional y valor del coeficiente de correlación, y, iv) correlación perfecta y ángulo formado por las rectas de regresión.

Relación de la intensidad de la dependencia entre dos variables y las rectas de regresión

Algunos estudiantes consideran que hay una conexión entre la intensidad de la relación y la pendiente de la recta de regresión. Así, en el ítem 3, el 16'6 % de los alumnos eligen la opción b, es decir, afirman que cuando la intensidad de la relación decrece la pendiente de la recta de regresión de X sobre Y crece y, el 12'0 % de los estudiantes eligen la opción a, o sea, cuando la intensidad de la relación disminuye la pendiente de la recta de regresión de Y sobre X crece.

El 49'7 % de los alumnos (opción e del ítem 6) responden adecuadamente que si el coeficiente de correlación es nulo ambas rectas de regresión de Y sobre X y de X sobre Y son perpendiculares. Sin embargo, el 14'5 % de los alumnos afirman

que son paralelas (opción a del ítem 6) y el 9'8 % que coinciden (opción c del ítem 6). En todo caso, Sánchez (1.996a, 1.996b) ha señalado las dificultades de los alumnos con la idea de independencia. Entre estos últimos, probablemente, se encuentran los que creen que cuando existe asociación el coeficiente de correlación varía en el intervalo [-1,1], siendo máxima la correlación cuando $r = 0$. Hecho que hemos discutido en las secciones 5.3, 6.2 y 6.3, 7.2.2 y 7.3.1.

Valor del coeficiente de correlación y pendientes de las rectas de regresión

En el ítem 11, se les pregunta a los alumnos por el valor del coeficiente de correlación, si la pendiente de las dos rectas de regresión son idénticas. Transcribimos a continuación dicho ítem:

	Ítem 11 ¿Qué valor ha de tener r si las dos rectas de regresión tienen una pendiente idéntica?	Nº respuestas afirmativas
	a) 0	29
X	b) 1	149
X	c) -1	80
	d) 0'5	7

Es de subrayar que, dentro de las respuestas correctas, el 77'2 % de los alumnos responden 1, el 41'5 % responden -1, siendo la doble respuesta (1,-1) del 38'9 %. Una vez más observamos la dificultad de comprensión que produce en los alumnos la asociación negativa. El presente ítem es una extensión del correspondiente de la investigación de Morris (1.997). Esta autora planteaba a los estudiantes, únicamente, *¿qué valor toma el coeficiente de correlación positivo perfecto?*, que es una restricción del que presentamos nosotros. Encontró que el 10 % de los sujetos de la muestra no contestaban o bien por desconocimiento o por estar confusa la pregunta, dado que un alumno respondía con una interrogación. El resto, o sea el 90 %, dió la contestación correcta.

Es de destacar también que el valor 0 (opción a del ítem 11) lo dan un 15'0 % de los alumnos, entre los que se encuentran los que consideran que en caso de

existir asociación el coeficiente de correlación varía en el intervalo [-1,1], tomando su máximo en 0, error que ya analizábamos de forma exhaustiva en la sección 5.3 en el apartado dedicado a la intensidad del coeficiente de correlación, y que surge, y, por consiguiente, también es examinado, en los capítulos 6 y 7, secciones 6.2 y 6.3, 7.2.2 y 7.3.1.

Además, si relacionamos lo anterior con la opción c del ítem 6, que recordemos que era: *"Si el coeficiente de correlación entre dos variables es nulo, c) Ambas rectas de regresión de Y sobre X y de X sobre Y coinciden"*, vemos que en este caso 19 de los 193 sujetos de la muestra, o sea un 9'8 %, responden afirmativamente a esta cuestión.

Dependencia funcional y valor del coeficiente de correlación

En el ítem 7 opción c se les solicitaba a los alumnos que evaluaran si *"una relación funcional entre variables se corresponde con un valor de r de +1 ó -1"*. Esta pregunta, que relaciona la dependencia funcional con el valor del coeficiente de correlación, fue contestada correctamente por el 58 % de los estudiantes. Este resultado contrasta con el obtenido en el apartado anterior, en el que sólo el 38'9 % de los sujetos de la muestra interpretaban que si ambas rectas de regresión tenían pendientes coincidentes el coeficiente de correlación debía valer 1 ó -1. Podemos deducir de lo anterior, que para estos alumnos es más fácil relacionar los valores máximos de *r* con el tipo de asociación, funcional en este caso, que con el ángulo que forman las dos rectas de regresión, y, asimismo, tienen dificultades en ligar estas características.

Correlación perfecta y ángulo de las rectas de regresión

Un hecho estadístico notable es que la mayor o menor amplitud del ángulo formado por las rectas de regresión es indicativo de la menor o mayor intensidad de la relación lineal entre las componentes de una variable estadística bidimensional (Barbancho, 1.973).

Esta característica se ha planteado en el ítem 12, que transcribimos a continuación:

	Ítem 12 Si la correlación entre X e Y es perfecta, el ángulo que forman las rectas de regresión es de	Nº respuestas afirmativas
	a) 120º	5
	b) 90º	32
	c) 45º	24
X	d) 0º	129

Aunque el 66'8 % de los estudiantes incluyen dentro de sus respuestas la opción d, únicamente el 64'8 % de los alumnos responden sólo con la opción d, es decir, declaran que cuando la correlación entre X e Y es perfecta, el ángulo que forman las rectas de regresión es nulo. Es conveniente indicar que el 16'6 % estiman que el ángulo formado por las rectas de regresión es de 90º, opción b, a pesar de que éste correspondería a una pareja de variables incorreladas o independientes. Como exponíamos en el apartado anterior vemos que los sujetos de la muestra tienen dificultades para establecer una correspondencia entre la intensidad de la asociación y el ángulo formado por las rectas de regresión dado que, prácticamente, un tercio de ellos dan respuestas total o parcialmente inadecuadas.

5.6. CONCLUSIONES SOBRE EL CONOCIMIENTO CONCEPTUAL DE LOS ALUMNOS

En este capítulo hemos analizado las respuestas dadas por los alumnos de la muestra a los 12 items recogidos en el cuestionario, cuyo contenido se ha estudiado con detalle en el Capítulo 4. Estas respuestas han sido comparadas con los elementos de significado relacionados con la covarianza, dependencia e independencia, correlación y regresión identificados en dicho capítulo. A continuación exponemos las principales conclusiones obtenidas.

Covarianza, dependencia e independencia

La mayoría de los alumnos conocen que el signo de la covarianza denota la dirección de la correlación existente entre las componentes de una variable estadística bidimensional. En concreto, dos de cada tres alumnos, aproximadamente, explicitan la relación entre la covarianza positiva y la asociación directa. Asimismo, reconocen (tres de cada cinco alumnos) que el signo de la covarianza es un indicador del signo del coeficiente de correlación.

Un pequeño porcentaje de alumnos no toma en consideración el decrecimiento de la intensidad de dependencia al disminuir el valor absoluto de la covarianza, ni la posibilidad de que el tipo de relación de las variables, cuando la covarianza es positiva, sea no lineal.

Correlación

Los alumnos comprenden con facilidad la adimensionalidad del coeficiente de correlación r, así como la relación entre el signo de la correlación y el sentido en que covarían los valores de las componentes de una variable estadística bidimensional. No obstante, y aunque conocen que si la correlación es positiva los valores de las variables varían en el mismo sentido, algunos no son conscientes de que esta propiedad se aplica también al caso en el que las dos variables disminuyen simultáneamente. La covariación en sentido creciente (a un aumento en el valor de la variable explicativa le corresponde un aumento en el valor de la variable explicada) se comprende mejor que la covariación en sentido decreciente (a una disminución en el valor de la variable explicativa le corresponde una disminución en el valor de la variable explicada), que sólo es captada por menos del cincuenta por ciento de los sujetos de la muestra.

Los alumnos tienen muy presente que la intensidad de la dependencia se obtiene a partir del coeficiente de correlación. Así, nueve de cada diez sujetos de la muestra afirman que un valor alto del coeficiente de correlación implicaría una asociación fuerte. En particular, resulta sencillo la interpretación de los valores 0, 1

y -1, o próximos. Los alumnos relacionan correctamente la correlación nula con la independencia de variables, en el caso de relación de tipo lineal, aunque tienen más dificultad en comprender que podría, en este caso, existir una relación no lineal. Asimismo, relacionan de forma pertinente la correlación perfecta o dependencia funcional con el valor del coeficiente de correlación correspondiente, 1 y -1. En concreto, lo efectúan tres de cada cinco sujetos de la muestra.

Por el contrario, hemos encontrado dificultades en la comparación de diferentes valores del coeficiente de correlación. Al solicitar a los alumnos que ordenen, de mayor a menor intensidad de la asociación, una serie de valores del coeficiente de correlación *r*, hemos obtenido cuatro ordenaciones diferentes, que completan las señaladas por Morris (1.997): (a) -0'8, 0'5, -0'4, 0'2, 0, (b) 0'5, 0'2, 0, -0'4, -0'8, (c) 0'5, 0'2, -0'4, -0'8, 0, (d) 0, 0'2, -0'4, 0'5, -0'8. La clase (a) es la correcta. La clase (b) exhibe el orden usual de los números reales, que ha sido un obstáculo para que los alumnos interpreten el significado estadístico del coeficiente. La clase (c) indica que el alumno interpreta que 0 representa la no existencia de correlación y lo sitúa en el último lugar, pero el resto de los valores los ordena como números reales. Finalmente, la clase (d) pone de manifiesto una concepción inadecuada, consistente en juzgar que *r* es un parámetro, tal que si hay asociación el valor del coeficiente de correlación pertenecerá al intervalo [-1,1], siendo mayor la intensidad en el 0 y decreciendo conforme nos acerquemos a los valores frontera -1 y 1; si no hay correlación *r* tomará valores en el complementario del anterior intervalo.

Los alumnos son sensibles a la relación existente entre el coeficiente de correlación y la dispersión de los datos, ya que más de la mitad de los sujetos de la muestra dominan estos hechos. No obstante, es pertinente manifestar que los estudiantes tienen más facilidad en conectar la dispersión de la nube de puntos con la intensidad de la dependencia que con el coeficiente de correlación. Además, parece más sencillo de comprender la correspondencia entre la intensidad de la dependencia y la dispersión de la nube de puntos que la existente entre el signo de la correlación y la pendiente de la recta de regresión. Pensamos que esto puede ser debido a las dificultades de los alumnos con la noción de pendiente de una recta (Azcárate, 1.990).

Los alumnos no son conscientes de los diferentes tipos de covariación que pueden dar lugar a la existencia de correlación (Barbancho, 1.973). El tipo más conocido por ellos es el de interdependencia, pues tres de cada cinco, aproximadamente, lo detectan, resultados que son coherentes con los de Tversky y Kahneman (1.982a). En particular, hemos encontrado que los alumnos confunden correlación y causalidad, uno de cada cuatro alumnos, confirmándose los resultados obtenidos en las investigaciones de Estepa (1.994) y Morris (1.997). Estos alumnos exhiben una concepción causal de la asociación (Estepa, 1.994) y consideramos que sería conveniente mostrarles ejemplos de otros tipos de covariación que originen correlaciones fuertes, tales como la concordancia o la interdependencia.

Finalmente, hemos observado que algunos alumnos confunden los coeficientes de correlación y de determinación.

Regresión

Los alumnos tienen notables dificultades para distinguir la variable explicativa de la variable explicada en la ecuación de la recta de regresión, ya que sólo uno de cada cuatro alumnos, aproximadamente, responden correctamente a este punto. Además, dos de cada cinco sujetos de la muestra se confunden de recta de regresión.

Hemos encontrado asimismo que dos de cada cinco alumnos interpretan convenientemente que si la pendiente de las rectas de regresión son idénticas, o sea, si existe correlación perfecta, el valor del coeficiente de correlación será 1 ó -1. Es pertinente subrayar que el valor 1, en solitario, es tenido en cuenta por casi el doble de los estudiantes que consideran el valor -1, en solitario. Esto pone en evidencia las dificultades que tienen los alumnos con la correlación negativa (Estepa, 1.994).

Dos de cada cinco estudiantes, aproximadamente, relacionan que ambas rectas de regresión son perpendiculares cuando el coeficiente de correlación es nulo. También establecen, mayoritariamente, la relación entre el tipo de correlación

y la amplitud del ángulo formado por ambas rectas de regresión, en particular, la correlación perfecta y una amplitud nula de dicho ángulo.

Entre todos los tipos posibles de ajuste, los alumnos consideran casi exclusivamente el lineal, posiblemente porque es el que han trabajado con mayor frecuencia durante la enseñanza. En concreto, casi la mitad de los sujetos de la muestra consideran que de una correlación positiva se deduce que la dependencia es lineal, y, aunque, en menor proporción, que una covarianza positiva implica una dependencia lineal. A pesar de que los estudiantes conocen otros tipos de ajustes, al ser más numerosas las aplicaciones que han hecho de este último tipo, el cual asimismo aparece de forma casi exclusiva en los textos de bachillerato (Sánchez Cobo, 1.996), generalizan excesivamente. Puesto que en la actualidad, el uso del ordenador en la enseñanza de la estadística resuelve los problemas de cálculo, abogamos, por tanto, por una mayor diversificación en los tipos de ajustes en los problemas que los alumnos tengan que resolver en la enseñanza universitaria, siempre que el tiempo disponible lo permita.

Capítulo 6

Actividades de traducción

6.1. INTRODUCCIÓN

En el Capítulo 4 presentábamos las diferentes representaciones que hemos establecido con respecto a la noción de la correlación: i) Descripción verbal, ii) tabla, iii) diagrama de dispersión, y, iv) coeficiente de correlación. Uno de los objetivos de esta investigación es el examen de la capacidad de estimación del coeficiente de correlación a partir de las representaciones anteriormente mencionadas, así como la traducción entre las mismas y estudiar las dificultades que pueden aflorar en las respuestas de los alumnos en procesos de tipo dinámico como éste. En este sentido, nuestro trabajo extiende el de Janvier (1.978,1.987) sobre funciones. Por otro lado, también extiende otros trabajos previos sobre estimación del coeficiente de correlación, por ejemplo las investigaciones de Chapman y Chapman (1.969) y Jennings, Amabile y Ross (1.982), que sólo usan representaciones numéricas o verbales. El diseño del cuestionario, que se describe en el Capítulo 4, permite, de igual modo, evaluar el efecto de la intensidad y el signo de la correlación, teorías previas, tipo de covariación y tipo de ajuste sobre la precisión de la estimación y las estrategias que los alumnos emplean en las

mismas. Esto es una aportación de la Tesis que completa estudios anteriores en el campo de la psicología.

El cuestionario que se ha empleado solicita a los alumnos la estimación de un coeficiente de correlación, el dibujo de un diagrama de dispersión o que indiquen una variable estadística bidimensional y que aporten una argumentación razonada de por qué han elegido esa respuesta. El análisis llevado a cabo en este capítulo, en las actividades de traducción a las representaciones de diagramas de dispersión o de coeficientes de correlación, se ha realizado en dos dimensiones: a) Sobre los coeficientes de correlación que los alumnos ofrecen -tareas 2,3 y 4- o sobre los coeficientes de correlación que hemos determinado a partir de los diagramas de dispersión que los estudiantes han dibujado -tareas 1 y 6-, b) sobre las estrategias usadas por los alumnos, que se deducían a partir de los argumentos que los sujetos exponían en sus respuestas.

En la sección 6.2 se estudia el efecto de cada una de las variables de tarea que se tuvieron en cuenta al diseñar el cuestionario, Tabla 4.2.3, sobre el error en valor absoluto de la estimación del valor absoluto del coeficiente de correlación en cada una de las tareas. Cuando dichas variables de tarea toman más de dos valores se ha llevado a cabo un análisis de la varianza; es el caso de la intensidad y el tipo de covariación. Si las variables de tarea son dicotómicas se ha realizado una comparación de muestras; es el caso del resto de las variables de tarea: Tipo de ajuste, existencia de teorías previas y tipo de dependencia.

En la sección 6.3 se estudia la forma en que los sujetos de la muestra encuentran la existencia o ausencia de asociación y su signo, así como la posible relación que los procedimientos utilizados pueden tener con el tipo de tarea propuesto. Para ello, hemos efectuado un análisis de correspondencias, donde se han tomado como filas cada una de las subtareas de las tareas 1, 2, 3 y 4 y como columnas las estrategias empleadas. Como variables suplementarias se han considerado las variables de tarea incluidas en el diseño experimental.

En la sección 6.4 se describen los procedimientos seguidos por los alumnos de la muestra acerca de la construcción de diagramas de dispersión.

En la sección 6.5 estudiamos las respuestas de los estudiantes cuando se les solicita un ejemplo de variable estadística bidimensional, a partir del valor de r

ofrecido en la actividad, que sea apropiado, tanto en signo como en intensidad (tarea 5 del cuestionario). Se ha analizado si la respuesta es una variable estadística bidimensional, si el signo de la situación propuesta es adecuado o no y si la dependencia en la situación es funcional, aleatoria o hay independencia.

Finalizamos el capítulo con las conclusiones del estudio sobre las tareas propuestas en el cuestionario.

6.2. ESTUDIO CUANTITATIVO DEL ERROR DE ESTIMACIÓN DEL COEFICIENTE DE CORRELACIÓN

En primer lugar, hemos tomado los valores estimados del coeficiente de correlación (tareas 2,3 y 4) por cada uno de los alumnos, o bien, hemos calculado el coeficiente de correlación de la nube de puntos que diseñan los alumnos (tareas 1 y 6). En este último caso, hemos utilizado el programa escrito en GWBASIC, que ya empleamos en la investigación de Sánchez Cobo (1.996). A partir de estos valores hemos hallado las diferencias absolutas entre el coeficiente de correlación normativo de cada subtarea y el coeficiente de correlación estimado por los alumnos, con lo que hemos obtenido el valor absoluto de los errores en la estimación del coeficiente de correlación. Este error, en valor absoluto, será la variable dependiente (la hemos denominado ERRORES), que analizaremos en este apartado para estudiar la posible influencia, sobre la misma, de las variables de tarea incluidas en el cuestionario. Tanto las estimaciones de los alumnos del coeficiente de correlación, como el coeficiente de correlación normativo se han multiplicado por 100, con lo que los datos que se refieren al coeficiente de correlación varían entre -100 y +100. En consecuencia, los errores absolutos variarán entre 0 y +100.

Como resumen de esta variable, en la Tabla 6.2.1 se muestra la media correspondiente a cada subtarea y las variables de tarea de cada una de ellas, según el diseño de la Tabla 4.2.3. Los datos mostrados corresponden a 97 alumnos que han respondido a todas las subtareas.

Al considerar el conjunto de todas las subtareas, se obtuvo una media global de las 25 subtareas de 28'73 con un error típico de 1'265 en la variable error absoluto de estimación.

Tabla 6.2.1. Media del error absoluto en las distintas tareas de traducción

			Intensidad (valor absoluto)					Tipo de ajuste		Tipo de covariación					Teorías previas		Tipo de dependencia	
Sub-tarea	Media	Error típico	0 0'1	0'1 0'35	0'35 0'65	0'65 0'9	0'9 1	Lineal	No lineal	Dep. c.uni.	Inter dep.	Dep ind.	Con cord	Cov cas.	Coin ciden	No coinc.	+	-
T1B	10'124	2'012					X		X	X							X	
T4C	12'832	1'552					X	X		X						X	X	
T1A	14'196	1'227				X			X		X							X
T6B	14'536	2'584					X											X
T6D	19'175	1'141				X											X	
T3D	19'393	2'003				X		X				X			X		X	
T4B	19'933	1'520			X				X				X		X		X	
T2C	21'585	5'328	X						X					X			X	
T2A	23'423	2'932					X	X			X						X	
T4D	24'598	2'458		X				X						X	X			X
T6E	26'841	1'646			X													X
T2E	28'208	2'626				X		X				X						X
T4A	30'855	4'374	X					X			X					X	X	
T2B	31'151	8'275		X				X					X					X
T3E	31'170	3'922		X				X					X		X		X	
T3C	32'307	2'862			X			X						X		X	X	
T6A	33'237	2'189		X													X	
T1C	34'608	1'344			X			X					X				X	
T3A	36'366	3'297					X		X	X					X			X
T2D	37'470	1'193			X				X	X							X	
T4E	37'712	7'630				X			X			X			X			X
T6C	39'361	2'823	X															X
T3B	39'709	4'663	X						X		X					X		X
T1E	42'186	3'628	X						X					X				
T1D	57'093	1'877		X				X				X						X

A partir de esta variable dependiente, hemos hecho una serie de análisis estadísticos respecto a las variables de tarea, que son las variables independientes en nuestro estudio. Nuestro diseño no nos permite contrastar el efecto conjunto de todas estas variables, pero si podemos realizar el estudio de los efectos principales y algunas interacciones de segundo orden mediante una serie de análisis que describimos a continuación.

6.2.1. EFECTO DE LA INTENSIDAD DE LA CORRELACIÓN Y EL TIPO DE TAREA

Inicialmente, se llevó a cabo un análisis de varianzas de medidas repetidas, empleando el paquete SPSS. Se tomó como variable dependiente el error absoluto de estimación y como factores la intensidad de la correlación en la tarea propuesta (variable que constaba de 5 niveles que representan los 5 intervalos considerados en el diseño del cuestionario) y el tipo de tarea (también con 5 niveles, ya que no se considera la tarea 5).

El diseño del cuestionario permite un diseño factorial completo para estas dos variables, de modo que, para cada alumno se dispone de 25 observaciones del error en valor absoluto que corresponden a las 25 combinaciones de los niveles del factor. En la Tabla 6.2.1.1 se presentan los resultados de la tabla de análisis de varianza.

Tabla 6.2.1.1. Resultados del análisis de varianza respecto a tipo de tarea e intensidad de la correlación

	Suma de cuadrados tipo	g.l.	Media cuadrática	F	Sig.
TAREA	16943'214	4	4235'804	3'122	0'015
INTENSIDAD	94312'760	4	23578'190	16'417	0'000
TAREA*INTENSIDAD	168693'759	16	10543'360	11'261	0'000
Error (TAREA*INTENSIDAD)	1438051'183	1.536	936'231		

De los resultados del análisis de varianza, presentados en la Tabla 6.2.1.1, se puede colegir que existe efecto significativo tanto respecto a la intensidad de correlación como al tipo de tarea, encontrándose, asimismo, interacción entre estos dos factores. No obstante, la diferencia entre tareas es menos nítida, pues al nivel 0'05 existe diferencia, pero al nivel 0'01 no existe diferencia.

Para analizar estos resultados presentamos, en las Tablas 6.2.1.2 y 6.2.1.3, los valores medios del error para cada nivel de estos dos factores, así como el error típico.

Las medias respecto a las tareas se dan en la Tabla 6.2.1.2, donde podemos observar que las tareas que dan menor valor absoluto de error son la 4 (estimación del coeficiente de correlación a partir de una nube de puntos) y su inversa, la tarea 6 (construir una nube de puntos a partir del coeficiente de correlación).

Las medias respecto a los intervalos de intensidad se dan en la Tabla 6.2.1.3, donde se puede observar que, conforme la intensidad de la asociación crece, en valor absoluto, el error absoluto en la estimación del coeficiente de correlación disminuye, lo que es coherente con lo expresado en otros análisis previos y en investigaciones anteriores, como, por ejemplo, la de Crocker (1.981).

Tabla 6.2.1.2. Media y error típico del factor tarea

TAREA	1	2	3	4	6
MEDIA	31'641	28'367	31'789	25'186	26'630
ERROR TÍPICO	0'923	2'650	2'264	2'200	1'164

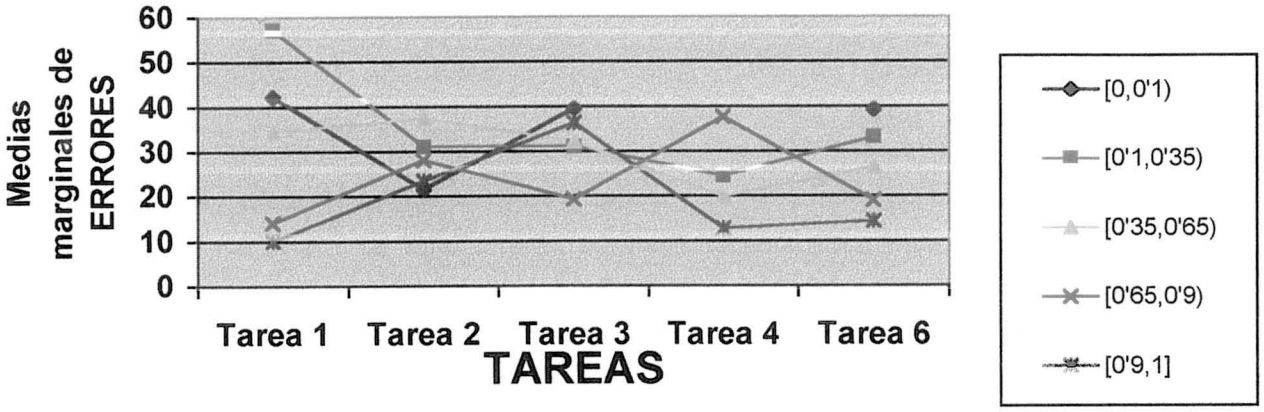
Tabla 6.2.1.3. Media y error típico del factor intensidad

INTENSIDAD	[0,0'1)	[0'1,0'35)	[0'35,0'65)	[0'65,0'9)	[0'9,1]
MEDIA	34'739	35'450	30'232	23'737	19'456
ERROR TÍPICO	2'644	2'774	0'901	1'674	1'245

Respecto a la interacción de los dos factores, los valores medios quedan representados en el Gráfico 6.1. Podemos observar como el efecto de la intensidad

no es homogéneo para las diversas tareas. Aunque, en general, a mayor intensidad hay menor error, para r = 1 el menor error se produce en las tareas 1, 4 y 6, mientras que para r = 0 se produce en la tarea 2.

Gráfico 6.1. Medias de ERROR de estimación según la intensidad y tipo de tarea



6.2.2. EFECTO DEL TIPO DE COVARIACIÓN

En segundo lugar, se llevó a cabo un análisis de varianzas de medidas repetidas, utilizando el paquete SPSS, tomándose como variable dependiente el error absoluto de estimación y como factores el tipo de covariación en la tarea propuesta (variable que constaba de 5 niveles) y el tipo de tarea (con 4 niveles, ya que en las tareas 5 y 6 no podemos considerar esta variable).

El diseño del cuestionario permite un diseño factorial completo para estas dos variables, de modo que, para cada alumno se dispone de 20 observaciones del error, en valor absoluto, que corresponden a las 20 combinaciones de los niveles del factor. En la Tabla 6.2.2.1 se presentan los resultados de la tabla de análisis de varianza.

Tabla 6.2.2.1. Resultados del análisis de varianza respecto a tipo de tarea y tipo de covariación

Fuente	Suma de cuadrados	g.l.	Media cuadrática	F	Sig.
TAREA	14791'741	3	4930'580	3'456	0'017
COVARIACIÓN	31275'332	4	7818'833	6'297	0'000
TAREA*COVARIACIÓN	214607'275	12	17883'940	16'264	0'000
Error (TAREA*COVARIACIÓN)	1359126'167	1.236	1099'617		

Tabla 6.2.2.2. Media y error típico del factor tipo de covariación

Tipo de covariación	Dependencia causal unilateral	Interde-pendencia	Concordancia	Dependencia casual	Dependencia indirecta
Media	24'384	26'879	29'290	30'383	35'934
ERROR TÍPICO	1'103	2'033	2'797	2'350	1'830

De nuevo podemos observar el efecto significativo de la tarea, tipo de covariación e interacción. Ocurre como en el análisis precedente, existe diferencia entre los diversos tipos de covariación y menos diferencia entre los distintos tipos de tareas.

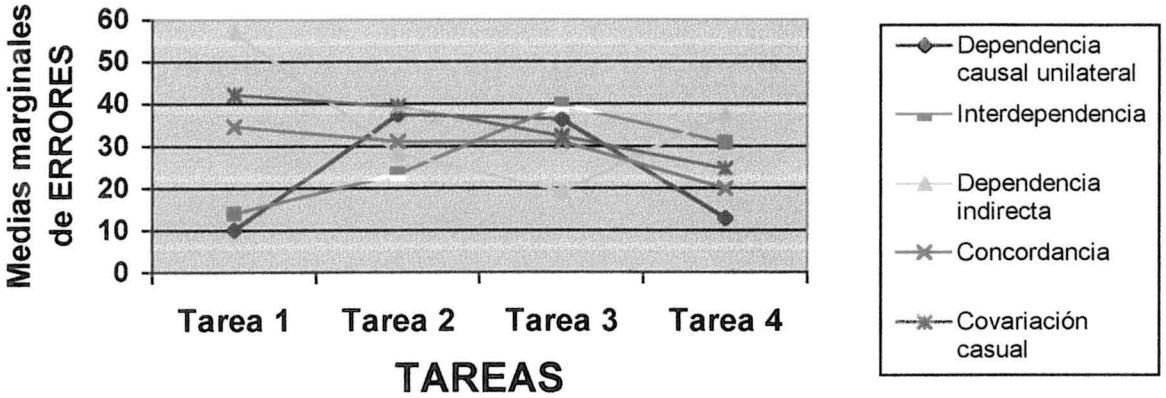
En cuanto al tipo de covariación, las medias del error absoluto de la estimación del coeficiente de correlación se dan en la Tabla 6.2.2.2, así como el error típico.

Los tipos de covariación donde el error absoluto en la estimación del coeficiente de correlación es más bajo son la dependencia causal unilateral e interdependencia, que son las que con más frecuencia aparecen en los ejercicios y ejemplos en los libros de texto (Sánchez Cobo, 1.996). Las otras tres aparecen menos en la bibliografía. Para analizar la interacción se presenta el Gráfico 6.2.

Observando el Gráfico 6.2, podemos inferir que los errores cometidos al estimar el coeficiente de correlación cuando existe concordancia o covariación casual, son poco sensibles al tipo de tarea, permaneciendo el error en un intervalo moderado. Sin embargo, los otros tres tipos de covariación son más sensibles al

tipo de tarea, especialmente este es el caso de la dependencia causal unilateral y la dependencia indirecta, que están contrapuestas según el tipo de tarea.

Gráfico 6.2. Medias de ERROR de estimación según el tipo de covariación y de tarea



6.2.3. EFECTO DEL TIPO DE AJUSTE

Para analizar el efecto del tipo de ajuste sobre el error de estimación del coeficiente de correlación, se tomó como variable dependiente el error medio cometido por cada alumno en el conjunto de tareas según el tipo (lineal o no lineal). El valor medio del error absoluto en la estimación del coeficiente de correlación cuando el ajuste es lineal es 31'524, con un error típico de 1'399. Cuando, en cambio, el ajuste es no lineal la media del error asciende a 28'981, siendo su error típico de 1'106.

Además, se ha obtenido un coeficiente de correlación entre las dos estimaciones de 0'529, siendo significativo, lo que indica que las estimaciones de los alumnos en los dos tipos de tareas están correlacionadas y esto nos faculta a utilizar la prueba T para muestras relacionadas, que hemos efectuado con el paquete estadístico SPSS, y cuyos resultados se presentan en la Tabla 6.2.3.1.

Tabla 6.2.3.1. Comparación de las muestra lineal y no lineal

	Media	Intervalo de confianza para la diferencia		t	g.l.	Sig. (bilateral)
		Inferior	Superior			
LINEAL-NO LINEAL	2'543	0'092	4'995	2'047	189	0'042

Se obtiene una diferencia muy pequeña en valor absoluto, aunque significativa, debido al tamaño de la muestra. A efectos prácticos, podemos aceptar la igualdad de las dos medias, lo que representa que no existe diferencia entre los errores absolutos de la estimación del coeficiente de correlación cuando la relación exhibida en la tarea es lineal o no lineal.

6.2.4. EFECTO DE LAS TEORÍAS PREVIAS

El mismo tipo de análisis se repitió para comparar los errores en las estimaciones del coeficiente de correlación de los alumnos en las tareas en que los datos coinciden o no coinciden con sus teorías previas. Se ha obtenido una media de 29'093, con un error típico de 1'571, para cuando las teorías previas están a favor, mientras que si las teorías previas están en contra se alcanzó una media de 29'071, con un error típico de 1'913. Entre las dos existe un coeficiente de correlación significativo de 0'574, por lo que realizamos una comparación de muestras relacionadas mediante la prueba T, cuyos resultados mostramos en la Tabla 6.2.4.1.

El contraste no es significativo, por lo que no se encuentra diferencia en el error de estimación del coeficiente de correlación según haya teorías previas a favor o en contra.

Tabla 6.2.4.1. Comparación de las muestras teorías previas a favor y en contra

	Media	Intervalo de confianza para la diferencia		t	g.l.	Sig. (bilateral)
		Inferior	Superior			
A FAVOR-EN CONTRA	0'022	-3'251	3'295	0'014	139	0'989

6.2.5. EFECTO DEL TIPO DE DEPENDENCIA

En cuanto al tipo de dependencia, hemos obtenido las siguientes medias de los errores absolutos en la estimación del coeficiente de correlación, con los errores típicos que se indican (directa: media = 26'426, error típico = 1'157; inversa: media = 31'705, error típico = 1'081). Además, hemos obtenido un coeficiente de correlación de 0'617. También, hemos realizado una comparación de muestras relacionadas aplicando, con el paquete estadístico SPSS, la prueba T para muestras relacionadas obteniéndose la Tabla 6.2.5.1.

Tabla 6.2.5.1. Comparación de los errores de estimación en las muestras directa e inversa

	Media	Intervalo de confianza para la diferencia		t	g.l.	Sig. (bilateral)
		Inferior	Superior			
DIRECTA-INVERSA	-5'279	-7'216	-3'341	-5'377	189	0'000

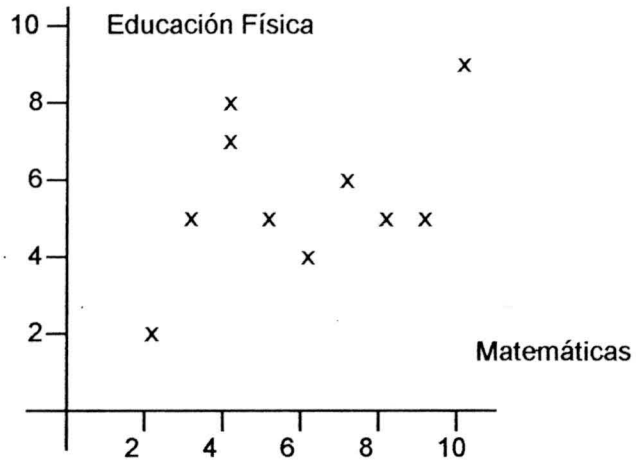
El contraste es significativo, lo que quiere expresar que existe diferencia al estimar el coeficiente de correlación entre una situación que presenta dependencia directa y otra que muestra dependencia inversa, siendo el valor absoluto del error más pequeño si se trata de una dependencia directa.

6.3. ANÁLISIS DE CORRESPONDENCIAS ENTRE TAREAS Y ESTRATEGIAS EMPLEADAS PARA ESTIMAR LA CORRELACIÓN

Uno de los objetivos de esta Tesis es el estudio de la resolución de tareas de traducción entre la descripción verbal, diagrama de dispersión, tablas y coeficiente de correlación de una variable estadística bidimensional. Desde el punto de vista didáctico, tiene especial interés el estudio de las estrategias que utilizan los alumnos cuando se enfrentan a un problema de traducción.

Además de analizar el error en valor absoluto del coeficiente de correlación estimado por los alumnos, hemos analizado las estrategias que estos emplean para estimar el signo y la intensidad del coeficiente de correlación. Los argumentos proporcionados por los alumnos se han clasificado en las siguientes categorías, que expresan sus estrategias implícitas:

- Estrategia 1 (E1). Los alumnos hacen referencia al signo de la correlación, no aportando más información de cómo han llegado a determinarlo. Por ejemplo, el sujeto 26 indica al responder a la tarea 1 apartado a): *"Varía con respecto al número de trabajadores"*. El sujeto 147 expresa cuando contesta a la tarea 1 apartado a): *"La relación entre pintores y tiempo es inversa"*. O el sujeto 31 manifiesta al responder a la tarea 1 apartado c) que: *"Existe dependencia entre las calificaciones de uno y otro profesor"*.
- Estrategia 2 (E2). Para detectar el signo de la correlación hacen referencia a alguna característica del diagrama de dispersión o lo dibujan, por ejemplo, a partir de la tabla de datos (marco gráfico). Así el sujeto 2 indica, al responder a la tarea 1 apartado a), que es una *"función decreciente"*. El sujeto 185 responde a la tarea 1 apartado e) diciendo que *"no es lineal"*. El sujeto 6 responde a la tarea 3 apartado c) dibujando la nube de puntos e indicando *"r = 0'5"*.



- Estrategia 3 (E3). Los alumnos se apoyan en algún hecho del marco numérico implícito. Por ejemplo, el sujeto 113, al contestar a la tarea 1 apartado a) indica: *"Cuantos más pintores haya, menos tiempo tardarán en pintar la habitación"*. El sujeto 16 responde a la tarea 3 apartado d): *"r = 0'8. Tiene relación una variable con otra a medida que aumenta una variable la otra aumenta con un incremento cada vez menor"*. El sujeto 28 dice, tarea 4 apartado a): *"r = 0'2. A mayor número de empates menor clasificación"*. Este alumno, a pesar de que indica que hay una dependencia inversa, luego valora el coeficiente de correlación con un signo positivo.
- Estrategia 4 (E4). Los alumnos utilizan varios de los argumentos de las estrategias anteriores en forma conjunta. Así el sujeto 185, al responder a la tarea 1 apartado d), dice: *"Cuanto más latitud haya menor temperatura habrá { ↑ latitud ↓ temperatura , relación indirecta decreciente }"*. Parece como si los sujetos de la muestra no fueran conscientes a que los hechos referidos al marco numérico implícito -*"cuanto más latitud haya menor temperatura habrá"*- son equivalentes a los referidos al marco gráfico -*"relación indirecta decreciente"*-, y por lo tanto lo que ofrecen es información redundante.
- Estrategia 5 (E5). Los alumnos sustentan sus argumentos en las teorías previas que poseen sobre el contexto del problema planteado. Así el sujeto 20 responde a la tarea 1 apartado b) de la siguiente manera: *"El tiempo depende de la memoria de cada persona y de la concentración que cada uno ponga. Todos no tenemos la memoria desarrollada igual, por lo tanto el tiempo es independiente del número de datos"*. El sujeto 85 responde a la tarea 2

apartado d) de la siguiente manera: *"r = -0'5. Los estudiantes hacen poco ejercicio y por tanto en una prueba de rendimiento físico no acabarían en buenas posiciones a no ser que haya alguno que pertenezca a algún club deportivo"*. El sujeto 97 expone, tarea 3 apartado e): *"r = -0'2. El entrenador B tiene en menor forma a su equipo"*. El sujeto 10 responde a la tarea 4 apartado e): *"r = -0'8. Inversa a más natalidad podrán comer menos cantidad de proteínas esto es lo que pasa en los países tercermundistas"*.

- Otras estrategias (OE). Esta categoría comprende todas aquellas estrategias que no se pueden incluir dentro de las anteriores. Por ejemplo, el sujeto 39 al responder a la tarea 1 apartado a) dice: *"Estas variables tienen relación funcional por tanto r = 1"*. El sujeto 37 al responder a la tarea 1 apartado b) indica: *"Si para (1 sola) persona aprender 10 números le cuesta 2 minutos, para aprender 50 le costará 10 minutos"*, o el sujeto 94 contesta a la misma pregunta exponiendo: *"En memorizar la lista máxima se tarda el tiempo máximo y se comprueba que hay una relación entre las variables de 5"*. Ambos alumnos tienen una concepción proporcional de la asociación. En particular este último al considerar que la lista máxima correspondería a la abscisa más grande (50), mientras que el tiempo máximo sería el de la mayor ordenada dada (10), ambos en los ejes de coordenadas ofrecidos en el cuestionario, concluye que la constante de proporcionalidad sería 5. Una respuesta interesante a la tarea 1 apartado e) es la ofrecida por el sujeto 10 cuando explica que: *"No represento nada porque creo que las calificaciones que se consigan no tienen nada que ver con la edad del alumno"*. Otra vez, algunos alumnos, si no hay dependencia, dan valores para el coeficiente de correlación fuera del intervalo [-1,1] , mientras que para la intensidad máxima indican que $r = 0$. Así, el sujeto 183 responde a la tarea 2 apartado b) de la siguiente forma: *"r = 2. No hay relación"*. Esta concepción errónea ya ha sido analizada en el Capítulo 5, sección 5.3, y emergerá, nuevamente, en el Capítulo 7, secciones 7.2.2 y 7.3.1.

Para cada una de las subtareas t1a, t1b, t1c, t1d, t1e, t2a, t2b, t2c, t2d, t2e, t3a, t3b, t3c, t3d, t3e, t4a, t4b, t4c, t4d, t4e, se han clasificado las respuestas de los

alumnos según las estrategias, obteniéndose la tabla de contingencia que se muestra en la Tabla 6.3.1. Solamente hemos tenido en cuenta las respuestas claras, de modo que las frecuencias obtenidas son pequeñas. No obstante, al calcular las frecuencias esperadas en esta tabla de contingencia se cumple las condiciones de aplicación del estadístico χ^2 ya que ninguna frecuencia esperada es menor que 1, siendo menos del 20 % de ellas menores que 5.

Tabla 6.3.1. Frecuencias absolutas de las estrategias observadas en las tareas 1, 2, 3 y 4

SUBTAREAS	ESTRATEGIAS						TOTAL
	E1	E2	E3	E4	E5	OE	
t1a	2	3	60	3	5	13	86
t1b	2	5	52	13	4	7	83
t1c	7	1	3	2	16	49	78
t1d	2	6	7	45	9	4	73
t1e	6	2	2	45	16	12	83
t2a	4	22	0	9	0	25	60
t2b	11	43	0	0	1	20	75
t2c	6	59	0	0	0	6	71
t2d	4	3	0	36	0	30	73
t2e	2	16	2	3	12	25	60
t3a	7	4	4	5	37	5	62
t3b	8	4	2	3	0	18	35
t3c	29	4	5	2	0	32	72
t3d	14	7	2	5	0	25	53
t3e	21	2	1	1	0	40	65
t4a	7	4	3	15	0	1	30
t4b	1	0	20	1	0	0	22
t4c	7	0	23	0	0	1	31
t4d	5	18	2	1	0	5	31
t4e	2	4	3	0	18	0	27
TOTAL	147	207	191	189	118	318	1.170

En el desarrollo que realizamos a continuación queremos estudiar las posibles relaciones existentes entre las estrategias utilizadas y las distintas tareas propuestas. En consecuencia, hemos considerado pertinente que un análisis de correspondencias es el procedimiento estadístico más adecuado. Este método estadístico permite visualizar las relaciones existentes entre las filas y las columnas de una tabla de contingencia (Lacasta y Brousseau, 1.995). Es uno de los métodos multivariantes factoriales y a partir de la distancia χ^2, definida sobre las dos nubes de puntos (fila y columna) que se obtienen mediante las frecuencias condicionales respecto a filas y columnas en la tabla de contingencia, permite determinar la dimensión del espacio vectorial definido por dichas distribuciones de puntos.

Una de las características más interesantes de este análisis es la posibilidad de representación conjunta de puntos fila y puntos columna. De este modo, la interpretación de los ejes factoriales se puede hacer, simultáneamente, sobre las filas y columnas de la tabla.

Tomando la Tabla 6.3.1 hemos llevado a cabo un análisis de correspondencias simples con el paquete estadístico BMDP.

El valor χ^2 -Tabla 6.3.2- es estadísticamente significativo, en consecuencia hay asociación entre las filas y las columnas, o sea, entre las tareas y subtareas y la estrategia empleada para dar la respuesta. Entonces las estrategias utilizadas dependen de las tareas y subtareas propuestas.

Tabla 6.3.2. Resultados del análisis de correspondencias

INERCIA TOTAL = SUMA DE LOS AUTOVALORES = 1'5354

EJE	AUTOVALOR	% DE INERCIA	% ACUMULADO	HISTOGRAMA
1	0'529	34'4	34'4	**********
2	0'393	25'6	60'0	**********
3	0'284	18'5	78'5	**********
4	0'267	17'4	95'9	**********
5	0'063	4'1	100'0	***

VALOR DE χ^2 CON 96 GL = 1796'37, P-VALOR = 0'000

Si observamos los autovalores -Tabla 6.3.2-, deducimos que nos encontramos ante un fenómeno no unidimensional, ya que no existe un único autovalor que acumule toda la inercia y tampoco la diferencia en porcentaje de explicación de la inercia entre los dos primeros es muy grande. Además, entre los tres primeros explican el 78'5 % de la inercia total. Por consiguiente, nos limitaremos a la interpretación de estos tres factores.

Tabla 6.3.3. Resultados del análisis de correspondencias (filas)

TAREA (Filas)	MASA	QLT	INERCIA	EJE 1		EJE 2		EJE 3	
				FACTOR	COR2	FACTOR	COR2	FACTOR	COR2
t1a	0'074	0'985	0'156	-1'346	0'856	0'519	0'127	-0'038	0'001
t1b	0'071	0'995	0'116	-1'214	0'903	0'312	0'060	-0'024	0'000
t1c	0'067	0'802	0'063	0'242	0'062	-0'248	0'065	0'124	0'016
t1d	0'062	0'998	0'101	-0'154	0'015	-0'946	0'554	0'046	0'001
t1e	0'071	0'996	0'093	-0'008	0'000	-1'051	0'840	0'017	0'000
t2a	0'051	0'849	0'029	0'607	0'642	0'245	0'105	0'118	0'024
t2b	0'064	1'000	0'083	0'800	0'495	0'728	0'410	-0'264	0'054
t2c	0'061	0'999	0'181	0'965	0'312	1'042	0'364	-0'687	0'158
t2d	0'062	0'942	0'072	0'216	0'040	-0'691	0'412	0'618	0'329
t2e	0'051	0'716	0'024	0'398	0'338	0'008	0'000	-0'319	0'218
t3a	0'053	0'983	0'146	-0'076	0'002	-0'826	0'248	-1'220	0'542
t3b	0'030	1'000	0'016	0'334	0'212	0'129	0'031	0'532	0'536
t3c	0'062	0'787	0'066	0'286	0'076	0'199	0'037	0'632	0'373
t3d	0'045	0'971	0'024	0'386	0'278	0'128	0'030	0'512	0'489
t3e	0'056	0'989	0'068	0'425	0'147	0'126	0'013	0'733	0'436
t4a	0'026	0'748	0'029	-0'008	0'000	-0'468	0'191	0'348	0'106
t4b	0'019	1'000	0'077	-1'882	0'866	0'728	0'130	0'015	0'000
t4c	0'026	0'939	0'074	-1'442	0'748	0'701	0'176	0'162	0'009
t4d	0'026	0'980	0'033	0'626	0'316	0'762	0'469	-0'303	0'074
t4e	0'023	0'990	0'084	-0'140	0'005	-0'671	0'123	-1'605	0'705

Tabla 6.3.4. Resultados del análisis de correspondencias (columnas)

ESTRATEGIA (Columnas)	MASA	QLT	INERCIA	EJE 1		EJE 2		EJE 3	
				FACTOR	COR2	FACTOR	COR2	FACTOR	COR2
E1	0'126	0'550	0'098	0'285	0'104	0'157	0'032	0'405	0'210
E2	0'177	1'000	0'309	0'779	0'348	0'768	0'338	-0'526	0'159
E3	0'163	1'000	0'424	-1'516	0'885	0'543	0'114	-0'026	0'000
E4	0'162	1'000	0'274	-0'073	0'003	-0'975	0'560	0'298	0'052
E5	0'101	1'000	0'276	-0'104	0'004	-0'910	0'302	-1'205	0'531
OE	0'272	0'879	0'154	0'353	0'220	0'018	0'001	0'441	0'344

Tabla 6.3.5. Columnas suplementarias

Subtareas	Intensidad (valor absoluto)					Tipo de ajuste		Tipo de dependencia		Tipo de covariación					Teorías previas	
	0 0'1	0'1 0'35	0'35 0'65	0'65 0'9	0'9 1	Lineal	No lineal	+	-	Dep. c.uni.	Inter dep.	Dep ind.	Con cord	Cov cas.	Coin ciden	No coinc.
T1A	0	0	0	1	0	0	1	0	1	0	1	0	0	0	-	-
T1B	0	0	0	0	1	0	1	1	0	1	0	0	0	0	-	-
T1C	0	0	1	0	0	1	0	1	0	0	0	0	1	0	-	-
T1D	0	1	0	0	0	1	0	0	1	0	0	1	0	0	-	-
T1E	1	0	0	0	0	0	1	0	0	0	0	0	0	1	-	-
T2A	0	0	0	0	1	1	0	1	0	0	1	0	0	0	-	-
T2B	0	1	0	0	0	1	0	0	1	0	0	0	1	0	-	-
T2C	1	0	0	0	0	0	1	1	0	0	0	0	0	1	-	-
T2D	0	0	1	0	0	0	1	1	0	1	0	0	0	0	-	-
T2E	0	0	0	1	0	1	0	0	1	0	0	1	0	0	-	-
T3A	0	0	0	0	1	0	1	0	1	1	0	0	0	0	1	0
T3B	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	1
T3C	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	1
T3D	0	0	0	1	0	1	0	1	0	0	0	1	0	0	1	0
T3E	0	1	0	0	0	1	0	1	0	0	0	0	1	0	1	0
T4A	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1
T4B	0	0	1	0	0	0	1	1	0	0	0	0	1	0	1	0
T4C	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	1
T4D	0	1	0	0	0	1	0	0	1	0	0	0	0	1	1	0
T4E	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0

Para favorecer la interpretación, se ha creído conveniente introducir como columnas suplementarias las variables de tarea que se incluyeron en el diseño, y que se muestran en la Tabla 4.2.3. Estas columnas suplementarias no intervienen en la determinación de los ejes factoriales. Sin embargo, una vez determinados éstos, se pueden proyectar sobre ellos para facilitar la interpretación. Debido a la propiedad baricéntrica, cada columna suplementaria se proyecta sobre la "media" de todas las filas que la definen, es decir, podemos interpretar estas columnas como el comportamiento promedio de las tareas caracterizadas por un cierto valor de una de las variables de tarea. Por ello, podremos relacionar los valores de las variables de tarea con las estrategias utilizadas por los alumnos. La calidad de la representación es alta en filas y columnas, excepto la estrategia 1 (E1), y columnas suplementarias.

EJE 1: OPOSICIÓN ENTRE RAZONAMIENTO NUMÉRICO Y GRÁFICO

Este eje opone algunas subtareas de las tareas 1 y 4 (parte negativa del eje), en concreto, [t1a (x = -1'346, r = 0'856); t1b (x = -1'214, r = 0'903); t4b (x = -1'882, r = 0'866); t4c (x = -1'442, r = 0'748)], a algunas subtareas de la tarea 2 (t2a, t2b, t2c y t2e), que aparecen en la parte positiva del eje, y que son las siguientes [t2a (x = 0'607, r = 0'642); t2b (x = 0'800, r = 0'495); t2c (x = 0'965, r = 0'312); t2e (x = 0'398, r = 0'338)], lo que nos indica que la estrategia de resolución sería distinta para cada grupo de subtareas. En efecto, si nos fijamos en las estrategias, en la parte negativa del eje aparece la estrategia E3 [marco gráfico] (x = -0'516, r = 0'885), asociada a las subtareas t1a, t1b, t4b y t4c, mientras que en la parte positiva aparece la estrategia E2 [marco numérico] (x = 0'779, r = 0'348), asociada a algunas subtareas de la tarea 2 (t2a, t2b, t2c y t2e).

Tabla 6.3.6. Resultados del análisis de correspondencias (columnas suplementarias)

Columnas suplementarias	QLT	EJE 1		EJE 2		EJE 3	
		FACTOR	COR2	FACTOR	COR2	FACTOR	COR2
I1	0'529	-0'371	0'242	0'013	0'000	-0'004	0'000
I2	0'676	-0'298	0'095	-0'205	0'045	-0'443	0'210
I3	0'940	-0'171	0'023	-0'401	0'129	-0'840	0'567
I4	0'977	-0'460	0'547	-0'239	0'148	-0'250	0'161
I5	0'919	-0'337	0'632	-0'079	0'034	-0'175	0'171
Lineal	0'650	-0'193	0'360	0'094	0'086	0'008	0'001
No Lineal	0'925	-0'589	0'891	-0'011	0'000	-0'103	0'027
Directa	0'963	-0'450	0'723	0'250	0'224	-0'007	0'000
Inversa	0'792	-0'441	0'578	-0'214	0'136	-0'128	0'049
Tcontra	0'936	-0'404	0'442	-0'018	0'001	-0'320	0'278
Favor	0'959	0'202	0'396	-0'011	0'001	0'163	0'258
Dcausal	0'997	-0'630	0'764	-0'053	0'005	-0'295	0'168
Inter	0'999	-0'533	0'510	0'081	0'012	-0'154	0'042
Dindir	0'926	-0'345	0'343	-0'297	0'255	-0'284	0'233
Concor	0'909	0'130	0'009	-0'629	0'211	-0'937	0'469
Casual	0'401	-0'237	0'065	-0'130	0'019	-0'186	0'040

Asimismo, esta diferenciación de tareas respecto a estas dos estrategias se puede observar en la tabla de porcentajes de distintas estrategias respecto a las subtareas -Tabla 6.3.7-, donde podemos observar que los porcentajes más altos de utilización de la estrategia E3 se dan en las subtareas t1a, t1b, t4b y t4c, mientras que los porcentajes más altos de uso de la estrategia E2 se dan en las subtareas t2a, t2b, t2c y t2e.

Así tenemos que para construir un diagrama de dispersión a partir de la descripción verbal, o bien, para estimar el coeficiente de correlación desde una nube de puntos, tareas 1 y 4, respectivamente, los alumnos tendrían preferencia por la estrategia E3, comprendida en el marco numérico implícito, consistente en examinar la variabilidad de ambas características comparando los valores numéricos hipotéticos de las mismas, presentando argumentaciones como, por

ejemplo, la siguiente: *"Cuántos más pintores haya, menos tiempo tardarán en pintar la habitación"* (sujeto 113, t1a).

Tabla 6.3.7. Porcentajes de las distintas estrategias según tarea

TAREAS	ESTRATEGIAS						TOTAL
	E1	E2	E3	E4	E5	OE	
t1a	1'4	1'4	31'4	1'6	4'2	4'1	7'4
t1b	1'4	2'4	27'2	6'9	3'4	2'2	7'1
t1c	4'8	0'5	1'6	1'1	13'6	15'4	6'7
t1d	1'4	2'9	3'7	23'8	7'6	1'3	6'2
t1e	4'1	1'0	1'0	23'8	13'6	3'8	7'1
t2a	2'7	10'6	0'0	4'8	0'0	7'9	5'1
t2b	7'5	20'8	0'0	0'0	0'8	6'3	6'4
t2c	4'1	28'5	0'0	0'0	0'0	1'9	6'1
t2d	2'7	1'4	0'0	19'0	0'0	9'4	6'2
t2e	1'4	7'7	1'0	1'6	10'2	7'9	5'1
t3a	4'8	1'9	2'1	2'6	31'4	1'6	5'3
t3b	5'4	1'9	1'0	1'6	0'0	5'7	3'0
t3c	19'7	1'9	2'6	1'1	0'0	10'1	6'2
t3d	9'5	3'4	1'0	2'6	0'0	7'9	4'5
t3e	14'3	1'0	0'5	0'5	0'0	12'6	5'6
t4a	4'8	1'9	1'6	7'9	0'0	0'3	2'6
t4b	0'7	0'0	10'5	0'5	0'0	0'0	1'9
t4c	4'8	0'0	12'0	0'0	0'0	0'3	2'6
t4d	3'4	8'7	1'0	0'5	0'0	1'6	2'6
t4e	1'4	1'9	1'6	0'0	15'3	0'0	2'3
TOTAL	100'0	100'0	100'0	100'0	100'0	100'0	100'0

Por el contrario, para estimar el coeficiente de correlación de dos variables presentadas mediante su descripción verbal (tarea 2), los alumnos emplean la estrategia E2, es decir, necesitan apoyarse en alguna característica del marco gráfico. Por lo tanto, en este caso no les es suficiente con la sola descripción verbal, sino que ellos mismos, traducen esta descripción verbal al marco gráfico, utilizando las características de éste para estimar el coeficiente de correlación.

Podríamos interpretar que los alumnos, en caso de que el problema no la proporcione, necesitan una imagen gráfica de la relación entre las variables, usando dicha visualización en sus respuestas. Así, el sujeto 8 responde a la tarea 2 apartado e) de la siguiente manera: *"r = 1. La relación es perfecta. Las rectas tienen la misma pendiente y una estaría sobre la otra"*, y el sujeto 48 indica que *"la relación es creciente"*.

En cuanto a las variables de tarea podemos observar que la dependencia directa (x = -0'450, r = 0'723) y el tipo de ajuste no lineal (x = -0'589, r = 0'891), en la parte negativa del eje, están asociadas a la estrategia E3 y a algunas subtareas de las tareas 1 y 4, en concreto, t1b, t4b y t4c tienen dependencia directa y t1a, t1b y t4b tienen un tipo de ajuste no lineal. Por lo que los alumnos requerirían, en mayor medida, efectuar comprobaciones y cálculos numéricos cuando la relación se explica de forma conveniente mediante un ajuste no lineal, lo cual parece lógico, debido a que el ajuste lineal es para ellos muy conocido.

En cuanto al tipo de covariación, la dependencia causal unilateral (dcausal x = -0'630, r = 0'764) y la interdependencia (interd x = -0'533, r = 0'510), aparecen asociadas a la parte negativa del eje, donde las subtareas referidas presentan este tipo de covariación; así, t1b y t4c tienen dependencia causal unilateral y t1a interdependencia. Esta clase de relaciones parecen prestarse más a los argumentos de tipo numérico por parte de los alumnos.

Una fuerte intensidad de la correlación presentada en la tarea está relacionada con el uso de la estrategia E3 [marco numérico], pues los intervalos de más alta intensidad de la correlación aparecen asociados a la parte negativa del eje I4 (x = -0'460, r = 0'547), I5 (x = -0'337, r = 0'632) igual que la estrategia E3.

En cuanto al resto de las variables de tarea consideradas en el análisis, no se observa una presencia destacada en este eje.

La representación gráfica del eje se puede observar en la Figura 6.1.

EJE 2: ARGUMENTOS GRÁFICOS FRENTE A TEORÍAS PREVIAS Y USO COMPLEMENTARIO DE ARGUMENTOS GRÁFICOS Y NUMÉRICOS

Opone las subtareas t2b (x = 0'728, r = 0'410), t2c (x = 1'042, r = 0'364), t4d (x = 0'762, r = 0'469) que se presentan en la parte positiva, a las t1d (x = -0'946, r = 0'554), t1e (x = -1'051, r = 0'840), en la parte negativa.

Se caracteriza por el uso de la estrategia E2 [marco gráfico] (x = 0'768, r = 0'338) en la parte positiva que se asocia a las subtareas t2b, t2c y t4d, frente a las estrategias E4 [marco gráfico y numérico] (x = -0'975, r = 0'560) y E5 [teorías previas] (x = -0'910, r = 0'302), que aparecen en la parte negativa del eje, y se asocian, por tanto, a las subtareas t1d y t1e. Esta asociación queda de manifiesto estudiando los porcentajes de la Tabla 6.3.7, donde se puede observar que los porcentajes mayores de cada estrategia se corresponden con las subtareas a las que se asocian en este eje.

En estas últimas tareas, los alumnos, para reforzar sus razonamientos, utilizan conjuntamente argumentaciones propias de los marcos numérico y gráfico, o bien sus teorías previas. Por consiguiente, para ellos no es suficiente el uso de una estrategia en solitario, teniendo que fortalecer sus argumentos con la utilización de una pluralidad de ellas.

En cuanto a las variables de tarea, las correlaciones con el mismo son muy bajas. Hay una leve correlación del tipo de dependencia (directa e inversa) que coincide con el presentado en las subtareas de este eje; t2c presentan correlación directa t2b y t4d correlación inversa.

Existe, no obstante, una pequeña asociación de la dependencia indirecta y concordancia con la parte negativa del eje, que es el tipo de covariación que estas subtareas presentan en el diseño del experimento, Tabla 4.2.3.

Figura 6.1. Representación gráfica del eje 1

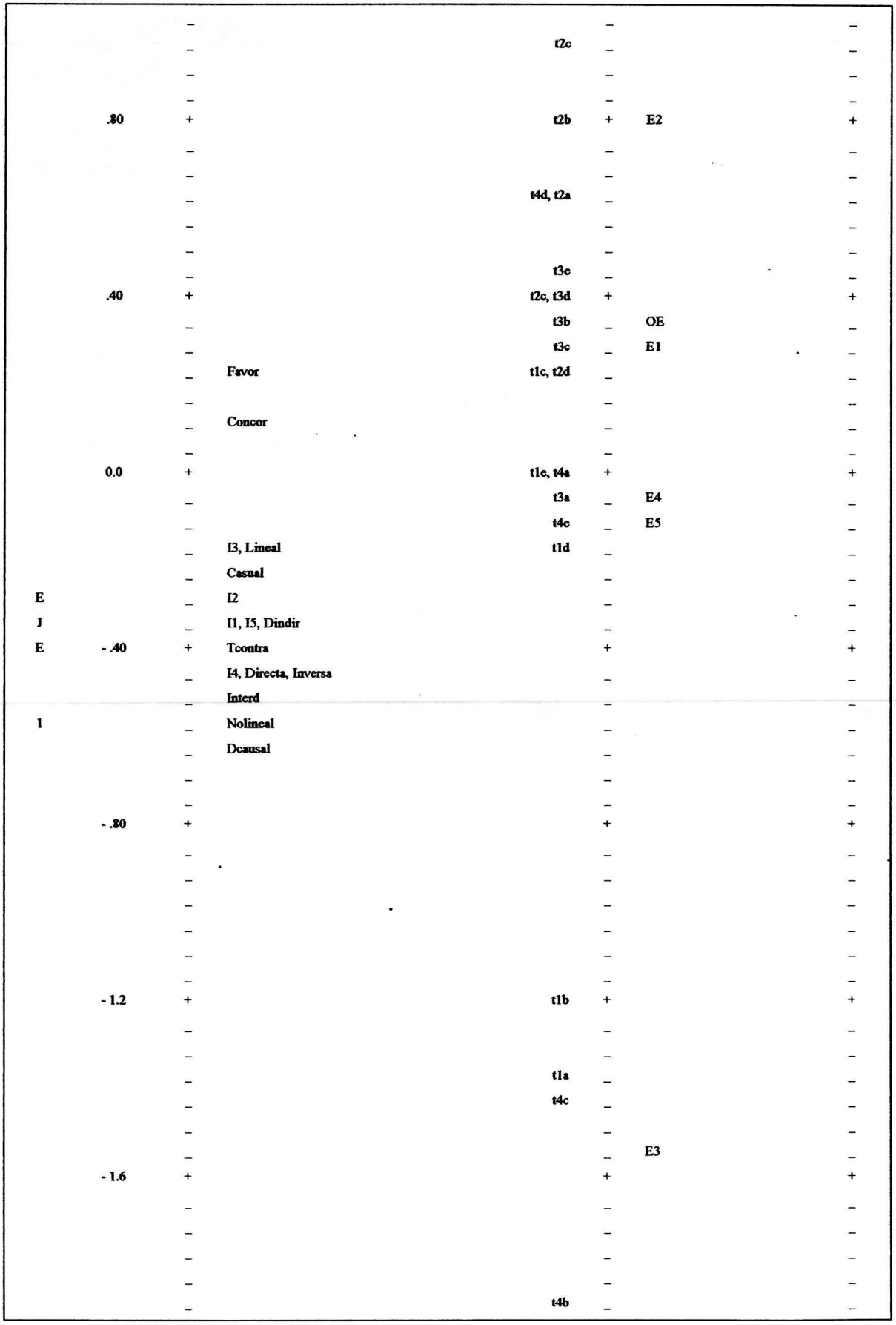
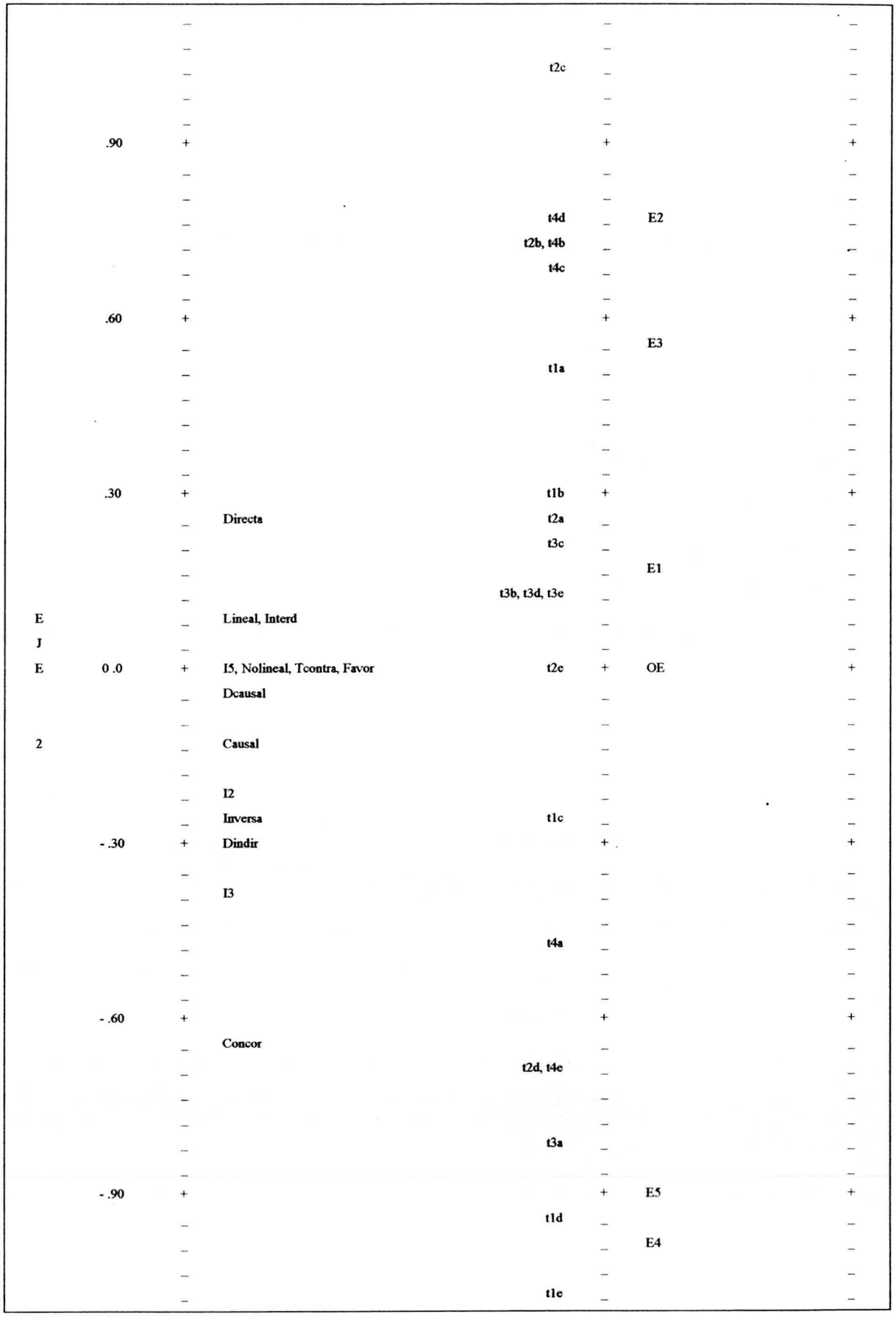


Figura 6.2. Representación gráfica del eje 2



EJE 3: PESO DE LAS TEORÍAS PREVIAS EN LAS ESTRATEGIAS DE LOS ALUMNOS

Opone las subtareas t2d (x = 0'618, r = 0'329); t3b (x = 0'532, r = 0'536); t3c (x = 0'632, r = 0'373); t3d (x = 0'512, r = 0'489); t3e (x = 0'733, r = 0'436) en la parte positiva, a las tareas t3a (x = -1'220, r = 0'542); t4e (x = -1'605, r = 0'705) en la parte negativa. Se caracteriza por el uso de la estrategia E5 (x = -1'205, r = 0'531), asociada a la parte negativa, que se opone a otras estrategias OE (x = 0'441, r = 0'344) y a la estrategia E1 (x = 0'405, r = 0'210), asociadas a la parte positiva. Es decir, el uso de las teorías previas frente al uso de otras estrategias, generalmente de tipo proporcional, o bien a la falta de argumentación para justificar la correlación que se da como resultado. Esta asociación entre subtareas y estrategias se puede observar en la Tabla 6.3.7, donde, en general, las estrategias mencionadas presentan un mayor porcentaje de uso en las subtareas con las que hemos señalado que se relacionan.

En consecuencia, cuando los alumnos se enfrentan a un problema de estimación del valor del coeficiente de correlación a partir de una tabla de valores contextualizados, en lugar de utilizarlos para valorar al coeficiente de correlación, mayoritariamente, prefieren utilizar sus teorías previas sobre el contexto de los datos ofrecidos (t3a), o bien dan la respuesta sin argumentación o usan otro tipo de estrategias. Juzgamos que esto es debido a que los alumnos emplean el marco numérico cuando ven una variación conjunta uniforme entre las variables -si una aumenta la otra también lo hace, si una aumenta la otra disminuye, etc,-.

Ahora bien, cuando el alumno observa que en la muestra ofrecida hay tanto rachas en las que si una variable aumenta la otra aumenta, como rachas en las que si una variable aumenta la otra disminuye, sufre un conflicto cognitivo, recurriendo a las estrategias E1, E5 y OE. Así el sujeto 26 aporta como argumentación a su repuesta a la tarea 3 apartado b):

$$\begin{array}{cccccc} & 1\,2\,2 & & 3\,4 & 4 & 5\,6\,7\,7 \\ " & \underbrace{3\,7\,6} & & \underbrace{2\,2} & \underbrace{7} & \underbrace{4\,5\,3\,5} \quad " \\ & r<1 & r>1 & & r<1 & r>1 \end{array}$$

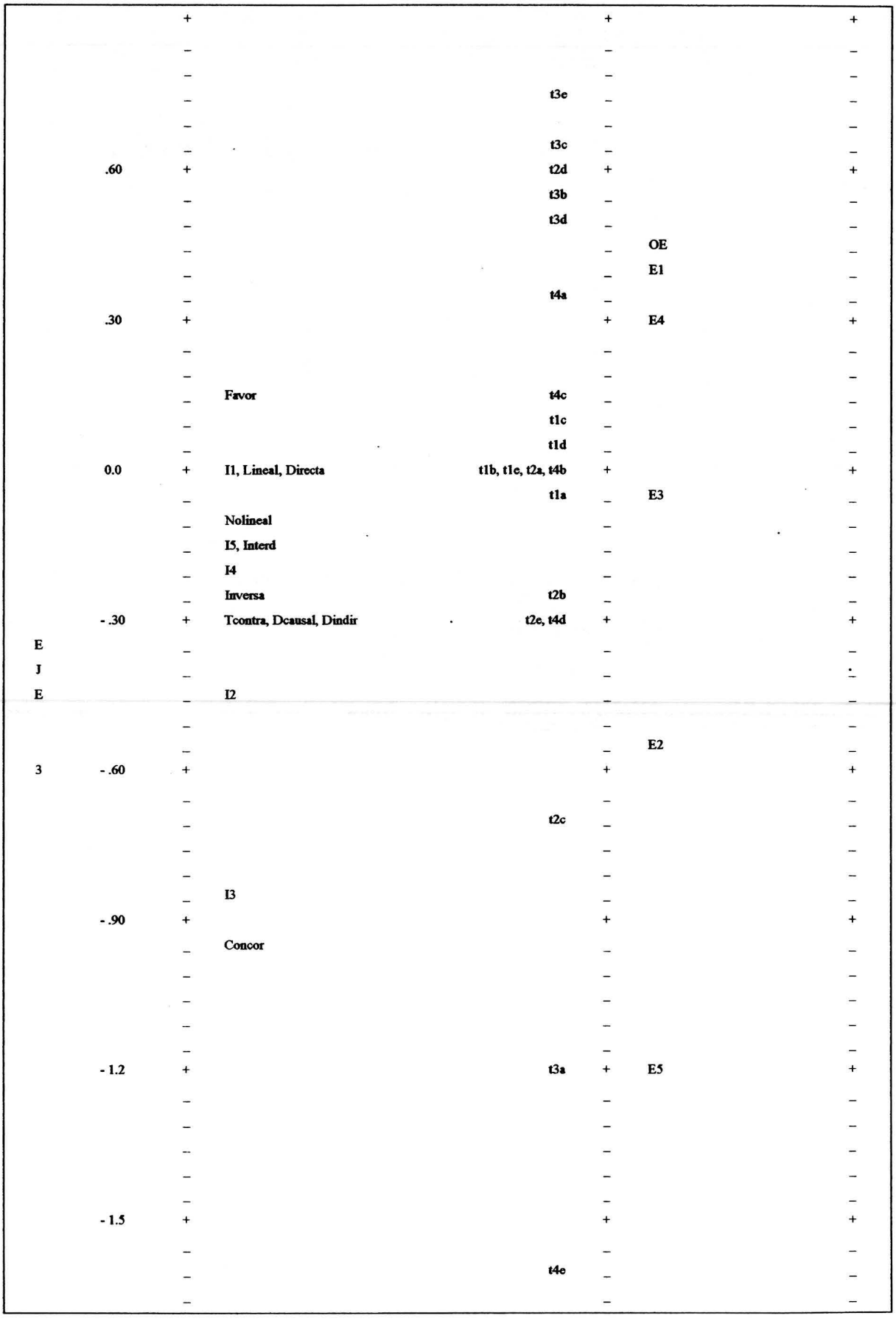
Nótese que, además, este alumno tiene la creencia errónea de que el coeficiente de correlación es la razón de los valores de las componentes de la variable estadística bidimensional.

En cuanto a las variables de tarea, utilizadas como variables suplementarias, se puede observar que tienen presencia destacada en este eje el intervalo [0'35,0'65) (x = -0'840, r = 0'567) de la intensidad de correlación, que se asocia a la parte negativa y, por tanto, al uso de las teorías previas, lo que concuerda con lo expuesto en el párrafo anterior, ya que, según hemos visto en la sección 6.2.1, los alumnos estiman bien las correlaciones altas o la independencia, pero sus estimaciones pierden seguridad cuando la intensidad de la correlación es intermedia (Shaklee y Mims, 1.982; Alloy y Tabachnik, 1.984).

Estos resultados son consistentes con otras investigaciones que han mostrado que los juicios de los sujetos son sensibles a la diferente intensidad de la correlación (Beach y Scopp, 1.966; Erlick y Mills, 1.967; Berman y Kenny, 1.976; Hamilton y Rose, 1.980; Jennings y cols., 1.982). Por lo tanto, es de esperar que los alumnos empleen sus teorías previas cuando no tienen seguridad en la estimación de la correlación.

Asimismo, destaca en este eje el tipo de covariación concordancia concor(x = -0'937, r = 0'469), que nos muestra asociación entre el uso de teorías previas y concordancia, hecho que ya fue observado en la tesis de Estepa (1.994). Además, los items en que las teorías previas coinciden con las de los alumnos (tfavor) se asocian a la parte positiva (otras estrategias), mientras que los que no coinciden (tcontra) se asocian a la parte negativa (uso de teorías previas). Por consiguiente, los alumnos usan, preferentemente, sus teorías previas cuando éstas no coinciden con los datos del problema.

Figura 6.3. Representación gráfica del eje 3



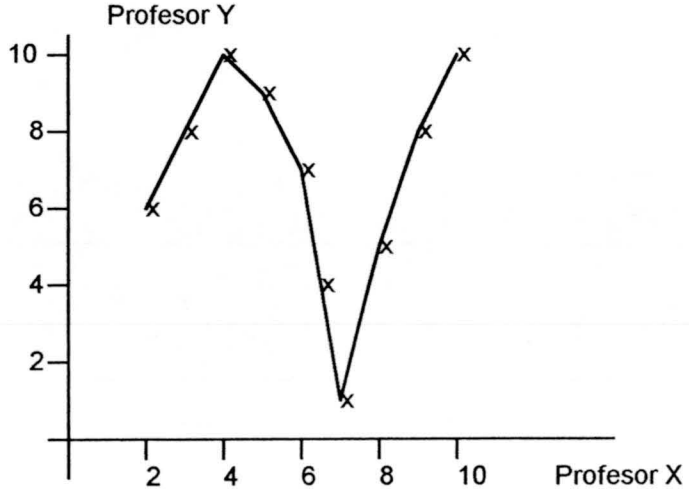
6.4. CONSTRUCCIÓN DE DIAGRAMAS DE DISPERSIÓN

La dependencia aleatoria es una extensión de la dependencia funcional. Los alumnos poseen nociones más o menos profundas sobre ésta última. Al plantearles situaciones nuevas, donde los conocimientos que tienen sobre la dependencia funcional se manifiestan incompletos, intentan, entonces, adaptarlos para resolver el problema propuesto.

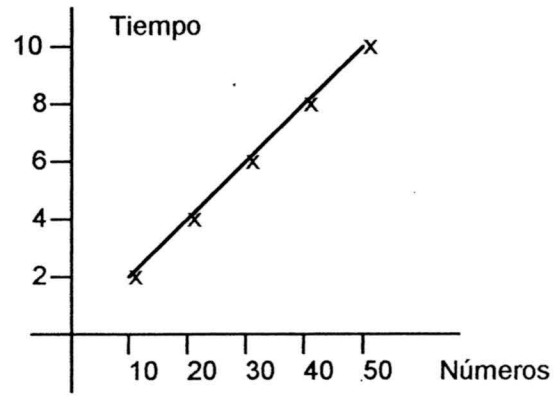
Cuando al alumno se le plantea el problema de construir un diagrama de dispersión a partir de la descripción verbal de la variable bidimensional (tarea 1) o a partir del coeficiente de correlación (tarea 6), ha usado las nociones que posee sobre la dependencia funcional modificadas y adaptadas para que le sean útiles en la resolución de la situación presente.

En esta adaptación, naturalmente, entrarán en juego las nociones que se poseen sobre la dependencia funcional y las adquiridas sobre la dependencia aleatoria, recibidas en el proceso de enseñanza. El resultado de esta conjunción se puede resumir en las cuatro estrategias que hemos observado en la construcción de los diagramas de dispersión:

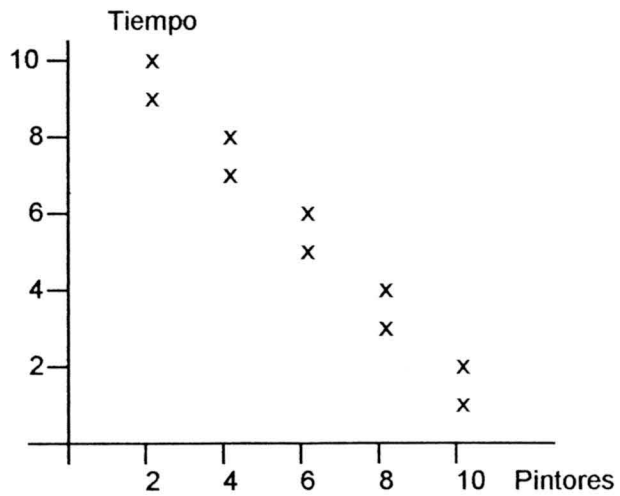
- Estrategia 1 (E1). Un sujeto usa esta estrategia cuando dibuja un diagrama de dispersión y une los puntos del mismo mediante una línea poligonal. Por ejemplo, el sujeto 24 responde de la siguiente forma a la tarea 1 apartado c) :



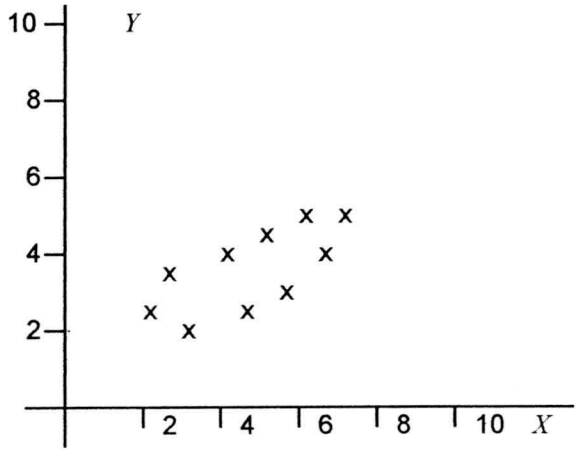
- Estrategia 2 (E2). Ahora el sujeto une los puntos mediante una línea, pero el diagrama de dispersión muestra una dependencia funcional perfecta, aunque ésta no corresponda a la situación demandada. Podría haber subyacente una concepción determinista en estos alumnos, ya que emplean un ajuste que caracteriza a una relación lineal perfecta, aunque esto es difícil de deducir pues no hemos realizado entrevistas. Por ejemplo, el sujeto 15 responde a la tarea 1 apartado b) :

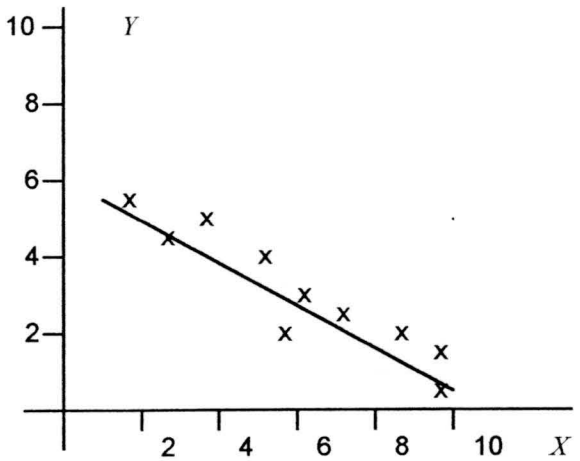


- Estrategia 3 (E3). Los sujetos recurren a un tipo de diagrama de dispersión muy peculiar con puntos en paralelo, que parece indicar que se desprenden de ciertas características de las funciones, como, verbigracia, que a una misma abscisa le correspondan diversos valores de la ordenada. Pero una tal constelación de puntos es difícil que represente, en general, a una variable estadística bidimensional. Por ejemplo, el sujeto 106 responde a la tarea 1 apartado a) :



- Estrategia 4 (E4). Los sujetos han diseñando un diagrama de dispersión típico de las situaciones de dependencia aleatoria. Hemos observado que en esta estrategia existen como dos variantes: i) El alumno, al dibujar la nube de puntos, parece sólo prestar atención al signo de la correlación y distribuye los puntos de forma aleatoria, y, ii) el alumno, previamente a dibujar el diagrama de dispersión, traza una recta a la que ajusta los puntos. Como ejemplos sirven las respuestas del alumno 125 (tarea 6 apartado d) y el alumno 23 (tarea 6 apartado e), respectivamente:





Apoyándonos en la taxonomía anterior, hemos cruzado las diversas estrategias que exhiben los sujetos de la muestra, a partir de los diagramas de dispersión que han dibujado, con todos los apartados de las tareas 1 y 6, habiendo obtenidos los resultados que se presentan en la Tabla 6.4.1.

Tabla 6.4.1. Frecuencia absoluta y porcentaje de las estrategias en el dibujo de diagramas de dispersión en cada una de las subtareas de las tareas 1 y 6

TAREA		ESTRATEGIAS				Total
		E1	E2	E3	E4	
1	a)	4 2'2	66 36'1	33 18'0	80 43'7	183 100'0
	b)	6 3'4	61 34'5	26 14'7	84 47'5	177 100'0
	c)	5 3'3	37 23'6	18 11'5	97 61'8	157 100'0
	d)	6 3'6	55 33'1	20 12'1	85 51'2	166 100'0
	e)	7 4'1	23 13'5	10 5'9	130 76'5	170 100'0
T. parcial		28 3'3	242 28'4	107 12'5	476 55'8	853 100'0
6	a)	1 0'6	4 2'5	7 4'4	146 92'4	158 100'0
	b)	0 0'0	80 49'1	16 9'8	67 41'1	163 100'0
	c)	1 0'6	6 3'7	11 6'8	138 85'7	156 100'0
	d)	2 1'2	6 3'7	20 12'4	133 82'6	161 100'0
	e)	1 0'7	6 3'9	10 6'6	135 88'8	152 100'0
T. parcial		5 0'6	102 12'9	64 8'1	619 78'3	790 100'0
Total		33 2'0	344 20'9	171 10'4	1095 66'6	1643 100'0

De la Tabla 6.4.1 se infiere que en las actividades de traducción a diagrama de dispersión desde una representación verbal (variable estadística bidimensional), tarea 1, los sujetos tienden, en mayor proporción, a dibujar nubes de puntos que representen dependencias funcionales que si parten de un valor del coeficiente de correlación, tarea 6, (tarea 1: E1 = 3'28 %, E2 = 28'37 %, frente a la tarea 6: E1 = 0'63 %, E2 = 12,91 %). En cambio, en la tarea 6 diseñan más diagramas de dispersión propios de una dependencia aleatoria que en la tarea 1, pues en la tarea

6 hay un 78'36 % de nubes de puntos de este tipo frente a la tarea 1 donde hay un 55'81 %.

Estos resultados parecen naturales ya que en la tarea 1 los alumnos, únicamente, cuentan con sus teorías previas para estimar el grado de relación entre las características y, por tanto, los juicios sobre covariación estarían influidos por las expectativas de los sujetos acerca de la asociación entre las variables en cuestión, sobreestimando las correlaciones que ellos esperaban encontrar y subestimándola en caso contrario (Berman y Kenny, 1.976; Chapman, 1.967; Chapman y Chapman, 1.967; Hamilton y Rose, 1.980; Trolier y Hamilton, 1.986). En la tarea 6 la información que se ofrece está descontextualizada obviando, por tanto, que semejantes espectativas sean posibles, llegándose a resultados similares a los alcanzados en algunos estudios en los que los juicios sobre correlación se presentaban a través de variables no contextualizadas, tales como pares de números (Erlick y Mills, 1.967; Jennings y cols., 1.982).

Asimismo, puede deducirse de la Tabla 6.4.1, que en todos los apartados de las dos tareas, salvo en el b) de la tarea 6, los alumnos representan, de forma mayoritaria, diagramas de dispersión correspondientes a una dependencia aleatoria entre las variables. Es interesante, sin embargo, subrayar que en el apartado b) de la tarea 6 se les ofrecía a los alumnos un coeficiente de correlación r = -1, y, a pesar de ello, únicamente 80 estudiantes, aproximadamente un 49'1 % de los alumnos que responden, han sabido dibujar una nube de puntos adecuada.

Estrategias y variables de tarea

A continuación, estudiamos la frecuencia y porcentaje de las diversas estrategias descritas con anterioridad, teniendo en cuenta las variables de tarea que se han considerado en el diseño que se muestra en la Tabla 4.2.3. Para ello nos basaremos en la Tabla 6.4.2.

Al cruzar los gráficos con los subintervalos de la intensidad del coeficiente de correlación, en valor absoluto, en todos los casos el mayor porcentaje se obtiene para la estrategia E4, pero es de destacar que en el subintervalo [0'9,1], se da

prácticamente un empate entre las estrategias E2 y E4, a pesar de lo expresado en el parágrafo anterior.

Tabla 6.4.2. Frecuencia absoluta y porcentaje de las estrategias en cada una de las subtareas de las tareas 1 y 6

VARIABLES		ESTRATEGIAS				
		E1	E2	E3	E4	Total
Intensidad	[0,0'1)	8 2'5	29 8'9	21 6'4	268 82'2	326 100'0
	[0'1,0'35)	7 2'2	59 18'2	27 8'3	231 71'3	324 100'0
	[0'35,0'65)	6 1'9	43 13'9	28 9'1	232 75'1	309 100'0
	[0'65,0'9)	6 1'8	72 20'9	53 15'4	213 61'9	344 100'0
	[0'9,1]	6 1'8	141 41'5	42 12'3	151 44'4	340 100'0
Tipo de ajuste	Lineal	11 3'4	92 28'5	38 11'8	182 56'3	323 100'0
	No lineal	17 3'2	150 28'3	69 13'0	294 55'5	530 100'0
Tipo de covariación	D.causal unilateral	6 3'4	61 34'5	26 14'7	84 47'4	177 100'0
	Interdepen-dencia	4 2'2	66 36'1	33 18'0	80 43'7	183 100'0
	Dependen. indirecta	6 3'6	55 33'1	20 12'1	85 51'2	166 100'0
	Concor-dancia	5 3'2	37 23'6	18 11'4	97 61'8	157 100'0
	Covariación casual	7 4'1	23 13'5	10 5'9	130 76'5	170 100'0
Signo	+	14 2'1	108 16'5	71 10'9	460 70'5	653 100'0
	-	12 1'5	213 26'0	90 11'0	505 61'5	820 100'0

Si cruzamos los gráficos con el tipo de ajuste sigue manteniéndose un mayor porcentaje para las nubes de puntos que muestran una dependencia aleatoria, pero no existe prácticamente diferencia entre el ajuste no lineal y lineal.

Cuando cruzamos las estrategias con el tipo de covariación, se observa que las representaciones de una dependencia aleatoria alcanzan su mayor porcentaje cuando se presenta una covariación casual, mientras que el menor es cuando hay interdependencia. Para las gráficas de tipo determinista, por el contrario, el máximo es cuando hay interdependencia y el mínimo es cuando hay covariación casual. Estos resultados parecen bastante propios, pues de los tipos de covariación (Barbancho, 1.973) la interdependencia comparte una característica muy importante con la dependencia funcional, la mutua relación entre las variables. Por otra parte, según Steinbring (1.991), el azar puede considerarse como la regla que explica aquellos efectos para los que desconocemos su causa, y dado que la covariación casual expresa que la conexión entre causa y efecto es debida al azar, parece natural su concatenación.

Si cruzamos los gráficos y el signo de la correlación de cada subtarea de las tareas 1 y 6, el porcentaje de los gráficos que muestran una dependencia aleatoria es mayor cuando el signo de la relación es positivo que cuando es negativo. Lo contrario ocurre con la dependencia determinista.

Por último, al observar el modo en que los alúmnos construyen la nube de puntos, hemos encontrado que no son sensibles a considerar la variable discreta como tal. En consecuencia, cuando dibujan el diagrama de dispersión toman las variables discretas como si fuesen continuas, haciéndoles dibujar los puntos del diagrama de forma aleatoria, lo cual, en ciertos casos, no correspondería a la situación real reflejada en el contexto del problema, por ejemplo, 2'5 pintores.

6.5. TRADUCCIÓN DEL COEFICIENTE DE CORRELACIÓN A UNA DESCRIPCIÓN VERBAL

En esta sección se exponen los resultados obtenidos en las actividades de traducción del coeficiente de correlación a una descripción verbal -variable estadística bidimensional-, es decir, las correspondientes a la tarea 5. Para ello hemos tomado en consideración si la variable indicada por el alumno es una variable estadística bidimensional o no, o si la presenta a través de un diagrama de

dispersión -marco gráfico- o a través de una tabla de datos -marco numérico-. Otra característica analizada es si el signo de la asociación entre las componentes de la variable estadística bidimensional ofrecida por el alumno coincide o no con él de la tarea. Además hemos estudiado si la dependencia entre las componentes de la variable estadística bidimensional aportada por el alumno es funcional, aleatoria o no hay relación entre ellas.

Como puede advertirse en la Tabla 6.5.1, en todos las subtareas de la tarea 5 los alumnos han respondido, de forma altamente mayoritaria, con una variable estadística bidimensional pertinente. Únicamente, en el apartado c) se alcanza un porcentaje de respuestas, algo significativo, que no son variables estadísticas bidimensionales. De forma muy minoritaria los alumnos responden ofreciendo una nube de puntos o una tabla de datos. En este último caso, consideramos que podría suceder que el alumno entienda que la actividad muestra el proceso inverso al de los ejercicios habituales que ellos han realizado en cursos anteriores, y que se han analizado en la investigación de Sánchez Cobo (1.996), en el que dado una tabla de datos ellos debían obtener el coeficiente de correlación.

En cuanto al signo, también podemos comprobar que en todas las subtareas, excepto en la b), él deducido de la variable estadística bidimensional denotada por el alumno, coincide con el signo correspondiente del valor del coeficiente de correlación dado en ella. Es conveniente subrayar que la coincidencia es netamente resaltada por los porcentajes en el caso de una relación positiva. Por el contrario, si el valor de r es negativo, los alumnos encuentran dificultades para precisar una variable estadística bidimensional de signo semejante.

Finalmente, si examinamos el tipo de dependencia, vemos que la dependencia funcional, que es la que se presenta en la subtarea a), sólo ha sido detectada por 30 estudiantes de 147 que responden, lo que equivale a un 20'4 %. En el caso de independencia, que es la correspondiente a la subtarea c), presentan una respuesta acertada 72 alumnos de 124 que responden (58'1 %), mientras que 46 alumnos de estos 124 (37'1 %) dan una dependencia aleatoria. Cuando la dependencia es aleatoria -subtareas b), d) y e)-, las variables estadísticas bidimensionales dadas por los alumnos concuerdan, en un alto

porcentaje -en los tres casos por encima del 70 % sobre respuestas ofrecidas-, con lo solicitado en la subtarea.

Tabla 6.5.1. Análisis de la tarea 5

TAREA 5		SUBTAREAS				
		a)	b)	c)	d)	e)
Variable Estadística Bidimensional	Si	134 69'4	93 48'2	94 48'7	106 54'9	99 51'3
	No	8 4'1	13 6'7	24 12'4	16 8'3	15 7'8
	Gráfica	1 0'5	1 0'5	1 0'5	1 0'5	2 1'0
	Tabla	4 2'1	4 2'1	5 2'6	5 2'6	5 2'6
	No responden	46 23'9	82 42'5	69 35'8	65 33'7	72 37'3
Total		193 100'0	193 100'0	193 100'0	193 100'0	193 100'0
Signo	Coincide	125 64'8	46 23'8	84 43'5	69 35'6	77 39'9
	No coincide	14 7'2	53 27'5	17 8'8	43 22'4	29 15'0
	No tiene	8 4'1	12 6'2	23 11'9	16 8'3	15 7'8
	No responden	46 23'9	82 42'5	69 35'8	65 33'7	72 37'3
Total		193 100'0	193 100'0	193 100'0	193 100'0	193 100'0
Tipo de dependencia	Funcional	30 15'5	5 2'6	6 3'1	9 4'7	6 3'1
	Aleatoria	107 55'4	79 40'9	46 23'8	92 47'6	87 45'1
	Independencia	10 5'2	26 13'5	72 37'3	27 14'0	27 14'0
	No responden	46 23'9	83 43'0	69 35'8	65 33'7	73 37'8
Total		193 100'0	193 100'0	193 100'0	193 100'0	193 100'0

6.7. CONCLUSIONES SOBRE LAS ACTIVIDADES DE TRADUCCIÓN

En este capítulo hemos analizado el valor absoluto del error de estimación del coeficiente de correlación a partir de diversas representaciones (verbal, gráfica, numérica y tabla), así como las estrategias de los alumnos en la traducción entre diferentes tipos de representación. Nuestro trabajo completa de esta manera los de Janvier respecto a la dependencia funcional y otros trabajos sobre estimación de la correlación, llevados a efecto en el campo de la psicología.

El estudio se ha llevado a cabo desde un punto de vista cuantitativo y cualitativo, evaluando, asimismo, el efecto de las diversas variables de tarea, de acuerdo con el diseño expuesto en el Capítulo 4. Una primera conclusión es que la actividad solicitada es compleja y no todos los estudiantes actúan de modo similar al enfrentarse con esta clase de problemas. Las conclusiones obtenidas las podemos clasificar en los siguientes apartados:

- Exactitud del valor absoluto del coeficiente de correlación estimado u obtenido del diagrama de dispersión dado por el alumno como respuesta.
- Estrategias utilizadas para decidir la existencia de dependencia y tipo (directa, inversa, independencia), teniendo en cuenta la tarea propuesta.
- Estrategias en la construcción de un diagrama de dispersión.
- Capacidad de proponer una variable estadística bidimensional que se adecue a un valor del coeficiente de correlación.

Estimación de la correlación

En general, los alumnos muestran una buena capacidad de estimación de la correlación, a pesar de no ser una actividad habitual en la enseñanza. La media global del error absoluto en la estimación (multiplicada por 100) ha ascendido a 28'73, con un error típico de 1'265, siendo su rango de variación de 10'124 a 57'093.

Se ha observado, de igual modo, el efecto sobre este error de las siguientes variables de tarea incluidas en el cuestionario: Tipo de tarea, intensidad de la correlación, tipo de covariación y tipo de dependencia. No hemos encontrado influencia del tipo de ajuste o de las teorías previas.

Respecto al tipo de tarea, la estimación es más fiable en las tareas 4 y 6, (estimar el valor del coeficiente de correlación a partir de un diagrama de dispersión y, su inversa, construir una nube de puntos a partir del coeficiente de correlación). Podemos considerar que la familiaridad de los alumnos con los diagramas de dispersión les ha llevado a realizar una buena estimación, tanto en sentido directo (del diagrama al coeficiente) como inverso (del coeficiente al diagrama). Los errores son mayores en las tareas 1 y 3 (construir un diagrama de dispersión a partir de una descripción verbal y estimar el valor del coeficiente de correlación a partir de una tabla de datos numéricos). En consecuencia, tanto el coeficiente de correlación como la nube de puntos parecen facilitar la tarea de traducción a los alumnos.

En cuanto a la intensidad, los errores son menores conforme aumenta la intensidad de la asociación presentada, es decir, se detecta con mayor facilidad las correlaciones intensas, lo que coincide con los resultados de las investigaciones en psicología.

En relación al tipo de covariación, las estimaciones con menor error se producen cuando la situación presenta dependencia causal unilateral, que es la situación que frecuentemente aparece en los ejemplos y ejercicios de los libros de texto. Por el contrario, el error es mayor cuando hay dependencia casual y dependencia indirecta, situaciones que los alumnos asocian con mayor asiduidad a dependencias de tipo aleatorio.

Respecto al tipo de dependencia, los alumnos de la muestra estiman mejor la dependencia directa que la inversa, lo cual es coherente con otras investigaciones, como la de Erlick y Mills (1.967).

Estrategias en los juicios sobre el signo e intensidad de la dependencia

En la investigación de Estepa se puso de manifiesto la importancia del estudio de las estrategias que siguen los alumnos en los juicios de asociación. Al igual que en dicho trabajo, hemos realizado un estudio cualitativo de la relación entre estas estrategias y las variables de tarea incluidas en el cuestionario.

El análisis de correspondencias nos ha permitido identificar tres ejes que explican las relaciones existentes entre las distintas subtareas y la estrategia utilizada para encontrar la asociación y el signo de la misma. La alta calidad de representación y los valores obtenidos de las correlaciones de filas, columnas y columnas suplementarias, indican la validez de los resultados.

El primero de los ejes opone el razonamiento numérico al gráfico, dependiendo del tipo de tarea. La tarea 2 (estimar el coeficiente de correlación a partir de la descripción verbal) está asociada a la estrategia 2 (E2), que pone en juego el marco gráfico. Los alumnos recurren a la representación gráfica para estimar el coeficiente de correlación cuando ésta no aparece en el enunciado del problema. Por el contrario, en algunas subtareas de las tareas 1 y 4, que ya contienen algunos elementos gráficos en el propio enunciado, se asocian a la estrategia 3 (E3) -marco numérico-. En consecuencia, deducimos la complementariedad de los marcos gráficos y numérico y la necesidad que sienten los alumnos de recurrir a ambos marcos en las actividades de estimación de la correlación.

En este eje se muestran también el efecto de las variables de tarea: Dependencia directa (marco numérico), el tipo de ajuste no lineal (marco numérico), la alta intensidad de la correlación (marco numérico), la dependencia causal unilateral y la independencia (marco numérico). Ello pone de manifiesto el interés de las variables consideradas, que no sólo hacen variar la dificultad de la tarea, sino incluso las estrategias seguidas en la misma.

El segundo eje muestra el empleo conjunto de varias argumentaciones, numéricas y gráficas y teorías previas para decidir la correlación presentada en las tareas, frente a los casos en que es suficiente un argumento gráfico. El segundo

caso se asocia a algunas subtareas de las tareas 2 y 4 y el primero a la 1, precisamente situaciones que son muy familiares al alumno (temperatura según latitud y relación entre calificaciones y edad). Parece que, puesto que el alumno conoce bien las situaciones le es menos necesario el recurso a estrategias complementarias.

Por último, el tercer eje muestra que el aumento o disminución no uniforme de las dos variables ofrecidas en una tabla (tarea 3), produce en algunos alumnos una confusión al compararlo con el crecimiento o decrecimiento uniforme de la dependencia funcional. Ello lleva a responder una correlación sin justificarla, a utilizar sus teorías previas o a usar diversos tipos de argumentaciones incorrectas. En todo caso, la dependencia indirecta y concordancia parecen asociados al empleo de teorías previas, lo que también se encontró en el trabajo de Estepa (1.994).

Construcción de diagramas de dispersión

Al analizar los diagramas de dispersión diseñados por los alumnos, hemos encontrado cuatro estrategias de construcción, que conducen a la siguiente tipología:

- **Diagrama propio de una dependencia de tipo funcional (1)**. Cuando el alumno dibuja un diagrama de dispersión y une los puntos con una línea poligonal, que no corresponde a una función conocida.
- **Diagrama propio de una dependencia de tipo funcional (2)**. Cuando el alumno une los puntos del diagrama de dispersión, mostrando un modelo funcional conocido, por ejemplo, una recta, a pesar de que éste no se corresponda con el subyacente en la situación propuesta.
- **Diagrama que representa una dependencia "casi aleatoria"**. Cuando el alumno construye una nube de puntos basándose en modelos deterministas, aunque con cierta dispersión.

- **Diagrama que representa una dependencia de tipo aleatorio**. Cuando los alumnos construyen diagramas de dispersión típicos de estas situaciones.

Para cada una de las subtareas de las tareas 1 y 6 hemos analizado las frecuencias con que se presentan este tipo de diagramas, observando el efecto de las variables intensidad, tipo de ajuste, covariación y signo.

Capacidad para proponer una variable estadística bidimensional

Los sujetos de la muestra exhiben, en general, una conveniente capacidad para asociar una variable estadística bidimensional a un valor del coeficiente de correlación. El signo de la variable estadística bidimensional con la que los sujetos de la muestra ejemplifican el valor del coeficiente de correlación dado en la subtarea es concordante con el signo del valor del coeficiente de correlación presentado en la tarea en las respuestas ofrecidas por dos de cada tres alumnos, aproximadamente. Existe una mayor correspondencia si $r > 0$, mientras que en caso de que $r < 0$ tienen notables dificultades para encontrar un ejemplo adecuado.

El orden de dificultad, de mayor a menor, que tienen los sujetos de la muestra cuando tienen que ejemplificar un tipo de asociación mediante una variable estadística bidimensional es el siguiente: i) Dependencia funcional, ii) independencia, y, iii) dependencia aleatoria.

Capítulo 7

Resolución de los problemas

7.1. INTRODUCCIÓN

En este capítulo abordamos el análisis de las soluciones aportadas por los alumnos a los dos problemas sobre las nociones de la correlación y, esencialmente, de la regresión, incluidos en el cuestionario. Esto es así, debido a que, en los dos capítulos anteriores, se estudiaban el conocimiento conceptual sobre hechos estadísticos conectados con la asociación en general, Capítulo 5, así como la capacidad de estimación de la correlación, Capítulo 6. Además, como la investigación de Sánchez Cobo (1.996) puso de manifiesto en su análisis de los ejercicios de libros de texto, existe un sesgo significativo hacia aquellas tareas que tenían una finalidad meramente algorítmica, en detrimento de otras, como la interpretación, representación gráfica, predicción, comparación y recogida y análisis de datos, que son esenciales para una formación estadística integral de nuestros alumnos desde un punto de vista educativo. En este capítulo queremos valorar no sólo la capacidad de cálculo de los diversos coeficientes que intervienen en la correlación y regresión, sino la competencia en la interpretación de los mismos y su aplicación a situaciones problemáticas.

Los análisis que presentamos a continuación, se han desarrollado respecto al cálculo de coeficientes y parámetros y las actividades de predicción e interpretación. Asimismo, las tablas, con las diversas respuestas de los alumnos de la muestra a los problemas 1 y 2, se encuentran en el Anexo VII.

7.2. CÁLCULO DE COEFICIENTES Y PARÁMETROS

En esta sección se estudian los procedimientos de cálculo de coeficientes y parámetros implicados en los dos problemas, y que giran en torno al:

- Cálculo de la media de la variable explicada, dados la pendiente de la recta de regresión, el punto de corte de esta recta con el eje de ordenadas y la media de la variable explicativa (Problema 1 del cuestionario).
- Cálculo del coeficiente de correlación a partir de una tabla de datos de una distribución estadística bidimensional (Problema 2 apartado a) del cuestionario).
- Determinación de la recta de regresión de Y sobre X a partir de una tabla de datos de una distribución estadística bidimensional (Problema 2 apartado c) del cuestionario).

7.2.1. CÁLCULO DE LA MEDIA DE LA VARIABLE EXPLICADA A PARTIR DE LOS DATOS OFRECIDOS EN EL PROBLEMA 1

Problema 1. Una recta de regresión tiene una pendiente de 16 y corta al eje de ordenadas en el punto y = 4. Si la media de la variable independiente es 8, ¿cuál es la media de la variable dependiente?

En el Problema 1 se les solicita a los alumnos que averigüen la media de la variable explicada, presentándoseles como datos: i) La pendiente de la recta de regresión, ii) el corte de esta recta con el eje de ordenadas, y, iii) la media de la variable explicativa. A continuación, efectuamos un análisis de contenido de los

procedimientos expuestos por los alumnos en sus respuestas, estudiando los resultados obtenidos.

La National Council of Teachers of Mathematics publicó en 1.983 *The agenda in action*, donde recogía una serie de recomendaciones sobre la enseñanza y aprendizaje de las Matemáticas. La primera de ellas se centraba sobre la resolución de problemas, sugiriendo que un buen problema matemático es *"una situación que implica un objetivo que hay que conseguir, hay obstáculos para alcanzar ese propósito, y requiere deliberación, ya que quien lo afronta no conoce ningún algoritmo para resolverlo. La situación es habitualmente cuantitativa o requiere técnicas matemáticas para su resolución y debe ser aceptado como problema por alguien antes de que pueda ser llamado problema"* (House, Wallace y Johnson, 1.983, pág. 10). En este sentido, creemos que el Problema 1 es un verdadero problema para los alumnos, dado que no disponen, de inmediato, de un procedimiento estándar, desarrollado explícitamente en la enseñanza recibida, para resolverlo. En consecuencia, deben recabar de su repertorio de conocimientos geométricos, particularmente aquellos que poseen sobre la recta en el plano cartesiano, y relacionarlos con las nociones adquiridas en la enseñanza que han recibido sobre la regresión lineal estadística. Si los primeros fallan nos encontraremos con métodos no adecuados y, muchas veces, no comprensibles para un observador externo.

En la Tabla 7.2.1.1, hemos clasificado las respuestas de los 129 alumnos que han resuelto el problema, por procedimiento seguido y tipo de solución. Todos los porcentajes que comentamos a continuación, mientras no se especifique otra cosa, se refieren a estos 129 alumnos.

Como puede deducirse de ella, y según está planteada la actividad, la inmensa mayoría de los sujetos de la muestra emplean una o ambas rectas de regresión para resolver el problema; en concreto, 122 de los 129 alumnos que responden, o sea el 94'57 %. De ellos, 106 optan por utilizar la recta de regresión de Y sobre X, el 82'17 %. La recta de regresión de X sobre Y es utilizada por 13, el 10'08 %, y aplican ambas rectas de regresión 3, el 2'32 %.

1. Uso de la recta de regresión de Y sobre X.

Dentro de los que se sirven únicamente de la recta de regresión de Y sobre X (tres primeras filas de la Tabla 7.2.1.1), 81 alumnos de 106 han utilizado, exclusivamente, la expresión de la ecuación explícita de la recta $y = mx+n$, 23 alumnos de 106 han usado, únicamente, la expresión punto-pendiente de la recta

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

y, 2 alumnos de 106 han aplicado ambas expresiones.

Tabla 7.2.1.1. Frecuencia y porcentaje de soluciones correctas e incorrectas, según los procedimientos de resolución del Problema 1

	SOLUCIONES				
PROCEDIMIENTOS DE RESOLUCIÓN	**Correcta**		**Incorrecta**		**Total**
	Frecuencia	%	Frecuencia	%	Frecuencia
$y = mx + n$	31	38'3	50	61'7	81
$y - y_m = [\, s_{xy} / s_x^2 \,](x - x_m)$	8	34'8	15	65'2	23
Usa ambas expresiones r.reg.	2	100			2
$x = m'y + n'$	2	11'1	7	88'9	9
$x - x_m = [\, s_{xy} / s_y^2 \,](y - y_m)$			4	100	4
Usa dos rectas de regresión	2	66'7	1	33'3	3
Usa un parámetro estadístico			7	100	7
Total	45	34'9	84	65'1	129

Ahora bien, según el tipo de expresión de la ecuación de la recta de regresión que emplean, los alumnos se enfrentan a dificultades de diversa índole que, en la mayoría de las ocasiones, les conducen a una resolución inadecuada del problema, pero que, en algunos casos excepcionales, les permiten descubrir ciertos métodos muy ingeniosos, constituyéndose en un conjunto de indicios de los escollos que han deseado salvar. Así, si el estudiante opta por la expresión de la

ecuación explícita de la recta de regresión, $y = m\,x + n$, el principal inconveniente surgiría del desconocimiento del hecho estadístico de que el centro de gravedad $(\bar{x},\bar{y})$ es un punto del plano que pertenece a la recta, y, por tanto, ha de verificar la ecuación anterior, es decir,

$$\bar{y} = m\bar{x} + n$$

En el análisis que en el epígrafe 3.3 se efectuó de los contenidos del curso de iniciación a la Estadística que habían recibido los alumnos de la muestra, se destacó que la noción de centro de gravedad no se trataba de forma tácita, sino que de manera indirecta se la denotaba como punto de corte de las dos rectas de regresión. Estimamos que, el efecto anteriormente mostrado, es una de las causas de la percepción que los estudiantes tienen del centro de gravedad.

Asimismo, entendemos que, desde un punto de vista epistemológico, a veces no se pone de manifiesto suficientemente en la enseñanza el estudio de la simetría de las características de la variable estadística bidimensional considerada, respecto a la correlación y la regresión. En relación con la correlación siempre son simétricas. Pero en relación con la regresión depende del tipo de covariación (Barbancho, 1.973) presentado en el contexto o en las condiciones del problema. Si la dependencia es causal unilateral, la característica explicativa y la característica explicada quedan determinadas de forma unívoca. Si el tipo de covariación es uno de los cuatro restantes - interdependencia, dependencia indirecta, concordancia y covariación casual - el resolutor debe decidir qué recta de regresión hay que emplear, Y sobre X o X sobre Y, y, desde ese momento, quedan determinadas ambas características, explicativa y explicada.

En el Problema 1, las condiciones expresadas en el enunciado no determinan de forma unívoca las variables dependiente e independiente. De todo lo anterior, no es de extrañar que una de las dificultades más notables que han aflorado en este caso, es confundir la variable explicativa con la explicada, con una frecuencia de 27 alumnos de los 50 que han usado la ecuación $y = mx + n$. Esto supone el 54 % de las respuestas incorrectas entre estos alumnos, lo que ha conducido a estos sujetos a intercambiar la media de X con la de Y, y, por consiguiente, a una solución errónea. Esto es uno de los actos de comprensión identificados en los trabajos de Estepa (1.994) y Batanero y cols. (1.997), Batanero,

Godino y Estepa (1.998) y Batanero, Estepa y Godino (1.998). Por ejemplo, el sujeto 46 dice que:

$$"a = 4,\ b = pendiente = 16,\ y = a + bx,\ \bar{y} = 8"$$

Otra dificultad a tener en cuenta es la interpretación incorrecta de los parámetros m y n en la ecuación de la recta $y = mx + n$, conmutando sus valores (Azcárate, 1.990), o bien, asignándoles valores inadecuados, la cual presenta una frecuencia de 9 alumnos. Por ejemplo, el sujeto 146 contesta a esta cuestión de la siguiente manera:

$$"y = a + bx,\ y = 4,\ a = 16,\ x = ?,\ 4 = 16 - 8x,\ 12 = 8x,\ x = 12 / 8 = 1'5"$$

Las demás dificultades, relativas al uso de la recta de regresión de Y sobre X, no son significativas debido a su baja frecuencia.

Si, por el contrario, los alumnos han empleado la expresión punto-pendiente de la recta de regresión de Y sobre X, el escollo fundamental que han encontrado es que, como necesitan un punto, no comprenden que ese es el punto de corte con el eje de ordenadas y no saben interpretarlo de manera pertinente. Así, de los 23 alumnos que emplean esta expresión 8 responden correctamente a esta cuestión -34'8 %-. Las dificultades provienen, básicamente, de la utilización de una expresión inadecuada de la recta de regresión de Y sobre X, con una frecuencia de 7 alumnos, y la confusión de quién es la variable independiente y quién la variable dependiente, con una frecuencia de 8 alumnos. Un ejemplo de la primera dificultad es la respuesta del sujeto 4 que dice:

$$"\bar{x} = 8,\ y = 4,\ y - \bar{y} = \frac{S_{xy}}{S_x^2} - (x - \bar{x}),\ y - \bar{y} = 16 - 8,\ \bar{y} = 8 - 4,\ \bar{y} = 4"$$

Por otra parte, la contestación del sujeto 57 nos puede servir para ejemplificar una dificultad del segundo tipo:

$$"\frac{S_{xy}}{S_x^2} = 16,\ (0,4),\ \bar{y} = 8,\ y - 8 = 16(0 - \bar{x}),\ -4 = -16\,\bar{x},\ \bar{x} = 1/4"$$

La forma de expresión de la ecuación de la recta de regresión parece jugar un papel notable en la interpretación de la pendiente de ella, pues mientras la ecuación de la recta sea $y = mx + n$, los alumnos muestran bastantes dificultades para interpretar que m es la pendiente. Por el contrario, si la ecuación de la recta de regresión de Y sobre X es

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

los estudiantes han mostrado menos inconvenientes para comprender que la pendiente viene expresado por $\frac{\sigma_{xy}}{\sigma_x^2}$.

Dos alumnos se sirven de ambas expresiones de la recta de regresión de Y sobre X, la ecuación explícita y la ecuación punto-pendiente, siguiendo un plan muy interesante: 1º. Desconocen que el centro de gravedad pertenece a la recta de regresión $y = mx+n$, 2º. Observan que la media de la variable dependiente se encuentra en la expresión de la ecuación punto-pendiente, 3º. Comparan ambas expresiones y determinan la media de la variable explicada. De esta forma se desarrolla la contestación del sujeto 33:

$$\text{"}y \text{ sobre } x,\ y = ax + b,\ y = 16x + 4,\ y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}),\ a = \frac{S_{xy}}{S_x^2} = 16,$$

$$b = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x},\ b = \bar{y} - 16 \cdot 8,\ \bar{y} = 132\text{"}$$

2. Uso de la recta de regresión de X sobre Y.

Trece alumnos, de los 129 que han respondido este problema (10'1 %), han utilizado la recta de regresión de X sobre Y, lo que en términos generales, y, dado que el problema se plantea en forma abierta, no se puede considerar que confundan la variable dependiente con la independiente. De éstos 13, 9 emplean la expresión de la ecuación explícita de la recta de regresión, $x = m'y + n'$, y sólo 2 de ellos responden de forma correcta. Los 4 restantes usan la ecuación punto-pendiente de la recta de regresión, $x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$. Nuevamente, se repiten las dificultades reseñadas anteriormente, habiendo sido encontradas la confusión entre la variable explicativa y la explicada y el intercambio de papeles entre el punto de corte con el eje de ordenadas y la pendiente de la recta de regresión.

Como ejemplo de la confusión entre las variables dependiente e independiente podemos citar la respuesta dada por el sujeto 105 que dice:

$$\text{"}x = 8,\ y = 4,\ b' = 16,\ x = a' + b'y,\ 8 = a' + 16 \cdot 4,\ 8 = a' + 64,\ a' = 64 - 8 = 56\text{"}$$

Y para la dificultad relacionada con los parámetros de la recta de regresión, podemos ejemplificarla con la contestación del sujeto 89:

$$"a = \bar{x} - b\,\bar{y}\ ,\ 16 = 8 - 4\,\bar{y}\ ,\ \bar{y} = (-16+8)/4 = 2"$$

Con respecto a la expresión de la ecuación punto-pendiente de la recta de regresión de X sobre Y, la única dificultad que es destacable es la de la confusión entre variable explicativa y explicada, con la mitad de los 4 alumnos que la utilizan, como le ocurre al sujeto 122 cuando responde así:

$$"x = 0,\ y = 4,\ \bar{x} = 8,\ x - \bar{x} = \frac{Cov(X,Y)}{S_y^2}\,(y - \bar{y}),\ x - 8 = 16(4 - \bar{y}),$$

$$0 - 8 = 16(4 - \bar{y}),\ 16\,\bar{y} = 16 \cdot 4 + 8 = 72,\ \bar{y} = 4'5$$

La ecuación de la recta viene definida como $x - x_1 = m\,(y - y_1)$ siendo m la pendiente de la recta. Como nos dan la pendiente, el valor de y, y sabemos que corta el eje Y en el punto 4, en ese punto $x = 0$. Al sustituir, la media de la variable dependiente sale 4'5"

Las demás dificultades no son significativas dada su baja frecuencia.

3. Uso de las dos rectas de regresión Y sobre X y de X sobre Y.

Tres alumnos se sirven de ambas rectas de regresión, Y sobre X y X sobre Y. Podemos considerar que ésta es la mejor respuesta, pues, como indicábamos anteriormente, el problema se planteó de forma abierta y, por tanto, es plausible considerar a X como variable explicativa o como variable explicada. En estos términos se expresa el sujeto 67:

$$"Suponemos\ la\ recta\ Y/X:\ x = 0,\ y = 4,\ y = ax + b,\ 4 = 0 + b,\ b = 4,\ b = \bar{y} - a\bar{x},$$

$$4 = \bar{y} - 16 \cdot 8,\ 4 = \bar{y} - 128,\ \bar{y} = 132.\ Suponemos\ la\ recta\ X/Y:\ x = a'y + b',$$

$$0 = 16 \cdot 4 + b,\ b = -64,\ b = \bar{x} - a'\,\bar{y}\ ,\ -64 = \bar{x} - 16 \cdot 8,\ \bar{x} = 64"$$

En resumen, podemos concluir que la dificultad más importante con la que se enfrentan los sujetos de la muestra que utilizan una sola expresión para una única recta de regresión, es la de discriminar entre la variable explicativa y la variable explicada, no habiéndolas diferenciado de forma pertinente el 50 % de los estudiantes que responden de forma incorrecta a este punto (38 alumnos de 76).

Aunque las expresiones de la ecuación de la recta de regresión, punto-pendiente y explícita, son, obviamente, equivalentes, estos alumnos emplean como herramienta más operativa la punto-pendiente, pues en ella identifican mejor

la pendiente que en la forma explícita; sin embargo, utilizan con más frecuencia la forma explícita, probablemente debido al enunciado del problema.

Finalmente, hemos de advertir que, aunque no presenten una frecuencia significativa, 7 alumnos de 129 intentan resolver este problema apoyándose en conceptos estadísticos, algunos de los cuales se encuentran en el enunciado del problema; para ello usan parámetros estadísticos como, por ejemplo, la media, la covarianza, ... , siendo aplicados todos estos procedimientos de forma inadecuada.

7.2.2. CÁLCULO DEL COEFICIENTE DE CORRELACIÓN A PARTIR DE UNA TABLA DE DATOS DE UNA DISTRIBUCIÓN ESTADÍSTICA BIDIMENSIONAL

En este parágrafo se estudian los procedimientos utilizados por los alumnos para determinar el coeficiente de correlación a partir de una tabla de datos de una distribución estadística bidimensional, que es la actividad correspondiente al Problema 2 apartado a).

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

a) Calcule el coeficiente de correlación lineal

En primer lugar, hay que destacar el índice de no respuestas existente dado por una frecuencia de 25 alumnos, lo que representa el 12'95 %. Consecuentemente, han respondido 168 alumnos, lo que sería el 87'05 % del total. De estos últimos, en la Tabla 7.2.2.1 resumimos la frecuencia de respuestas, indicando su corrección. Con tal fin, hemos considerado como respuesta correcta aquélla en la cual la fórmula y cálculo han sido correctos.

El averiguar el valor del coeficiente de correlación a partir de una tabla de datos correspondientes a una variable estadística bidimensional, es una actividad típica del tópico sobre la correlación y la regresión, y que, naturalmente, ha sido tratada en el curso de introducción a la Estadística que han recibido los alumnos universitarios que conforman la muestra. De la Tabla 7.2.2.1 se deduce que los alumnos emplean una única expresión para determinar el coeficiente de correlación de entre los algoritmos, que para tal fin, se les proporcionaba en la enseñanza recibida, como queda recogido en la sección 3.3.2: i) $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, que es el utilizado, y, ii) $r^2 = b \cdot b'$, que, evidentemente, es menos operativo que el anterior.

Tabla 7.2.2.1. Frecuencia y porcentaje de respuestas correctas e incorrectas, según los procedimientos de resolución del Problema 2 apartado a)

	RESPUESTA				
PROCEDIMIENTOS DE RESOLUCIÓN	**Correcta**		**Incorrecta**		**Total**
	Frecuencia	%	Frecuencia	%	Frecuencia
Utiliza la fórmula $r = s_{xy} / (s_x s_y)$	106	65'8	55	34'2	161
Utiliza una expresión inadecuada del coeficiente de correlación			5	100	5
Utiliza el coeficiente de determinación			2	100	2
Total	106	63'1	62	36'9	168

Además, aproximadamente, un tercio de los 161 alumnos que han utilizado la fórmula correcta para calcular el coeficiente de correlación, han dado una respuesta inadecuada. Los cálculos mal efectuados es el principal error que cometen, pues le ha ocurrido a 22 estudiantes de los 55, lo que supone un 40 % de las respuestas incorrectas. Entre las respuestas que presentan errores es interesante reseñar la interpretación del coeficiente de correlación que da el sujeto 117 cuando expone:

"r = 1'2. Este resultado no es posible ya que r es un valor comprendido entre ± 1, $-1 < r > 1$. Por lo que estas son independientes, no tienen relación"

Nuevamente, como ya exponíamos en el epígrafe de la correlación del Capítulo 5, este alumno evidencia una concepción errónea consistente en considerar que r es un parámetro matemático que tiene como dominio de definición el conjunto de los números reales $\mathbf{R}$. Si hay relación entre las variables, el valor del coeficiente de correlación pertenecerá al intervalo [-1,1], siendo ésta más intensa cuanto más próximo se encuentre r al valor 0. Mientras que cuando no existe asociación, el valor del coeficiente de correlación pertenece al complementario de dicho intervalo, es decir, al conjunto $(-\infty, -1) \cup (1, \infty)$. Esta concepción es tan resistente para este estudiante, que no se plantea el revisar sus cálculos, por si se hubiera equivocado, sino que interpreta el valor obtenido a la luz de su percepción del campo de existencia del coeficiente de correlación.

Todo lo contrario le sucede al sujeto 173 que llega a manifestar:

"r = 2'24. Debe salir entre -1 ≤ r ≤ 1. Este dato debe estar mal"

dibuja el diagrama de dispersión y efectúa una estimación, aunque errónea, de este parámetro estadístico: *"En realidad, o eso creo, r = 0'039"*.

Otro caso interesante es la respuesta dada por el sujeto 187, que al equivocarse al hallar la desviación típica de la variable X, y salirle nula, razona de la siguiente manera:

"$r = \frac{23'8}{0}$*. Ésto no puede ser, luego no existe correlación alguna"*

A continuación dibuja la nube de puntos y ante la observación de la misma dice:

"Debe estar mal porque se ve claramente que si existe correlación. El error posiblemente se debe a que S_X *no sea cero"*

y da una estimación de r.

Por otro lado, obsérvese que, dado que el uso de la calculadora fue permitido para responder al cuestionario, cuatro de cada diez alumnos de la muestra presentan serias deficiencias en destrezas básicas en la utilización de esta herramienta en el momento de determinar los parámetros estadísticos requeridos en esta actividad. Otras fuentes de errores hay que buscarlas en la utilización de expresiones inadecuadas para determinar la varianza y la covarianza, con unos porcentajes del 16'36 % y 10'91 % de las 55 respuestas incorrectas,

respectivamente - de esta manera, verbigracia, el sujeto 2 dice: *"No se hallar la covarianza"* - y en el uso de la varianza cuando lo adecuado es emplear la desviación típica (12'73 % de las 55 respuestas incorrectas). Otros tipos de respuestas equivocadas no son significativas por su baja frecuencia.

Es conveniente mencionar como el sujeto 110, antes de hallar el valor del coeficiente de correlación, dibuja la nube de puntos y a partir de ella realiza una estimación de *r*, explicando que:

"Por el diagrama de la nube de puntos vemos a priori que el coeficiente de correlación va a estar próximo a +1 (0'8-0'9). Se va a tratar de una relación lineal directa"

Sin embargo, cuando resuelve las tareas 3 y 4 no aplica este procedimiento a ninguno de sus apartados.

Dentro de las respuestas que emplean una expresión inadecuada del coeficiente de correlación, podemos citar la del sujeto 7 en la que usa las varianzas marginales en vez de las desviaciones típicas: $"r = \frac{S_{xy}}{S_x^2 S_y^2} = \frac{35'8}{2 \cdot 299'44} = 0'059"$. También, hay algunos alumnos que averiguan el coeficiente de determinación como hace el sujeto 54: $"r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{566'44}{2 \cdot 299'44} = 0'94"$.

7.2.3. DETERMINACIÓN DE LA RECTA DE REGRESIÓN DE *Y* SOBRE *X* A PARTIR DE UNA TABLA DE DATOS DE UNA DISTRIBUCIÓN ESTADÍSTICA BIDIMENSIONAL

Ahora vamos a analizar cómo determinan los alumnos la recta de regresión de Y sobre X a partir de una tabla de datos de una distribución estadística bidimensional, que es la actividad correspondiente al problema 2 apartado c).

La tasa de no respuestas de este problema ha sido del 16'1 % de los alumnos de la muestra, lo que hace un total de 31 alumnos. Por lo tanto, 162 alumnos han contestado a esta cuestión, lo que representa un total del 83'9%.

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

c) Determine la recta de regresión de y, número de bacterias por unidad de volumen, sobre x, número de horas

Como puede observarse en la Tabla 7.2.3.1, los estudiantes, para contestar esta cuestión, han elegido o bien una recta de regresión, con un porcentaje del 82'1 % del total de respuestas, o ambas rectas, con un porcentaje del 17'9 % del total de respuestas. De los 133 que usan una recta de regresión, utilizan la expresión de la ecuación punto-pendiente de la recta regresión de Y sobre X un 97'0 %, mientras que, por el contrario, usan la recta de regresión de X sobre Y un 3'0 % de estas 133 respuestas.

Hay un notable porcentaje de los sujetos que han respondido el problema - 17'9 % - que hallan tanto la recta de regresión de Y sobre X como la de X sobre Y. Esto puede ser debido a que no saben claramente cuál recta hay que determinar en este apartado o a que, como posteriormente - Problema 2 apartado e) - necesitarán la recta de regresión de X sobre Y, deciden obtenerla en este apartado. Los alumnos incluidos en este grupo no hacen ninguna indicación a tal respecto.

Tabla 7.2.3.1. Frecuencia y porcentaje de respuestas correctas e incorrectas, según las estrategias de resolución del Problema 2 apartado c)

	RESPUESTA				
ESTRATEGIAS DE RESOLUCIÓN	**Correcta**		**Incorrecta**		**Total**
	Frecuencia	%	Frecuencia	%	Frecuencia
Usa recta de regresión de Y sobre X	81	62'8	48	37'2	129
Usa recta de regresión de X sobre Y			4	100	4
Ambas rectas de regresión			29	100	29
Total	81	50	81	50	162

Es de resaltar que la mitad de los estudiantes que responden (81 de 162) lo hacen de forma inadecuada, siendo esta una actividad tan primordial para este tema. En cuanto a los errores que cometen los alumnos son, esencialmente, similares a los expuestos en el parágrafo anterior.

7.3. PREDICCIÓN E INTERPRETACIÓN

En este epígrafe se analizan las tareas de predicción e interpretación que se incluyen en el Problema 2 y que giran en torno a:

- Expresar el tipo de relación que existe en una tabla de datos de una variable estadística bidimensional, después de calcular el coeficiente de correlación (Problema 2 apartado b) del cuestionario).
- Predecir dos valores de la variable explicada a partir de dos valores dados de la variable explicativa (Problema 2 apartado d) del cuestionario).
- Predecir a partir de la recta de regresión X sobre Y (Problema 2 apartado e) del cuestionario).

7.3.1. JUICIO DE ASOCIACIÓN A PARTIR DEL CÁLCULO PREVIO DEL COEFICIENTE DE CORRELACIÓN

En el apartado b) del Problema 2, se solicitaba a los alumnos que, después de determinar el coeficiente de correlación, expresaran el tipo de dependencia que consideran que existe entre ambas variables.

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

b) Decir qué tipo de relación (directa, inversa o independencia) existe entre ambas variables

Los procedimientos que han empleado los sujetos de la muestra se han agrupado en 6 categorías, cuyas frecuencias y respuestas obtenidas se ofrecen en la Tabla 7.3.1.1 y que, a continuación, se exponen:

Tabla 7.3.1.1. Frecuencia y porcentaje de las estrategias y respuestas en el Problema 2 apartado b)

ESTRATEGIAS	RESPUESTA						NO DA JUICIO		TOTAL
	Dependencia directa*		Dependencia inversa		Independencia				
	Frecuencia	%	Frecuencia	%	Frecuencia	%	Frec	%	
Coef. correlación	46	90'2	1	2'0	3	5'8	1	2'0	51
Variación conjunta	17	94'4			1	5'6			18
Covarianza	17	94'4	1	5'6					18
Coeficiente determinación	3	50	1	16'7	2	33'3			6
Otras	3	42'9			3	42'9	1	14'2	7
Sin estrategia	57	61'3	2	2'2	5	5'4	29	31'1	93
Total	143	74'1	5	2'6	14	7'2	31	16'1	193

* respuesta correcta

1. Usa el coeficiente de correlación

Es una estrategia correcta. Si el cálculo del coeficiente ha sido correcto, la respuesta aparece como correcta, en otro caso como incorrecta. Un aspecto a destacar es que aunque 168 alumnos han calculado el coeficiente de correlación en el problema 2 apartado a), menos de un tercio de ellos, 51 estudiantes, lo utilizan para averiguar el tipo de asociación que existe entre las variables. Así, verbigracia, el sujeto 33 responde:

"La relación es directa ya que el coeficiente de correlación es positivo"

A pesar de que en el cuestionario se les aportaba la terminología para expresar la dirección de la asociación, hay alumnos que se explican a través de expresiones ambiguas, como, por ejemplo, el sujeto 3 que dice:

"La relación existente entre ambas variables es muy buena ya que el coeficiente de correlación es 0'9, es decir, está próximo a 1".

Un hecho estadístico destacado es que *"la magnitud* (fuerza) *del coeficiente de correlación es por completo independiente de su dirección* (positiva o negativa)" (Phillips, 1.992, pág. 113). Es importante subrayar que hay alumnos que confunden, de manera más o menos explícita, la intensidad y el sentido de la dependencia, como, por ejemplo, el sujeto 21 que contesta:

"Existe una relación directa ya que el coeficiente de correlación es alto"

o el sujeto 58 que responde:

"La relación es directa ya que r es positivo y se acerca a 1"

Una explicación plausible del por qué los alumnos confunden estas dos características podría ser que, en general, la definición de correlación que se da, tanto en los textos universitarios como en los apuntes del profesor y los tomados por las alumnas, hace referencia a la correlación como la medida de la intensidad de la relación. Asimismo, en los textos de bachillerato analizados por Sánchez Cobo (1.996), una buena parte de ellos, definen este concepto en términos del grado de dependencia entre las variables o mayor o menor dependencia de las variables.

Parece natural, entonces, que los alumnos infieran que el tipo de asociación esta relacionado con la intensidad de la misma.

2. Variación conjunta de las variables

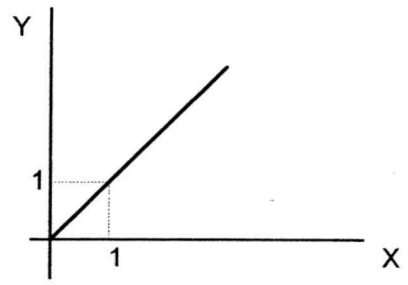
Es una estrategia correcta. En este caso, el alumno, al observar la tabla de datos argumenta, en función del marco numérico, expresando la variación conjunta entrambas variables, es decir, enlazando el aumento (o disminución) de los valores de una variable con el aumento (o disminución) de los valores de la otra. En este razonamiento se apoya el sujeto 8 cuando expone:

"Relación directa puesto que a medida que va pasando el tiempo, el número de bacterias aumenta"

La determinación del signo de la asociación basándose en la variación conjunta podría convertirse en una dificultad para los alumnos, puesto que a veces les hace confundir asociación con proporcionalidad. Así el sujeto 119 responde:

"La relación que existe entre el número de horas y el número de bacterias por unidad de volumen es dependiente, y además directamente proporcional, es decir, si

una aumenta la otra también y viceversa, porque $0 < r < 1$; $r = 0'9725$ y hay que decir que al aproximarse a uno es casi la máxima relación de dependencia directamente proporcional, aproximándose su recta a



El alumno que responde independencia argumenta que:

"r = 0'056 (el valor correcto es *r = 0'97) La relación entre las variables no es buena, hay muy poca relación entre ellas. Las variables varían en el mismo sentido"*

3. Usa la covarianza

Es una estrategia correcta. En este caso el alumno se apoya en el signo de la covarianza para contestar a esta cuestión. De esta manera se pronuncia el sujeto 28 cuando indica:

"Es una relación directa ya que la covarianza es positiva"

Hay algunos alumnos que consideran, erróneamente, que el coeficiente de correlación, fundamentalmente, sirve para decidir si hay o no relación de tipo lineal, siendo la covarianza la que debe utilizarse para especificar la dirección de ella. Como ejemplo, el sujeto 112 se expresa en los siguientes términos:

"A través del coeficiente de correlación se puede afirmar que existe una dependencia lineal alta. Para determinar si es directa o inversa se utiliza la covarianza $S_{XY} = 23'8$. Al ser positiva será directa"

Claramente, este alumno desconoce una propiedad, básica, del coeficiente de correlación.

4. Usa el coeficiente de determinación

En esta argumentación el alumno usa la bondad del ajuste como criterio para juzgar la relación, así el sujeto 14 afirma:

"La relación es directa e intensa ya que r^2 se aproxima a 1"

También el sujeto 87 expone:

"$r^2 = 0'97^2 = 0'9409$, relación directa ya que se acerca a 1, la relación es buena"

Nuevamente aparece la creencia de que cuando la magnitud de r^2 o r es alta en valor absoluto la relación es directa. No obstante, en este punto obtenemos información adicional ya que, parece como si algunos alumnos juzgaran que cuando el coeficiente de determinación está en la vecindad del 1 la relación es positiva, y, en cambio, debería ocurrir que la dependencia fuera negativa si este valor está alejado del 1. Por ejemplo, el sujeto 83 responde:

"Relación inversa ya que el valor de $r^2 = 0'4568$ no está próximo a 1"

5. Otras estrategias

En esta categoría se hayan comprendidos diversos procedimientos y argumentaciones como son utilizar un intervalo de variación del coeficiente de correlación, usar el diagrama de dispersión o comparar la asociación y la proporcionalidad. Un ejemplo del primer caso es la respuesta dada por el sujeto 80 al indicar que:

"Existe una relación directa ya que el valor del coeficiente de correlación se encuentra comprendido entre 0 y 1, por lo que existe también una dependencia aleatoria"

Dentro de este grupo podemos incluir algunos alumnos que tienen la creencia de que cuando el coeficiente de correlación está fuera del intervalo [-1, 1], existe independencia, como ya vimos en el análisis de los items en el Capítulo 5 y como también aparece en la sección 7.2.2 de este capítulo. Así el sujeto 63 se expresa en los siguientes términos:

"No hay relación ninguna, pues el coeficiente de correlación lineal se mueve entre -1 y 1 y aquí ha salido -1'8061"

En el segundo caso, los alumnos elaboran el diagrama de dispersión y según la forma de la nube de puntos deciden cuál es el tipo de relación existente. Así, por ejemplo, el sujeto 100 dice:

"Según el gráfico que obtenemos al representar los datos la regresión es directa"

En el tercer caso, se observa una falta de razonamiento proporcional, bien cuando el alumno argumenta que no existe dependencia porque las variables no son proporcionales, como, verbigracia, el sujeto 120 que reflexiona de la siguiente manera:

"Entre ambas variables existe una relación de independencia (no son proporcionales)"

o bien cuando el sujeto 119 expresa *"dependencia directamente proporcional"*.

6. No explicita estrategia

Ahora el alumno responde, sin justificación razonada, sobre cuál es el signo de la relación. Por ejemplo, el sujeto 15 dice:

"Relación directa bastante dependientes"

o como el sujeto 44:

"La relación que existe entre las dos variables es directa con intensidad fuerte"

De los resultados de la Tabla 7.3.1.1 se deduce que una mayoría de los estudiantes, el 88'3 % de los que responden, han determinado de manera correcta el signo de la dependencia, sólo el 11'7 % de los mismos no han sabido determinarlo y 16'1 % de los sujetos de la muestra no han respondido a esta cuestión, lo que estimamos es un dato para la reflexión. De los que han respondido de forma adecuada los procedimientos más empleados son, al margen de los que no han mostrado argumentación alguna, el estudio del coeficiente de correlación, con un 32'2 % de 143 respuestas, y el examen de la variación conjunta de las variables y el estudio de la covarianza, con similar porcentaje 11'9 % de 143 respuestas.

Asimismo, es de subrayar la confusión que manifiestan algunos estudiantes al relacionar la intensidad de la correlación con el signo de la misma.

7.3.2. PREDICCIÓN DE VALORES DE LA VARIABLE EXPLICADA A PARTIR DE VALORES DE LA VARIABLE EXPLICATIVA

En el Problema 2 apartado d) se requería a los alumnos una doble actividad de predicción: i) Que lleven a cabo una interpolación - primera pregunta -, y, ii) que efectúen una extrapolación - segunda pregunta -.

Interpolación

En el Problema 2 apartado d), la primera pregunta pide que se realice una predicción a partir de un valor de la variable explicativa perteneciente al rango de la distribución estadística bidimensional dada. Es, por ello, una tarea de interpolación.

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

d) ¿ Qué número de bacterias cabe esperar que habrá, transcurridas 2'5 horas ?

Como ya expresábamos en la sección 4.6 del Capítulo 4, creemos que, además de la utilización de la recta de regresión, aflorarán otros métodos en la resolución de este punto como, verbigracia, procesos de interpolación y aplicación de la proporcionalidad. Las contestaciones de los estudiantes han confirmado nuestras previsiones, pues, por ejemplo, el sujeto 64 responde:

$$\left.\begin{array}{l} 2 \rightarrow 21 \\ 2'5 \rightarrow x \end{array}\right\} \; x = 26'25 \; bacterias$$

Otro caso es el del sujeto 102 que observa que 2'5 horas es el punto medio del intervalo [2,3], y estima que el número de bacterias que habrá será la mitad de las que existen a las 2 y 3 horas:

$$\frac{21+33}{2} = 27 \; bacterias$$

A pesar de que el valor obtenido es 27 bacterias, que es el mismo del apartado e), sólo hay tres alumnos que responden a la pregunta de ese apartado remitiendo al resultado obtenido en el apartado d), como ocurre con el sujeto 118 que dice:

"Habría 27 como se ha calculado en el apartado anterior a las 2'5 horas y siempre suponiendo la relación de valor 1"

En la Tabla 7.3.2.1 se agrupan estas estrategias y expresamos la frecuencia de los distintos tipos de respuestas encontrados.

Tabla 7.3.2.1. Frecuencia y porcentaje de las estrategias y respuestas en la pregunta de interpolación del Problema 2 apartado d)

ESTRATEGIAS	RESPUESTA						NO RESPUESTA		TOTAL
	Correcta		Incorrecta		Nº decimal				
	Frec	%	Frec	%	Frec	%	Frec	%	
Recta regresión Y sobre X	5	3'5	49	34'5	88	62'0			142
Recta regresión X sobre Y			8	100					8
Uso de proporcionalidad			10	100					10
No estrategia			2	6'1			31	93'9	33
Total	5	2'6	69	35'7	88	45'6	31	16'1	193

Como resultado correcto hemos considerado 31 ó 32 bacterias, dado que el valor obtenido a partir de la recta de regresión de Y sobre X es 31'625 bacterias. Un aspecto, importante, que recoge esta pregunta es la integración del resultado obtenido dentro del contexto que se les ofrece a los alumnos. Por respuesta un número decimal nos referimos a que el estudiante ha aplicado el modelo estadístico adecuado, pero ha dado como contestación un número decimal, en vez de responder con un número natural como exige el contexto (no puede haber un numero racional de bacterias).

Ahora bien, esta dificultad la tendrán, también, los alumnos que hayan utilizado otros métodos distintos del uso de la recta de regresión de Y sobre X. Así, todos los que se sirven de la recta de regresión de X sobre Y, que, evidentemente,

presentan una confusión entre variable explicativa y explicada, exhiben por respuesta un número decimal. Por ejemplo el sujeto 188 expone:

$$"2'5 = 0'09y + 0'0778119,\ 2'5 - 0'0778119 = 0'09Y,\ 2'422 = 0'09y,$$

$$y = \frac{2'422}{0'09} = 26'9\ bacterias"$$

Es conveniente señalar que, entre los que aplican procedimientos de proporcionalidad, únicamente aquellos que determinan directamente el número de bacterias, transcurridas las 2'5 horas, son los que se expresan mediante una respuesta decimal - por ejemplo el sujeto 64 -, mientras que los demás, como el sujeto 102, no tienen esta dificultad.

Creemos que los alumnos de la muestra prestan muy poca atención al resultado de esta cuestión y se encuentran atraídos por el modelo estadístico utilizado, no poniendo en duda el valor decimal obtenido, siendo posible que se sientan hipnotizados por el "juego" que les plantea el investigador (Freudenthal, 1.991; Schoenfeld, 1.988) y no presten atención al sentido común.

Extrapolación

En la segunda pregunta del Problema 2 apartado d), se les plantea a los alumnos que obtengan el valor de la variable explicada a partir de un valor de la variable explicativa, pero ahora este dato cae fuera del rango de ella, es decir, se trata de una tarea de extrapolación.

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

d) ¿ Y cuando pasen 6 horas ?

Hemos considerado como respuesta correcta aquella que no sólo da el número de bacterias cuando hayan pasado 6 horas y el criterio para el conjunto

numérico en el que hay que encontrar la solución, sino que aporten alguna justificación sobre si es o no conveniente la extrapolación en este caso. No olvidemos, que *"el modelo puede ser un perfecto resumen de la nube de puntos observada; por tanto, facilitará excelentes interpolaciones. Pero de ello no se puede concluir que las extrapolaciones hayan de ser excelentes"* (Barbancho, 1.973, pág. 239).

Vemos que, como en otras investigaciones (Truran, 1.997), los alumnos son escasamente conscientes de las restricciones que deben imponerse a la extrapolación como, verbigracia, la intensidad de la correlación, lo que no es este caso, y el tamaño de la muestra.

Hay algunos alumnos que utilizan la recta de regresión de X sobre Y lo que, evidentemente, muestra que no discriminan de forma pertinente entre la variable independiente y la dependiente. De esta manera, el sujeto 188 responde:

$$"6 = 0'009y + 0'0778119,\ 6 - 0'0778119 = 0'09y,\ 5'922 = 0'09y,$$

$$y = \frac{5'922}{0'09} = 65'80\ bacterias"$$

Asimismo, algunos sujetos de la muestra aplican la proporcionalidad para resolver este proceso de extrapolación, sucediendo, como en el apartado sobre interpolación, que los alumnos confunden asociación con proporcionalidad. Ejemplo de este hecho es la respuesta del sujeto 26:

"(2,3] Si en una amplitud de 1 hay 12 (33 - 21) en una amplitud de 0'5 habrá x y utilizamos una regla de tres simple

$$\left.\begin{matrix} 1 \rightarrow 12 \\ 0'5 \rightarrow x \end{matrix}\right\} \ x = 12 \cdot 0'5 = 6$$

$$\left.\begin{matrix} 1 \rightarrow 12 \\ 6 \rightarrow x \end{matrix}\right\} \ x = 72"$$

Es conveniente reseñar que ahora hay un pequeño aumento de alumnos que no responden en comparación al parágrafo anterior, alcanzando un 19'7 % del total de la muestra, a pesar de que, técnicamente, es una cuestión análoga a la precedente.

Tabla 7.3.2.2. Frecuencia y porcentaje de las estrategias y respuestas en la pregunta de extrapolación del Problema 2 apartado d)

ESTRATEGIAS	RESPUESTA						NO RESPUESTA		TOTAL
	Correcta		Incorrecta		Nº decimal				
	Frec	%	Frec	%	Frec	%	Frec	%	
Recta regresión Y sobre X			138	99'3	1	0'7			139
Recta regresión X sobre Y			7	100					7
Uso de proporcionalidad			7	100					7
No estrategia			2	5			38	95	40
Total			154	79'8	1	0'5	38	19'7	193

Asimismo, tampoco prestan atención al conjunto numérico en que hay que expresar la solución de este problema, como sucedía en la pregunta de interpolación.

7.3.3. PREDICCIÓN A PARTIR DE LA RECTA DE REGRESIÓN X SOBRE Y

El objetivo que se ha marcado esta investigación con el apartado e) del Problema 2 es estudiar las estrategias que utilizan los alumnos cuando tienen que estimar el valor de una variable a partir de la otra en la recta de regresión X sobre Y.

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

e) ¿ Qué tiempo deberá pasar para que el número de bacterias del cultivo sea de 27 ?

En el análisis "a priori" que hemos realizado, y que se encuentra en la sección 4.6 del Capítulo 4, estimamos que:

- algunos alumnos emplearían la estrategia correcta, utilizar la recta de regresión de X sobre Y,
- otros alumnos, no tendrían en cuenta la existencia de las dos rectas de regresión y usarían la recta de regresión de Y sobre X, ya calculada
- habría alumnos que se servirían de alguna estrategia interpolatoria basada en la proporcionalidad

En la Tabla 7.3.3.1 se dan las frecuencias y corrección de estos procedimientos. Los alumnos que utilizan la recta de regresión de X sobre Y y no obtienen respuesta correcta es debido a algún error en los cálculos o en la expresión de la recta de regresión.

Tabla 7.3.3.1. Frecuencia y porcentaje de las estrategias y respuestas en el Problema 2 apartado e)

ESTRATEGIAS	RESPUESTA				NO RESPUESTA		TOTAL
	Correcta		Incorrecta				
	Frec	%	Frec	%	Frec	%	
Recta regresión X sobre Y	49	50'5	46	47'4	2	2'1	97
Recta regresión Y sobre X			45	100			45
Uso de proporcionalidad			9	100			9
Otras estrategias			2	100			2
No estrategia			2	5	38	95	40
Total	49	25'4	104	53'9	40	20'7	193

Los alumnos que utilizan la recta de regresión de Y sobre X, sustituyen y por su valor , 27, y calculan el valor de x que, obviamente, es incorrecto. Estos estudiantes no discriminan de forma pertinente la variable independiente de la dependiente. Por ejemplo, el sujeto 190 responde de la siguiente manera:

"Tiempo para que bacterias $= 27$, $y = 1'685 + 11'905x$, $27 = 1'685 + 11'905x$,

$$x = \frac{27 - 1'685}{11'905} = 2'126,$$ *tendrán que pasar 2'126 horas"*

Los alumnos que utilizan algún procedimiento interpolatorio de tipo proporcional, tienen la creencia de que la recta de regresión es una aplicación lineal, hecho que, en general no es correcto. Así, el sujeto 84 expone que:

$$\text{"} \quad \left.\begin{array}{c} 21\ bacterias \rightarrow 2\ horas \\ 27\ bacterias \rightarrow x \end{array}\right\} \ x = 2'57\ horas\text{"}$$

En consecuencia, se advierte que estos alumnos presentan un razonamiento proporcional (Giménez, 1.988) que manifiesta ciertas carencias. Dentro de éstos hay una minoría que comparan el resultado obtenido en la segunda pregunta del apartado d) y, sin hacer ningún cálculo, responden que el tiempo que deberá transcurrir es de 2'5 horas, el mismo que el ofrecido en dicha cuestión. Esto es indicativo de una concepción funcional de la dependencia (Estepa, 1.994) o bien no discriminan las dos rectas de regresión.

Asimismo puede advertirse que el 32'02 % de los sujetos que responden lo hacen de forma correcta, es decir, han aprehendido la necesidad de la existencia de las dos rectas de regresión y cuando deben obtenerlas. En cambio, el 67'98 % de los alumnos que responden a esta pregunta dan una respuesta incorrecta, de los cuales el 43'27 % nuevamente utilizan la recta de regresión de Y sobre X, lo que nos induce a considerar que no discriminan las rectas de regresión. Por último, el 20'72 % del total de la muestra no responden a esta pregunta, resultado bastante semejante al obtenido en el apartado d) en cada una de las dos preguntas presentadas. No obstante, si lo comparamos con los alumnos que no han respondido al Problema 2 apartado c), que recordemos consistía en obtener la recta de regresión de Y sobre X, vemos que se aumenta en cinco puntos el porcentaje de no respuestas, lo que parece indicarnos que, primero, no comprenden que sentido tiene el determinar dicha recta de regresión, y por ende la misma noción de la regresión, y/o, segundo, no saben emplearla como útil para efectuar una predicción, el cual es uno de sus fines fundamentales.

7.4. CONCLUSIONES SOBRE LA RESOLUCIÓN DE PROBLEMAS

En este capítulo hemos analizado las soluciones dadas por los alumnos a dos problemas sobre correlación y regresión, así como las estrategias que han obtenido para determinar esta solución. Los problemas abarcan las actividades de cálculo de diversos parámetros a partir de otros o a partir de la tabla de valores, interpretación y predicción. A continuación resumimos las principales conclusiones obtenidas sobre cada uno de estos puntos.

Actividades de cálculo

a) Media de la variable dependiente

Una primera actividad ha sido la determinación de la media de la variable dependiente conocidos la pendiente de la recta de regresión, el punto de corte con el eje de ordenadas y la media de la variable independiente.

La mejor respuesta a este problema sería hallar el valor pedido a partir de ambas rectas de regresión. Sin embargo, son pocos los alumnos que siguen este procedimiento, debido a las dificultades que han mostrado para interpretar convenientemente los parámetros de la recta de regresión (pendiente y el punto de corte con el eje de ordenadas).

La inmensa mayoría de los estudiantes que han intentado resolver este problema emplean sólo una recta de regresión. A pesar de que la expresión que se les ha enseñado es la punto-pendiente y de haberla aplicado de forma casi exclusiva al responder al Problema 2 apartado c), en el Problema 1 han usado preferentemente la ecuación explícita.

Tanto en la recta de regresión de Y sobre X como en la de X sobre Y, las dificultades más significativas que se encuentran los alumnos es la discriminación entre la variable independiente y la dependiente, así como la interpretación inadecuada de los parámetros m y n, en especial la pendiente de la recta. En

relación con este último punto, nuestros resultados coinciden con otras investigaciones (Azcárate, 1.990 sobre funciones; Truran, 1.997 sobre regresión).

b) Coeficiente de correlación

Con respecto a la determinación del coeficiente de correlación *r*, los sujetos de la muestra exhiben un dominio pertinente de las destrezas básicas de cálculo a partir de una tabla de datos de una variable estadística bidimensional. En particular, dos de cada tres alumnos, aproximadamente, han hallado correctamente el valor del coeficiente de correlación.

El resto ha mostrado fuertes carencias en destrezas básicas de cálculo, uso de expresiones inadecuadas para el coeficiente de correlación y otros parámetros estadísticos que intervienen en el algoritmo para determinar a éste. Además, vuelve a surgir la idea de que el coeficiente de correlación pertenecerá al intervalo [-1,1] exclusivamente en el caso de que haya asociación, en caso contrario pertenecerá al complementario.

c) Recta de regresión

A pesar de que la determinación de la ecuación de la recta de regresión de Y sobre X a partir de una tabla de datos de una variable estadística bidimensional es una actividad esencial, sólo la mitad de los alumnos que responden hallan la expresión adecuada de la recta de regresión. Además, hay un grupo significado de alumnos que presentan fuertes déficits en destrezas fundamentales de cálculo y confunden la variable independiente y la dependiente.

Prácticamente la totalidad de los estudiantes han empleado la ecuación punto-pendiente de la recta de regresión Y sobre X, que es la que se les ha enseñado. Esto está en contraposición con lo que hemos observado, y explicado anteriormente, en la resolución del Problema 1.

Juicios de asociación

Una amplia mayoría de los alumnos, cuando se les permite efectuar cálculos numéricos, muestran que están capacitados para emitir juicios de asociación a partir de una tabla de datos de una variable estadística bidimensional. En particular, nueve de cada diez estudiantes han averiguado el signo de la dependencia presentada.

Las estrategias más utilizadas para efectuar dicho juicio de asociación son, de mayor a menor porcentaje de aplicación: i) La utilización del signo del coeficiente de correlación, ii) la variación conjunta de los valores de la variable explicativa y de la variable explicada, y, iii) la utilización del signo de la covarianza. La relación entre el signo de la covarianza y de la correlación también era muy utilizado por los alumnos en los items, como ya se describió en el Capítulo 5, sección 5.2. También se estudiaba allí la variación conjunta de los valores de las dos componentes de una variable estadística bidimensional.

Una mayoría de los sujetos de la muestra no han usado el coeficiente de correlación para determinar el signo de la dependencia, a pesar de haberlo calculado previamente.

Otras dificultades son confundir la magnitud del coeficiente de correlación con la dirección de la asociación (Phillips, 1.992), estimar que del valor del coeficiente de correlación puede deducirse si la relación es o no lineal - este error también aparecía en el análisis de los items -, mientras que el signo de la covarianza es el que nos indica el signo de la dependencia y la utilización del coeficiente de determinación.

Predicción

Los alumnos conocen bien cuál es el procedimiento que deben aplicar para predecir el valor de la variable explicada a partir de otro valor de la variable explicativa, tanto en un proceso de interpolación como de extrapolación, pero evidencian una muy escasa sensibilidad sobre la relación entre el resultado y el

contexto del problema. Los estudiantes se encuentran atrapados por el modelo matemático utilizado, a través del cual cierta clase de fenómenos de la realidad se traducen a representaciones abstractas. Finalmente hay un proceso de *desconceptualización* por el cual los resultados alcanzados se transforman y adecuan a la realidad original (Ríos, 1.977). Es obvio que los sujetos de la muestra, casi en su totalidad, no llevan a cabo esta última etapa.

Los alumnos no tienen en cuenta las limitaciones propias de todo proceso de extrapolación (Wallace, 1.993; Nicholson, 1.997), pues ningún alumno muestra reserva alguna al aplicarlo. De nuevo los estudiantes aplican un modelo sin cuestionarse su campo de validez.

Tanto en la predicción de Y a partir de X como en la de X a partir de Y emerge el conflicto de la discriminación entre variable independiente y dependiente. También observábamos idéntico hecho en el análisis de los items en el Capítulo 5.

Capítulo 8

Aportaciones y líneas de investigación abiertas

8.1. APORTACIONES DE LA INVESTIGACIÓN

En esta Memoria hemos llevado a cabo un estudio teórico y experimental sobre la enseñanza de la correlación y regresión en los cursos introductorios de estadística a nivel universitario.

El estudio teórico ha comprendido un resumen de los antecedentes de nuestro trabajo, que se presenta en el Capítulo 2, e incluye tanto las investigaciones en el campo de la Psicología como las realizadas desde la Educación Matemática. Consideramos que este estudio es una primera aportación de nuestro trabajo, que puede servir de punto de partida a otros investigadores interesados por los problemas de enseñanza y aprendizaje de la estadística.

El estudio experimental ha tenido dos componentes:

- Un análisis del significado de la correlación y regresión en bachillerato y un curso introductorio de iniciación a la estadística a nivel universitario, llevado a efecto a partir de una muestra de libros de texto de bachillerato, así como de los apuntes del profesor de la asignatura cursada por los alumnos de la muestra y de los apuntes tomados en clase por dos de estas alumnas (Capítulo 3).

- Una evaluación del significado personal que sobre la correlación y regresión ponen de manifiesto los alumnos participantes al finalizar la enseñanza, que comprende su conocimiento conceptual y procedimental, así como la competencia en la estimación de la correlación a partir de sus diversas representaciones y la capacidad de traducción entre estas diversas representaciones.

En este capítulo describimos las principales aportaciones del trabajo, en relación con los objetivos que se especificaron en el Capítulo 1 y sus implicaciones didácticas, finalizando con la descripción de algunos puntos a partir de los cuales podría continuarse esta investigación.

Aportaciones del trabajo en relación con los objetivos de la investigación

Aportaciones respecto al primer objetivo

Como ya exponíamos en el Capítulo 1, el primer objetivo de nuestro trabajo era analizar los contenidos incluidos en el estudio descriptivo de la correlación y regresión, tanto en bachillerato, como en los cursos "típicos" de iniciación a la estadística en la universidad. Partíamos de nuestro trabajo previo (Sánchez Cobo, 1.996), que se ha resumido, organizado y completado con el análisis de los apuntes del profesor y dos de las alumnas participantes en la investigación.

Respecto a este objetivo, en el Capítulo 3 hemos mostrado un análisis detallado de los contenidos y metodología de presentación de la correlación y regresión en los dos niveles de enseñanza citados, de igual modo hemos efectuado el estudio de los ejercicios que se presentaban sobre este tópico en los libros de texto de bachillerato.

Sobre los libros de texto, hemos mostrado las diferentes formas de abordar el tema, así como los distintos itinerarios docentes. Como consecuencia, hemos abogado por una secuenciación que trate primero la correlación y después la regresión, pues aunque a toda nube de puntos podamos asociarle una recta de

regresión, existen casos en que la recta de regresión hallada nos daría predicciones escasamente fiables e información poco relevante (Grupo Azarquiel, 1.985). En la mayoría de los libros de texto y en la programación del profesor se ha optado por una secuenciación contraria, primero se desarrolla la regresión y después la correlación.

Hemos realizado también una aportación novedosa al elaborar una taxonomía de las definiciones de los diversos conceptos (regresión, correlación, covarianza), que nos muestra que apenas se presentan definiciones de tipo intrumento-relacional. Dado que en los textos de bachillerato fundamentalmente se exhiben definiciones de tipo instrumental, pensamos que se pudiera transmitir una visión de las matemáticas como disciplina conformada por una colección de reglas y hechos que deben ser recordados y que se refieren sobre todo al cálculo (Buxton, 1.981). Por el contrario, la programación del profesor y su posterior exposición en el aula presenta una mayoría de definiciones de tipo relacional o intrumento-relacional.

También es novedoso el análisis efectuado de las demostraciones. Aunque en los textos de bachillerato se indican, implícitamente, las componentes de ellas, no obstante, básicamente se alude de forma casi exclusiva a sus funciones de convicción y explicativa, las cuales no tienen interés para los alumnos (De Villiers, 1.993). Puesto que lo *"fundamental del proceso de demostración* (está) *en la habilidad para argumentar de una manera lógica"* (MacNab y Cummine, 1.992, pág. 91), consideramos que el planteamiento anterior pudiera responder a las creencias que los autores poseen sobre las matemáticas. Además, parece cuestionable el desarrollo de demostraciones que implican conceptos que los alumnos no han adquirido, como puede ser el caso de la correspondiente a la determinación de la recta de regresión mínimo-cuadrática, la cual conllevaría el estudio de los extremos relativos de una función de varias variables, noción no explicitada tanto a un nivel de bachillerato (Romero y López, 1.998) como al de universidad, pues cuando los alumnos se enfrentan con el tema de la correlación y regresión únicamente han trabajado en matemáticas con funciones reales de variable real.

En caso de que se estime oportuno la necesidad del aprendizaje de las demostraciones, abogamos que se emplee las demostraciones a dos columnas

(NCTM, 1.991). De igual modo, debe tenerse en cuenta que *"los obstáculos ligados con el aprendizaje de la demostración se encuentran a dos niveles:*

- el enseñante que, por una parte, tiene, en general, dificultades para localizar e identificar los tipos de errores cometidos por los alumnos a fin de formular hipótesis sobre sus concepciones, y por otra parte, en construir situaciones que permitan la emergencia de ciertos procedimientos y el desequilibrio de los procedimientos erróneos.

- el alumno que tiene dificultades para comprender el interés y sentido de la demostración, en encontrar argumentos y en formularlos y articularlos racionalmente" (Ag Almouloud, 1.992, pág. 271).

El estudio de los ejercicios y ejemplos ha permitido mostrar las variables de tarea de los mismos, así como sus carencias y limitaciones. Creemos que este aspecto debe cuidarse en el desarrollo de materiales didácticos para la enseñanza del tema, pues los ejemplos y ejercicios sirven para mostrar al alumno la riqueza de los conceptos y ayudarles a construir concepciones adecuadas sobre los mismos.

El análisis de la enseñanza ha permitido describir una serie de elementos de significado de la correlación y regresión, que conforman el significado institucional del tema en los cursos de bachillerato y cursos introductorios de estadística. Este resultado ha sido utilizado en la construcción de los elementos de evaluación usados en la última parte de la investigación. Estimamos que, asimismo, puede servir de punto de partida para la construcción de otros instrumentos de evaluación o de secuencias didácticas para la enseñanza del tema.

Finalmente, consideramos que el modelo de investigación utilizado para el análisis de los libros de texto puede ser punto de partida a otros trabajos que tengan como campo de aplicación los libros de texto, e igualmente valer como referencia al profesorado para la elaboración de instrumentos que les ayude en la elección de los manuales.

Aportaciones respecto al segundo objetivo

El segundo objetivo de nuestro trabajo era caracterizar el significado personal que los alumnos universitarios dan a la correlación y regresión

estadísticas al finalizar un curso de introducción a la estadística en la universidad. En particular, estábamos interesados en describir los errores conceptuales y procedimentales de estos estudiantes, la capacidad de estimación del coeficiente de correlación a partir de distintas representaciones de la correlación (descripción verbal, tabla, diagrama de dispersión) y de traducción entre estas representaciones.

Este objetivo se ha alcanzado con la evaluación realizada sobre una muestra de 193 alumnos, con los instrumentos construidos para esta investigación y que se describen en el Capítulo 4. Creemos que los mismos instrumentos son otro resultado valioso del trabajo, pues cubren una amplia gama de contenidos conceptuales y procedimentales, y podrían ser empleados, en su totalidad o en parte, en otras investigaciones o para la evaluación del aprendizaje de los alumnos.

Como se indica en el Capítulo 1, la evaluación se ha centrado en los siguientes puntos:

- Comprensión de las propiedades más sobresalientes de la covariación, dependencia estadística, coeficiente de correlación lineal y recta de regresión.
- Errores conceptuales asociados a elementos de significado relacionados con el coeficiente de correlación, covarianza, rectas de regresión, tipos de covariación y relaciones entre correlación y causalidad.
- Estimación de la correlación que hacen los alumnos a partir de diversas representaciones (descripción verbal, diagrama de dispersión, tabla) e interpretación de valores numéricos del coeficiente de correlación y construcción de situaciones asociadas, representadas en forma verbal y gráfica.
- Estrategias de los alumnos en el ajuste de una recta a un conjunto de datos, estimación de valores de las variables y de las rectas de regresión en una situación problemática.

Para cada uno de estos puntos hemos descrito pormenorizadamente los resultados del proceso de evaluación en los Capítulos 5, 6 y 7. Estos resultados incluyen los conocimientos y capacidades exhibidos por los alumnos, así como sus

dificultades y errores. Remitimos al lector a las conclusiones de los Capítulos 5, 6 y 7 para un resumen de estos resultados que, desde nuestro punto de vista, proporcionan una información útil a los profesores encargados de la enseñanza de este tema, que hasta la fecha había recibido una atención muy restringida desde el campo de la educación matemática.

En este sentido juzgamos que nuestra investigación contiene aportaciones originales en la descripción de las dificultades y errores conceptuales y procedimentales de los alumnos, que completan las de otros trabajos previos, particularmente los de Estepa (1.994), Morris (1.997, 1.998) y Truran (1.995, 1.997).

Asimismo, nuestro estudio de las tareas de traducción -Capítulo 6- constituye un complemento del realizado por Janvier (1.978) para el caso de las funciones, tanto desde el punto de vista teórico como práctico. Un punto original es la determinación de las estrategias que los alumnos emplean en estas tareas, que ha sido puesto en relación con las variables utilizadas en el diseño del cuestionario, mediante el análisis de correspondencias.

Hemos diseñado, también, una clasificación de las estrategias usadas por los alumnos en la construcción de un diagrama de dispersión, que se describe, igualmente, en el Capítulo 6, y analizado la dependencia de estas estrategias respecto a las variables de tarea de nuestro estudio.

8.2. IMPLICACIONES PARA LA ENSEÑANZA DEL TEMA Y NUEVAS LÍNEAS DE INVESTIGACIÓN EN EL ÁREA DE DIDÁCTICA DE LAS MATEMÁTICAS

Implicaciones para la enseñanza del tema

Es natural que un trabajo de investigación en el Área de Didáctica de las Matemáticas contenga algunas sugerencias para la enseñanza del tema sobre el que versa. Esta Tesis no es una excepción, ya que en los capítulos que la configuran se han ido presentando distintas reflexiones para insinuar una mejora de la enseñanza de la correlación y de la regresión. En lo que sigue haremos un resumen de estas reflexiones.

Aunque a toda nube de puntos podamos hacerle corresponder una curva de regresión, carecería de interés hallarla si las características de la variable estadística bidimensional no estuvieran correlacionadas o si nos diera predicciones poco fiables e información de escasa trascendencia. Es por ello que propugnamos que la secuenciación en la enseñanza de este tema siga un orden en el que primero se trabaje la correlación, para, después, abordar la regresión.

Abogamos que no se haga una aproximación al concepto de variable estadística bidimensional a partir de dos variables estadísticas unidimensionales, como se efectuaba en algunos manuales de bachillerato, ya que podría favorecer el que los alumnos sean poco conscientes de que dichas variables unidimensionales tienen que estar referidas a la misma unidad estadística, como ha puesto de manifiesto los resultados de esta investigación.

Creemos que debe darse la oportunidad a los alumnos de enfrentarse con situaciones problemáticas que les muestre una diversidad de tipos de covariación, tipos de dependencia e intensidad y signo de la asociación. Esto contribuiría a eliminar ciertas concepciones erróneas en los estudiantes, que se han descrito, por ejemplo, en las investigaciones de Estepa, Batanero, Godino y Morris. En especial, es oportuno que trabajen en profundidad con situaciones que correspondan a una dependencia negativa y de independencia, respecto a las cuales los alumnos han mostrado dificultades singulares (Estepa, 1.994).

Asimismo, debe efectuarse una aproximación al tema de la correlación y de la regresión que haga aflorar la confusión de la asociación y la causalidad. No olvidemos que una de las características fundamentales del razonamiento correlacional es el conocer que una asociación fuerte entre las componentes de una variable estadística bidimensional no conlleva una relación de causa-efecto entre ellas (Garfield, 1.998).

Un punto que debe cuidarse es la interpretación del valor del coeficiente de correlación y su conexión con otros conceptos estadísticos que juegan un rol destacado en este tema -covarianza, tipo de dependencia, dispersión de los datos, etc.-. Por otro lado, se ha de tener en cuenta que al valorar la intensidad de la asociación se produce una inversión del orden del conjunto de los números reales

R, lo cual podría influir, en ciertas situaciones, en la génesis de un obstáculo didáctico.

Dada la dificultad que los sujetos de la muestra han exhibido a la hora de distinguir entre variable independiente y dependiente, abogamos por una intensificación de la enseñanza que tenga como objetivo el discriminar, de forma conveniente, la variable explicativa de la explicada. Esto supondría, además, que se potenciaría la diferenciación entre ambas rectas de regresión, siendo conscientes de la utilización de la recta de regresión X sobre Y, cuando éste sea el modelo estadístico pertinente.

Estimamos que los ejercicios que realicen los alumnos deben plantear ajustes tanto lineales como no lineales, lo cual ayudará a que los estudiantes no establezcan relaciones inadecuadas entre el tipo de ajuste y otros conceptos estadísticos -covarianza, coeficiente de correlación, etc-.

Un aspecto relevante, con respecto a la regresión, es que a veces comporta el trabajar dentro de un marco gráfico, lo cual, como ha puesto de manifiesto la presente Memoria, ocasiona bastantes dificultades a los alumnos. Sería conveniente, por lo tanto, que se reforzara la enseñanza en este sentido.

Por otra parte, también es importante que los ejercicios estimulen tareas fundamentales para la educación estadística como la interpretación, representación gráfica, predicción, comparación y recogida y análisis de datos. Además, dado que *"el buen razonamiento estadístico es tan necesario para ejercer una ciudadanía eficiente como la capacidad de leer y escribir"* (Campbell, 1.981, pág. 14), debemos proponer problemas que sean catalizadores de este tipo de razonamiento, lo cual implicaría que tengan que incluir las tareas anteriormente reseñadas. Como indican Espinel y cols. (1.995), adecuar el sistema educativo a un nuevo contexto social, económico, tecnológico y demográfico demanda herramientas matemáticas útiles para interpretar y resolver problemas en otro área de conocimiento.

Es importante que los ejercicios que tengan que resolver los alumnos sean dados en contexto y con datos reales. Lo primero, aparte de mostrar que la estadística es una ciencia de amplio espectro de aplicación, permitiría a los alumnos realizar tareas como las mencionadas anteriormente, lo cual sería imposible si los problemas se dan descontextualizados. Lo segundo sería un

indicador de la utilidad de la estadística, no siendo óbice el trabajar con ellos dado los medios tecnológicos -calculadoras, ordenadores, etc.- de los que actualmente se puede disponer.

Somos partidarios de que, siempre que las condiciones materiales lo posibiliten, la enseñanza de la estadística se realice con ordenador. Esto permitiría, con respecto a nuestro tema, el trabajar, por ejemplo, con valores atípicos (Nurhonen y Puntanen, 1.992) u observar las múltiples ventajas del software estadístico específico. Tratar estos aspectos mediante las técnicas tradicionales no sería recomendable.

Sugerencias para otras investigaciones

En el desarrollo de nuestra investigación hemos obtenido respuestas parciales a los problemas que en ella se plantearon. Cada una de ellas nos sugiere una línea de investigación futura, que describimos a continuación.

Una de las dificultades mostradas en todos los bloques del cuestionario de esta investigación, es la que tienen los alumnos para discriminar entre la variable explicativa y la variable explicada en una recta de regresión. Puesto que esta diferenciación es básica en las aplicaciones prácticas, creemos que constituye un tema en el que se podría profundizar. Una sugerencia es estudiar si los conocimientos que sobre las funciones poseen los alumnos actúan de forma poco conveniente y pueden constituirse en obstáculo para comprender las diferencias entre las dos rectas de regresión y el papel de ambas variables. Otras posibles explicaciones de esta dificultad son el hecho de que la correlación es una relación simétrica, mientras que la regresión es asimétrica o las dificultades de los alumnos en la interpretación de probabilidades condicionales, idea que está implícita en la recta de regresión.

Se ha elaborado un catálogo de errores conceptuales relacionados con elementos de significado de la asociación, a partir del análisis de los textos de bachillerato y de los apuntes del profesor de la asignatura y dos alumnas. Un camino a seguir, podría ser, completar dicha lista, mediante el análisis de libros de texto a nivel universitario. Asimismo, sugerimos encontrar una interpretación de los

errores asociados a los mismos, y alguna estrategia que posibilite a los estudiantes su superación.

En esta investigación hemos llevado a efecto una primera aproximación a los procesos de traducción entre las diferentes representaciones que podemos establecer sobre la noción de correlación. Consideramos que es conveniente una estudio más intensivo de estas actividades, que nos aporte cuáles son las concepciones previas que tienen los alumnos sobre las mismas. Dado que los resultados obtenidos indican una cierta complejidad de estas tareas para los estudiantes, pensamos que serían necesarias nuevas investigaciones empleando entrevistas y otras técnicas cuantitativas que nos permitan profundizar en los procesos de razonamiento de los estudiantes.

Como la correlación y la regresión son un tema que está incluido tanto en los currícula de nivel universitario como en los nuevos currícula de enseñanza secundaria, una posible vía de investigación sería el diseño de unidades didácticas para estos niveles de enseñanza y su evaluación.

Referencias

ABRAMSON, L. y ALLOY, L. B. (1.980). Judgment of contingency: Errors and their implications. En A. Baum y J. Singer (Eds.), *Advances in Environmental Psychology* (vol. 2, págs. 11-139). Hillsdale, New Jersey: Erlbaum.

ADI, H., KARPLUS, R. y LAWSON, A. (1.978). Intellectual development beyond elementary school VI: Correlational reasoning. *School Science and Mathematics*, 75, 675-683.

AG ALMOULOUD, S. (1.992). Ayuda informática para la resolución de problemas con prueba: Secuencias didácticas para la enseñanza de la demostración. *Recherches en Didactique des Mathématiques*, 12 (2-3), 271-318.

ALLAN, L. G. y JENKINS, H. M. (1.983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, 14, 381-405.

ALLOY, L. B. y TABACHNIK, N. (1.984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, 91, 112-149.

ANDERSON, J. R. y FINCHAM, J. M. (1.996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22 (2), 259-277.

ARAÚJO, J. B. y CHADWICK, C. B. (1.988). *Tecnología educacional. Teorías de instrucción*. Barcelona: Ed. Paidós Educador.

ARKES, H. R. y HARKNESS, A. R. (1.983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General*, 112 (1), 117-135.

ARKES, H. R. y ROTHBART, M. (1.985). Memory, retrieval and contingency judgments. *Journal of Personality and Social Psychology*, 49 (3), 598-606.

AZCÁRATE, C. (1.990). *La velocidad: Introducción al concepto de derivada*. Tesis Doctoral. Universidad Autónoma de Barcelona.

AZCÁRATE, C. y DEULOFEU, J. (1.990). *Funciones y gráficas*. Madrid: Síntesis.

AZORÍN, F. y SÁNCHEZ CRESPO, J. L. (1.986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza.

BABBIE, E. (1.989). *The practice of social research.* Belmont, Ca: Wadsworth Publishing Company.

BARBANCHO, A. G. (1.973). *Estadística elemental moderna*. Barcelona: Ed. Ariel (Cuarta edición. Reimpresión de 1975).

BARDIN, L. (1.986). El análisis de contenido. Madrid: Akal Universitaria.

BATANERO, C., DIAZ GODINO, J. y ESTEPA, A. (1.991). Análisis exploratorio de datos: Sus posibilidades en la Enseñanza Secundaria. *Suma*, 9, 25-31.

BATANERO, C., ESTEPA, A. y GODINO, J. D. (1.991). Estrategias y argumentos en el estudio descriptivo de la asociación usando microordenadores. *Enseñanza de la Ciencia* , 9 (2), 145-150.

BATANERO, C., ESTEPA, A. y GODINO, J. D. (1.995). Correspondence analysis as tool to analyse the relational structure of students' intuitive strategies in judging statistical association. En R. Gras (Ed.), *Méthodes d'Analyses Statistiques Multidimmensionnelles en Didactique des Mathématiques* (págs. 155-166). Rennes: A.R.D.M.

BATANERO, C., ESTEPA, A. y GODINO, J. D. (1.997). Evolution of students' understanding of statistical association in a computer based teaching environment. En J. B. Garfield y G. Burril (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (págs. 191-205). Voorburg: International Statistical Institute.

BATANERO, C., ESTEPA, A. y GODINO, J. D. (1.998). La construcción del significado de la asociación mediante actividades de análisis de datos: reflexiones sobre el papel del ordenador en la enseñanza de la estadística. *II Seminario de la Sociedad Española en Educación Matemática.* Pamplona.

BATANERO, C., ESTEPA, A., GODINO, J. D. y GREEN, D. R. (1.996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, 27 (2), 151-169.

BATANERO, C. y GODINO, J. D. (1.998). Understanding graphical and numerical representations of statistical association in a computer environment. En L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee y W. Wong (Eds.), *Proceedings of the Fifth Conference on Teaching Statistics* (vol. 2, págs. 1.017-1.024). Voorburg: International Statistical Institute.

BATANERO, C., GODINO, J. D. y ESTEPA, A. (1.988). *Curso de estadística aplicada basado en el uso de ordenadores.* Jaén: Los autores.

BATANERO, C., GODINO, J. D. y ESTEPA, A. (1.998). Building the meaning of association through data analysis activities. Research Forum. *22 Conference of the International Group for the Psychology of Mathematics Education.* Stellembosch, Sudáfrica.

BEACH, L. R. y SCOPP, T. (1.966). Inferences about correlations. *Psychonomic Science*, 6, 253-254.

BELL, A. W. (1.976). Stages in generalisation and proof. *Proceedings of the CIMI Conference.* Nyireghaza, Hungría.

BERMAN, J. S. y KENNY, D. A. (1.976). Correlational bias in observer ratings. *Journal of Personality and Social Psychology*, 34, 263-273.

BEYTH-MAROM, R. (1.982). Perception of correlation reexamined. *Memory and Cognition*, 10 (6), 511-519.

BOERO, P. (1.989). Utilización de la Historia de las Matemáticas en clase con alumnos de 6 a 13 años. *Suma* , 2 (1), 17-28.

BOLGER, F. y HARVEY, N. (1.993). Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology*, 46A (4), 779-811.

BOTELLA, L. M. (1.996). La calculadora gráfica en correlación y regresión. *Suma* , 22, 71-78.

BOWER, G. H. y MASLING, M. (1.978). *Causal explanations as mediators for remembering correlations.* Unpublished manuscript. Stanford University.

BOYER, C. B. (1.986). *Historia de la matemática*. Madrid: Alianza Universidad.

BROUSSEAU, G. (1.986). *Théorisation des phénomènes d'enseignement des mathématiques.* Thèse d'Etat. Université de Bordeaux.

BURRILL, G., BURRILL, J. C., COFFIELD, P., DAVIS, G., LANGE, J. DE, RESNICK, D. y SIEGEL, M. (1.992). *Data analysis and statistics across the curriculum.* Reston, Virginia: N.C.T.M.

BUTTS, T. (1.980). Posing problems properly. En S. Krulik (Ed.), *Problem Solving in School Mathematics*, 1.980 Yearbook (págs. 23-33). Reston (Virginia): National Council of Teachers of Mathematics.

BUXTON, L. (1.981). *Do you panic about maths ?* Londres: Heinemann.

CALOT, G. (1.982). *Curso de estadística descriptiva.* Madrid: Paraninfo (3ª edición).

CAMPBELL, S. K. (1.981). *Equívocos y falacias en la interpretación de estadísticas.* Naucalpan de Juárez: Limusa.

CAÑIZARES, Mª J. (1.997). *Influencia del razonamiento proporcional y combinatorio y de las creencias subjetivas en las intuiciones probabilísticas primarias.* Tesis Doctoral. Departamento de Didáctica de la Matemática. Universidad de Granada.

CARMINES, E. G. y ZELLER, R. A. (1.979). *Reliability and validity assessment.* Londres: Sage University Paper.

COBB, G. W. (1.987). Introductory textbooks: A framework for evaluation. *Journal of the American Statistical Association* , 82, 321-339.

COCKCROFT, W. H. (1.985). *Las matemáticas si cuentan (Informe Cockcroft).* Madrid: Ministerio de Educación y Ciencia.

CONTRERAS, A. y SÁNCHEZ, C. (en prensa). Estudio de manuales universitarios de la segunda mitad del siglo XX sobre el concepto de límite de una función, en cuanto a los ejemplos. *VI Simposio de enseñanza e historia de las ciencias.*

COOK, R. D. y WEISBERG, S. (1.994). *An introduction to regression graphics.* New York: John Wiley & Sons, Inc.

CROCKER, J. (1.981). Judgment of covariation by social perceivers. *Psychological Bulletin*, 90 (2), 272-292.

CROCKER, J. (1.982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin*, 8, 214-220.

CRUISE, R. J., DUDLEY, R. L. y THAYER, J. D. (1.984). *A resource guide for introductory statistics.* Dubuque, Iowa: Kendall / Hunt Publishing Company.

CHAPMAN, L. J. (1.967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, 6, 151-155.

CHAPMAN, L. J. y CHAPMAN, J. P. (1.967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, 193-204.

CHAPMAN, L. J. y CHAPMAN, J. P. (1.969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271-280.

CHAPMAN, G. B. y ROBBINS, S. J. (1.990). Cue interaction in human contingency judgment. *Memory and Cognition*, 18 (5), 537-545.

CHATLOSH, D. L., NEUNABER, D. J. y WASSERMAN, E. A. (1.985). Response-outcome contingency: Behavioral and judgment effects of appetitive and aversive outcomes with college students. *Learning and Motivation*, 16, 1-34.

DANE, F. C. (1.990). *Research methods.* Pacific Grow, California: Thomson Information Publishing Group.

DAVIS, P. J. y HERSH, R. (1.988). *Experiencia matemática.* Barcelona: MEC-Labor.

DAVIS, P. J. y HERSH, R. (1.989). *El sueño de Descartès.* Barcelona: MEC-Labor.

DE VILLIERS, M. (1.993). El papel y la función de la demostración en matemáticas. *Epsilon* , 26, 15-30.

DEWDNEY, A. K. (1.985). Juegos de ordenador. *Investigación y Ciencia* , 107, 87-93.

DICKINSON, A., SHANKS, D. y EVENDEN, J. (1.984). Judgment of act-outcome contingency: the role of selective attribution. *Quarterly Journal of Experimental Psychology*, 36A, 29-50.

DOUADY, R. (1.986). Jeux de cadres et dialectique outil-objet. *Recherches en Didactique des Mathématiques*, 7 (2), 5-31.

DUVAL, R. (1.993). Semiosis et Noesis. *Lecturas en Didáctica de la Matemática: Escuela Francesa.* México: Sección de Matemática Educativa del CINVESTAV-IPN.

DUVAL, R. (1.995). Semiosis et Pensée humaine. Bern: Peter Lang SA.

ERLICK, D. E. y MILLS, R. G. (1.967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology*, 73 (1), 9-14.

ESPINEL, Mª C., BRUNO, A. y GARCÍA CRUZ, J. A. (1.995). Diagramas para visualizar desigualdades y clasificaciones. *Uno*, 5, 57-66.

ESTEPA, A. (1.990). *Enseñanza de la Estadística basada en el uso de ordenadores: Un estudio exploratorio.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

ESTEPA, A. (1.994). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores.* Tesis Doctoral. Departamento de Didáctica de la Matemática. Universidad de Granada.

ESTEPA, A. (1.995a). Algunas consideraciones sobre la enseñanza de la asociación estadística. *Uno. Revista de Didáctica de las Matemáticas*, 5, 69-79.

ESTEPA, A. (1.995b). Las tablas de contingencia y su enseñanza. ¿Qué podemos aprender de las investigaciones? *Uno* , 3, 89-100.

ESTEPA, A. y BATANERO, C. (1.994). Judgments of association in scatterplots. En *Proceedings of the Fourth International Conference Psychology on Teaching Statistics* (pág. 587). Marruecos: The National Institute of Statistics and Applied Economics. (Copia completa -8 páginas- en J. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics.* The International Study Group for Research on Learning Probability and Statistics.

ESTEPA, A. y BATANERO, C. (1.995). Concepciones iniciales sobre la asociación estadística. *Enseñanza de las Ciencias*, 13 (2), 155-170.

ESTEPA, A. y BATANERO, C. (1.996). Judgments of correlation in scatter plots: student's intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education*, 4, 25-41.

ESTEPA, A., GREEN, D. R., BATANERO, C. y GODINO, J. D. (1.994). Judgments of association in contingency tables. An empirical study of students' strategies and preconception. En J. P. Ponte y J. F. Matos (Eds.), *Proceedings of the XVIII International Conference on the Psychology of Mathematics Education* (vol. 2, págs. 312-319). Universidad de Lisboa.

ESTEPA, A. y SÁNCHEZ COBO, F. T. (1.994). Desarrollo histórico de la idea de asociación estadística. *Epsilon*, 30, 61-74.

ESTEPA, A. y SÁNCHEZ COBO, F. T. (1.996a). Association judgements in the comparison of two samples. En L. Puig y A. Gutiérrez (Eds.), *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education*, (v. 2, págs. 337-344). Universidad de Valencia.

ESTEPA, A. y SÁNCHEZ COBO, F. T. (1.996b). Sesgos en la enseñanza de la asociación. *Book of abstracts of short presentation ICME VIII* , (pág. 426). Sevilla.

ESTEPA, A. y SÁNCHEZ COBO, F. T. (1.998). Correlation and regression in secondary school text books. En L. Pereira-Mendoza, L. Seu, T. Wee y W. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics* (vol. 2, págs. 671-676). Voorburg: International Statistical Institute.

EVANS, J. St. B. T. (1.982). On statistical intuitions and inferential rules. A discussion of Kahneman and Tversky. *Cognition*, 12, 323-326.

FALK, R. (1.986). Conditional probabilities: Insights and difficulties. En R. Davidson y J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (págs. 292-297). University of Victoria.

FALK, R. y WELL, A. D. (1.997). Many faces of the correlation coefficient. *Journal of Statistics Education*, 5(3) (http: www.stat.ncsu.edu/info/jse).

FERNÁNDEZ, A. y RICO, L. (1.992). *Prensa y educación matemática.* Madrid: Síntesis.

FIOL, Mª L. y FORTUNY, J. M. (1.990). *Proporcionalidad directa. La forma y el número.* Madrid: Síntesis.

FOX, D. J. (1.987). *El proceso de investigación en educación.* Pamplona: Universidad de Navarra (2ª edición).

FRANKLIN, L. A. (1.988). Clarifying regression concepts using 3 point data sets. *Teaching Statistics*, 10 (1), 8-12.

FREUDENTHAL, H. (1.991). *Revisiting Mathematics Education.* Dordrecht: Kluwer.

GAL, I. y GARFIELD, J. (1.997). Curricular goals and assessment challenges in statistical education. En I. Gal y J. B. Garfield (Eds.), *The assessment challenge in statistical education.* Amsterdam: IOS Press and International Statistical Institute.

GARFIELD, J. (1.998). The statistical reasoning assessment: Development and validation of a research tool. En L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee y W. Wong (Eds.), *Proceedings of the Fifth Conference on Teaching Statistics* (vol. 2, págs. 781-786). Voorburg: International Statistical Institute.

GARFIELD, J. y AHLGREN, A. (1.988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, 19 (1), 44-63.

GIMÉNEZ, J. (1.986). Una aproximación didáctica a las fracciones egipcias. *Números*, 14, 57-62.

GIMÉNEZ, J. (1.988). ¿Proporcionalidad o razonamiento proporcional? *Números*, 18, 19-44.

GIMÉNEZ, J., RICO, L., GIL, F., FERNÁNDEZ, F., CASTRO, E., DEL OLMO, A., MORENO, F. y SEGOVIA, I. (1.997). ¿Por qué y para qué evaluar en Matemáticas? En J. Giménez (Ed.), *Evaluación en Matemáticas. Una integración de perspectivas* (págs. 15-38). Madrid: Síntesis.

GIMENO, L. (1.985). Estudio de algunos modelos interdisciplinares basados en matemáticas de 2º de bachillerato. *Aspectos didácticos de matemáticas* (vol. 1, págs. 11-30). Zaragoza: I.C.E.

GODINO, J. D. y BATANERO, C. (1.994). Significado institucional y personal de los objetos matemáticos. *Recherches en Didactique des Mathématiques*, 14 (3), 325-355.

GODINO, J. D., BATANERO, C. y ESTEPA, A. (1.990). Estrategias y argumentos en el estudio descriptivo de la asociación usando microordenadores. En G. Booker, P. Cobb y T. N. Mendicuti (Eds.), *Proceedings of the XIV International Conference of Psichology of Mathematics Education* (págs. 157-164). México: P.M.E. Program Committee.

GODINO, J. D., BATANERO, C. y ESTEPA, A. (1.991). Task variables in statistical problem solving using computers. En J. P. Ponte y J. F. Matos (Eds.), *Mathematical problem solving and new information technologies. Research in Contexts of Practice* (págs. 193-203). Berlin: Springer-Verlag.

GOLDSTEIN, L. J., LAY, D. C. y SCHNEIDER, D. I. (1.990). *Cálculo y sus aplicaciones.* México: Prentice-Hall Hispanoamericana S.L. (4ª edición).

GONZÁLEZ, J. L.; IRIARTE, M. D.; JIMENO, M.; ORTIZ, A.; ORTIZ, A.; SANZ, E.; VARGAS-MACHUCA, I. (1990). *Números enteros.* Madrid: Ed. Síntesis.

GONZÁLEZ, R. Mª (1.993). *A descriptive study of verbal problems in selected mathematics textbooks at the high school.* Ph. D. Dissertation. State University of New York at Buffalo.

GOODE, S. M. y GOLD, E. J. (1.987). Linear regression and correlation-An enlightening approach. *Teaching Statistics*, 9 (2), 60-62.

GREENACRE, M. J. (1.984). *Theory and applications of correspondence analysis.* London: Academic Press.

GREENACRE, M. J. y HASTIE, T. (1.987). The geometric interpretations of correspondence analysis. *Journal of the American Statistical Association*, 82, 398, págs. 437-447.

GRUPO AZARQUIEL (1.985). *Regresión y correlación una introducción intuitiva.* Madrid: Monografías del I.C.E.

GUTIÉRREZ CABRIA, S. (1.994). *Filosofía de la estadística.* Valencia: Universidad de Valencia.

GUZMÁN, M. DE, COLERA, J. y SALVADOR, A. (1.988). *Matemáticas 3º de B.U.P.* Madrid: Anaya.

HAMILTON, D. L. y ROSE, T. R. (1.980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology*, 39, 832-845.

HAWKINS, A. (1.991). Success and failure in statistical education - An U.K. perspective. En D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (vol. 1, págs. 24-32). Voorburg: International Statistical Institute.

HAWKINS, A.; JOLLIFFE, F.; GLICKMAN, L. (1992). *Teaching Statistical Concepts.* New York: Logman.

HERMOSO, J. A. y HERNÁNDEZ, A. (1.989). *Introducción a la Estadística.* Granada: Los autores.

HOFFMANN, L. D. y BRADLEY, G. L. (1.994). *Cálculo aplicado a administración, economía, contaduría y ciencias sociales.* Santafé de Bogotá: McGraw-Hill (5ª edición).

HOUSE, P. A., WALLACE, M. L. y JOHNSON, M. A. (1.983). Problem solving as a focus: How ? When ? Whose responsibility ? En G. Shufelt y J. R. Smart (Eds.), *The agenda in action* (págs. 9-19). Reston, Virginia: N.C.T.M.

HUBERMAN, A. M. y MILES, M. B. (1.994). Data management and analysis methods. En N. K. Denzin y Y. S. Lincoln (Eds.), *Handbook of qualitative research* (págs. 445-462). London: Sage.

INHELDER, B. y PIAGET, J. (1.955). *De la logique de l'enfant à la logique de l'adolescent.* Paris: Presses Universitaires de France (Traducción castellana, Primera reimpresión, 1.985. Barcelona: Paidós).

JANVIER, C. (1.978). *The interpretation of complex cartesian graph representing situations, studies and teaching experiments.* Tesis Doctoral. Universidad de Québec.

JANVIER, C. (1.987). Procesos de traducción en educación matemática. En C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (págs. 27-32). Londres: LEA Publ. (Traducción de Moisés Coriat Benarroch).

JENKINS, H. M. y WARD, W. C. (1.965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, 79, 1-17.

JENNINGS, D. L., AMABILE, T. M. y ROSS, L. (1.982). Informal covariation assessment: Data-based versus theory-based judgments. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (págs. 211-230). New York: Cambridge University Press.

JONG, P. J. DE, MERCKELBACH, H. y ARNTZ, A. (1.995). Covariation bias in phobic women: The relationship between a priori expectancy, on-line expectancy, autonomic responding and a posteriori contingency judgment. *Journal of Abnormal Psychology*, 104 (1), 55-62.

JULLIEN, M. y NIN, G. (1.989). L'E.D.A. au secours de L'O.G.D. ou quelques remarques concernant l'énseignement de la statistique dans les colleges. *Petit X*, 19, 29-41.

KAPUT, J. J. (1.987). Representation systems and mathematics. En C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (págs. 19-26). Hillsdale: LEA.

KAREEV, Y. (1.995). Positive bias in the perception of covariation. *Psychological Review*, 102 (3), 490-502.

KELLEY, H. (1.973). The process of causal attribution. *American Psychologist*, 28, 107-128.

KILPATRICK, J. (1.978). Variables and methodologies in research on problem solving. En L. Hatfield y D. Brandbard (Eds.), *Mathematical Problem Solving: Papers from a research workshop.* Columbus, Ohio: Eric-Smeac.

KLINGER, M. R. y GREENWALD, A. G. (1.995). Unconscious priming of association judgments. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21 (3), 569-581.

KONOLD, C., POLLATSEK, A., WELL, A. y GAGNON, A. (1.996). Students analyzing data: Research of critical barriers. En J. B. Garfield y G. Burrill (Eds.), *Research on the Role of Technology in Teaching ang Learning Statistics* (págs. 151-169). Voorburg: International Statistical Institute.

KRIGOWSKA, A. Z. (1.971). Treatment of the axiomatic method in class. En W. Servais y T. Varga (Eds.), *Teaching school mathematics* (págs. 124-150). London: Penguin-Unesco.

KRIPPENDORFF, K. (1.990). *Metodología de análisis de contenido. Teoría y práctica.* Barcelona: Paidós.

LACASTA, E. (1.995). Les graphiques cartésiens de fonctions dans l'enseignement secondaire des mathématiques: Illusions et contrôles. Tesis Doctoral. Université Bordeaux I.

LACASTA, E. y BROUSSEAU, G. (1.995). Utilisation de la contingence par l'analyse factorielle. Traitement d'un cas: Le graphique. En R. Gras (Ed.), *Méthodes d'analyses statistiques multidimmensionnelles en didactique des mathématiques* (págs. 53-90). Rennes: A.R.D.M.

LAKATOS, I. (1.986). *Pruebas y refutaciones.* Madrid: Alianza Universidad.

LANE, D. M., ANDERSON, C. A. y KELLAM, K. L. (1.985). Judging the relatedness of variables: The psychophisics of covariation detection. *Journal of Experimental Psychology. Perception and Performance*, 11 (5), 640-649.

LAVIOLETTE, M. (1.994). Linear regression: The computer as a teaching tool. *Journal of Statistics Education*, 2 (2) (http: www2.ncsu. edu/ncsu/pams/stat/info/jse).

LÓBEZ URQUÍA, J. y CASA ARUTA, E. (1.975). *Estadística intermedia.* Madrid: Vicens-Vives.

LÓPEZ CACHERO, M. (1.990). *Fundamentos y métodos de Estadística.* Madrid: Pirámide (9ª edición).

LÓPEZ FEAL, R. (1.986). *Construcción de instrumentos de medida en ciencias conductuales y sociales.* Barcelona: Alamex.

MacNAB, D. S. y CUMMINE, J. A. (1.992). *La enseñanza de las matemáticas de 11 a 16.* Madrid: Visor.

MARTÍNEZ, P. S. (1.991). *Correlación y regresión.* Sevilla: Consejería de Educación y Ciencia de la Junta de Andalucía.

MARTINÓN, A. y SAURET, Mª D. (1.990). La asimetría en el aprendizaje de las Matemáticas. *Epsilon*, 21, 45-46.

MENDENHALL, W., SCHEAFFER, R. L. y WACKERLY, D. D. (1.986). *Estadística matemática con aplicaciones.* México: Grupo Editorial Iberoamérica.

MILES, M. B. y HUBERMAN, A. M. (1.984). *Qualitative data analysis: A sourcebook of new methods.* Londres: Sage Publications.

MOORE, D. S. (1.990). Uncertainty. En L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (págs. 95-137). Washington, D.C.: National Academy Press.

MOORE, D. S. (1.995). *The basic practice of statistics.* New York: Freeman.

MORRIS, E. J. (1.997). An investigation of students' conceptions and procedural stills in the statistical topic correlation. *Centre for Information Technology in Education*, Report nº 230. The Open University.

MORRIS, E. J. (1.998). Link: The principled design of a computer assisted learning program for correlation. En L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee y W. Wong (Eds.), *Proceedings of the Fifth Conference on Teaching Statistics* (vol. 2, págs. 1.033-1.040). Voorburg: International Statistical Institute.

MUÑOZ, J. y PASCUAL, A. (1.986). Análisis singular de un conjunto de datos. Las observaciones outliers. *Thales* , 6, 33-39.

MURPHY, G. L. y MEDIN, D. L. (1.985). The role of theories in conceptual coherence. *Psichological Review*, 92 (3), 289-316.

NAVARRO-PELAYO, V. (1.991). *La enseñanza de la combinatoria en Bachillerato.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

NAVARRO-PELAYO, V. (1.994). *Estructura de los problemas combinatorios simples y del razonamiento combinatorio en alumnos de secundaria.* Tesis Doctoral. Departamento de Didáctica de la Matemática. Universidad de Granada.

N.C.T.M. (1.983). *The agenda in action.* Virginia. N.C.T.M.

N.C.T.M. (1.987). *How to evaluate mathematics textbooks.* Virginia. N.C.T.M. (1ª edición 1.982).

N.C.T.M. (1.991). *Estándares curriculares y de evaluación para la educación matemática.* Sevilla: S.A.E.M. Thales.

NICHOLSON, J. (1.997). Developing probabilistic and statistical reasoning at secondary level through the use of data and technology. En J. B. Garfield y G. Burril (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (págs. 29-44). Voorburg: International Statistical Institute.

NISBETT, R. y ROSS, L. (1.980). *Human inference: strategies and shortcomings of social judgment.* Nueva Jersey: Prentice Hall.

NISHISATO, S. (1.980). *Analysis of categorical data: Dual scaling and its applications.* Toronto: University of Toronto Press.

NODA, Mª A. y ESPINEL, Mª C. (1.992). Análisis de datos a través de métodos gráficos. *Números*, 22, 29-36.

NORTES CHECA, A. (1.987). *Encuestas y precios.* Madrid: Síntesis.

NURHONEN, M. y PUNTANEN, S. (1.992). Illustrating regression concepts. *Teaching Statistics*, 14 (1), 20-23.

ORTEGA MARTÍNEZ, A. R. (1.991). *Contingencia y juicios de covariación en humanos.* Tesis Doctoral. Departamento de Psicología Experimental y Fisiología del Comportamiento. Universidad de Granada.

ORTIZ, J. J. (1.996). *Significado de los conceptos probabilísticos elementales en los textos de bachillerato.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

ORTON, A. (1.990). *Didáctica de las matemáticas.* Madrid: MEC-Morata.

PÉREZ ECHEVERRÍA, M. P. (1.990). *Psicología del razonamiento probabilístico.* Madrid: Ediciones de la Universidad Autónoma de Madrid.

PETERSON, C. R. (1.980). Recognition of non contingency. *Journal of Personality and Social Psychology*, 38 (5), 727-734.

PFANNKUCH, M. y BROWN, C. (1.996). Building on and challenging students' intuitions about probability: Can we improve undergraduate learning ? *Journal of Statistics Education*, 4 (1) (http: www2.ncsu.edu/ncsu/pams/stat/info/jse).

PHILLIPS, J. L. (1.992). *How to think about statistics.* New York: W.H. Freeman and Company.

PIAGET, J. y GARCÍA, R. (1.973). *Las explicaciones causales.* Barcelona: Barral.

PORKESS, R. (1.996). Bivariate data: Lessons from students' coursework. *Teaching Statistics*, 18 (3), 76-80.

POZO, J. I. (1.987). *Aprendizaje de la ciencia y pensamiento causal.* Madrid: Visor.

PRICE, P. C. y YATES, J. F. (1.995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21 (6), 1.639-1.655.

RICO, L. (1.990). Diseño curricular en Educación Matemática: Una perspectiva cultural. En S. Llinares y Mª V. Sánchez (Eds.), *Teoría y práctica en Educación Matemática* (págs. 17-61). Sevilla: Alfar.

RICO, L. y CASTRO, E. (1.994). Difficulties and errors in number reasoning develpment. En N. A. Malara y L. Rico (Eds.), *Proceedings of the First Italian-Spanish Research Symposium in Mathematics Education* (págs. 123-130). Universitá di Modena.

RICO, L., CASTRO, E. y ROMERO, I. (1.996). The role of representation systems in the learning of numerical structures. En L. Puig y A. Gutiérrez (Eds.), *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education* (vol 1, págs. 87-102). Universidad de Valencia.

RÍOS, S. (1.976). *Análisis estadístico aplicado.* Madrid: Paraninfo.

RÍOS, S. (1.977). *Métodos estadísticos.* Madrid: Castillo.

RIUS, F., BARÓN, J., PARRAS, L. y SÁNCHEZ, E. (1.997). *Bioestadística: Métodos y aplicaciones.* Málaga: Servicio de Publicaciones e Intercambio Científico de la Universidad de Málaga.

ROBERT, A. y ROBINET, J. (1.989). *Enoncés d'exercices de manuels de seconde et représentations des auteurs de manuels*. IREM. Université Paris VII.

ROMERO, J. B. y LÓPEZ, Mª A. (1.998). Aspectos geométricos de la regresión y correlación lineal. *Números*, 35, 32-43.

ROSS, J. A. y SMYTH, E. (1.995). Thinking skills for gifted students: The case for correlational reasoning. *Roeper Review*, 17 (4), 239-243.

RUBIN, A. (1.989). Reasoning under uncertainty: developing statistical reasoning. *Journal of Mathematical Behavior*, 8, 205-219.

RUIZ HIGUERAS, L. (1.991). *Una aproximación a las concepciones de los alumnos de secundaria sobre la noción de función*. Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

RUIZ HIGUERAS, L. (1.994). *Concepciones de los alumnos de Secundaria sobre la noción de función: Análisis epistemológico y didáctico.* Tesis Doctoral. Departamento de Didáctica de la Matemática. Universidad de Granada.

SÁNCHEZ, E. (1.996a). *Conceptos teóricos e ideas espontáneas sobre la noción de independencia estocástica en profesores de bachillerato: Un estudio de casos.* Tesis Doctoral. CINVESTAT. México.

SÁNCHEZ, E. (1.996b). Dificultades en la comprensión del concepto de eventos independientes. En F. Hitt (Ed.), *Investigaciones en Matemática Educativa* (págs. 389-404). México: Grupo Editorial Iberoamericano.

SÁNCHEZ COBO, F. T. (1.996). *Análisis de la exposición teórica y de los ejercicios de correlación y regresión en los textos de bachillerato.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

SÁNCHEZ COBO, F. T. y ESTEPA, A. (1.996). Análisis de ejercicios de correlación y regresión en libros de texto de bachillerato. En M. de la Fuente y M. Torralbo (Eds.), *VII Jornadas Andaluzas de Educación Matemática "THALES"* (págs. 303-316). Córdoba.

SÁNCHEZ COBO, F. T. y ESTEPA, A. (1.997a). Estudio de la presentación de la correlación en los libros de texto. En L. García Areitio (Ed.), *El material impreso en la enseñanza a distancia* (págs. 287-297). Madrid: U.N.E.D.

SÁNCHEZ COBO, F. T. y ESTEPA, A. (1.997b). Demostraciones y definiciones en Enseñanza Secundaria. En Sociedad Castellano-Leonesa de Profesorado de Matemáticas (Ed.), *VIII Jornadas para el aprendizaje y la enseñanza de las Matemáticas* (págs. 507-511). Salamanca: HERGAR S.L.

SÁNCHEZ COBO, F. T. y ESTEPA, A. (1.998). La regresión en los libros de texto de Secundaria. En F. J. Muñoz, D. Cárdenas y A. J. López (Eds.), *VIII Jornadas Andaluzas de Educación Matemática "THALES"* (págs. 333-340). Universidad de Jaén.

SANZ, I. (1.990). Comunicación, lenguaje y matemáticas. En S. Llinares y Mª V. Sánchez (Eds), *Teoría y práctica en educación matemática* (págs. 173-235). Sevilla: Alfar.

SAX, G. (1.989). *Principles of educational and psychological measurement and evaluation.* Belmont, Ca: Wadsworth Publishing Company.

SCHOENFELD, A. (1.988). Problem solving in context(s). En R. I. Charles y E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (págs. 82-92). Reston, Virginia: Lawrence Erlbaum, N.C.T.M.

SCHUYTEN, G. (1.991). Statistical thinking in psychology and education. En D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (vol 2., págs. 486-489). Voorburg: International Statistical Institute.

SCOTT, P. (1.989). *Introducción a la investigación y evaluación educativa. Lecturas en Educación Matemática.* México: UNAM.

SERRANO, L. (1.993). *Aproximación frecuencial a la enseñanza de la probabilidad y conceptos elementales sobre conceptos estocásticos: un estudio de concepciones iniciales.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemáticas. Universidad de Granada.

SERRANO, L. (1.996). *Significados institucionales y personales de objetos matemáticos ligados a la aproximación frecuencial de la enseñanza de la probabilidad.* Tesis Doctoral. Departamento de Didáctica de la Matemática. Universidad de Granada.

SHAKLEE, H. (1.979). Bounded rationality and cognitive development: upper limits on growth ? *Cognitive Psychology*, 11, 327-335.

SHAKLEE, H. (1.983). Human covariation judgment: Accuracy and strategy. *Learning and Motivation*, 14, 433-448.

SHAKLEE, H. y MIMS, M. (1.982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 8 (3), 208-224.

SHAKLEE, H. y TUCKER, D. (1.980). A rule analysis judgments of covariation between events. *Memory and Cognition*, 8 (5), 459-467.

SHANKS, D. R. (1.989). Selectional processes in causality judgment. *Memory and Cognition*, 17 (1), 27-34.

SHUARD, H. y ROTHERY, A. (1.988). *Children reading mathematics*. London: John Murray.

SHUTE, V. y GAWLICK-GRENDELL, L. (1.993). An experiential approach to teaching and learning probability: Stat lady. En P. Brna, S. Ohlsson y H. Pain (Eds.), *Artificial Intelligence in Education* (págs. 177-184). *Proceedings of AIED 93, World Conference on Artificial Intelligence in Education*. Edinburg August. Charlottesville, VA: Association for the Advancement of Computing in Education.

SKEMP, R. (1.980). *Psicología del aprendizaje de las matemáticas*. Madrid: Morata.

SMEDLUND, J. (1.963). The concept of correlation in adults. *Scandinavian Journal of Psychology*, 4, 165-174.

SMIRNOV, N. V. y DUNIN-BARKOWSKIJ, I. V. (1.978). *Cálculo de probabilidades y estadística matemática*. Madrid: Paraninfo.

STEINBRING, H. (1.991). The concept of chance in everyday teaching: A study of a social epistemology of mathematical knowledge. *Educational Studies in Mathematics*, 22, 503-522.

TAMURA, H. (1.994). Model comparison in regression. *Teaching Statistics*, 16 (2), 47-49.

THORNDIKE, R. L. (1.989). *Psicometría aplicada.* México: Limusa.

TOMARKEN, A. J., SUTTON, S. K. y MINEKA, S. (1.995). Fear-relevant illusory correlations: What types of associations promote judgmental bias. *Journal of Abnormal Psychology*, 104 (2), 312-326.

TROLIER, T. K. y HAMILTON, D. L. (1.986). Variables influencing judgments of correlational relations. *Journal of Personality and Social Psychology*, 50 (5), 879-888.

TRURAN, J. M. (1.995). Some undergraduates' understanding of the meaning of a correlation coefficient. En B. Atweh y S. Flavel (Eds.), *MERGA 18: Galtha* (págs. 524-529). *Proceedings of the Eigteenth Annual Conference of the Mathematics Education Research Group of Australasia* (MERGA). Northern Territory University, Darwin, Australia.

TRURAN, J. M. (1.997). Understanding of association and regression by first year economics students from two different countries as revealed in responses to the same examination questions. En J. B. Garfield y J. M. Truran (Eds.), *Research Papers on Stochastics Educations from 1.997* (págs. 205-212). Department Educational Psychology University of Minnesota.

TUKEY, J. W. (1.977). *Exploratory data analysis.* Nueva York: Addison-Wesley.

TVERSKY, A. y KAHNEMAN, D. (1.982a). Causal schemas in judgments under uncertainty. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (págs. 117-128). Nueva York: Cambridge University Press.

TVERSKY, A. y KAHNEMAN, D. (1.982b). Judgment under uncertainty: heuristics and biases. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (págs. 3-20). Nueva York: Cambridge University Press.

TVERSKY, A., SATTATH, S. y SLOVIC, P. (1.988). Contingent weighting in judgment and choice. *Psychological Review*, 95(3), 371-384.

VALLECILLOS, A. (1.992). *Nivel de significación en un contraste de hipótesis: estudio teórico-experimental de errores en estudiantes universitarios.* Memoria de Tercer Ciclo. Departamento de Didáctica de la Matemática. Universidad de Granada.

VALLECILLOS, A. (1.994). *Un estudio teórico-experimental de errores y concepciones sobre el contraste estadístico de hipótesis en estudiantes universitarios.* Tesis Doctoral. Universidad de Granada.

VALLÉE-TOURANGEAU, F., BAKER, A. G. y MERCIER, P. (1.994). Discounting in causality and covariation judgements. *The Quarterly Journal of Experimental Psychology*, 47B (2), 151-171.

VÁZQUEZ, C. (1.987). Judgment of contingency: cognitive biases in depressed and nondepressed subjects. *Journal of Personality and Social Psychology*, 52 (2), 419-431.

VIZMANOS, J. R. y ANZOLA, M. (1.988). *Matemáticas II. Opción C: Ciencias Sociales. Opción D: Humanística / Lingüística.* Madrid: SM.

WALLACE, E. (1.993). Exploring regression with a graphing calculator. *The Mathematics Teacher*, 86 (9), 741-743.

WARD, W. C. y JENKINS, H. M. (1.965). The display of information and the judgment of contingency. *Canadian Journal of Psychology*, 19, 231-241.

WASSERMAN, E. A., CHATLOSH, D. L. y NEUNABER, D. J. (1.983). Perception of causal relation in humans: Factors affecting judgment of response-outcome contingencies under free-operant procedures. *Learning and Motivation*, 14, 406-432.

WASSERMAN, E. A. y SHAKLEE, H. (1.984). Judging response-outcome relations: The role of response-outcome contingency, outcome probability and method of information presentation. *Memory and Cognition*, 12 (3), 270-283.

WELL, A. D., BOYCE, S. J., MORRIS, R. K., SHINJO, M. y CHUMBLEY, J. I. (1.988). Prediction and judgment as indicators of sensitivity to covariation of continuous variables. *Memory and Cognition*, 16 (3), 271-280.

WILD, C. y PFANNKUCH, M. (1.998). What is statistical thiking ? En L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee y W. Wong (Eds.), *Proceedings of the Fifth Conference on Teaching Statistics* (vol. 1, págs. 333-339). Voorburg: International Statistical Institute.

WILLET, J. B. y SINGER, J. D. (1.992). Providing a statistical "model" teaching applied statistics using real-world. En F. Gordon y S. Gordon (Eds.), *Statistics for the twenty-first century* (págs. 83-98). Nueva Jersey. The Mathematical Association of American.

WRIGHT, J. C. y MURPHY, G. L. (1.984). The utility of theories in intuitive statistics: The robustness of theory-based judgments. *Journal of Experimental Psychology General*, 113, 2, 301-322.

YATES, J. F. y CURLEY, S. P. (1.986). Contingency judgment: Primacy effects and attention decrement. *Acta Psychologica*, 62, 293-302.

Anexos

Anexo I

Datos de la muestra

Tabla I-1. Frecuencia y porcentaje de la edad de los sujetos de la muestra

Edad	Frecuencia	Porcentaje
18	68	35'8
19	50	26'3
20	28	14'7
21	10	5'3
22	11	5'8
23	8	4'2
24	4	2'1
25	2	1'1
26	1	0'5
27	1	0'5
28	1	0'5
32	1	0'5
34	1	0'5
35	2	1'1
37	2	1'1
Total	190	100

Tabla I-2. Frecuencia y porcentaje del sexo y titulación de los sujetos de la muestra

Titulación / Sexo	Diplomatura C. Empresariales	Diplomatura en Enfermería	Total
Varón	37	20	57 (29'5 %)
Hembra	67	69	136 (70'5 %)
Total	104 (53'9 %)	89 (46'1 %)	193 (100 %)

Tabla I-3. Frecuencia y porcentaje de la forma de acceso a la universidad de los sujetos de la muestra

Forma de acceso	Frecuencia	Porcentaje
COU (Opción C) Ciencias Sociales	42	21'9
COU (Opción A) Científico-Tecnológica	46	24'0
COU (Opción B) Biosanitaria	41	21'4
Formación Profesional II/Administrativa y C.	21	10'9
Bachillerato Tecnológico (Reforma)	9	4'7
Técnico en Informática de Gestión	1	0'5
Bachillerato Ciencias Sociales (LOGSE)	2	1'0
Mayores de 25 años	3	1'6
Formación Profesional/Técnico Laboratorio	8	4'2
Formación Profesional/Dietética y Nutrición	2	1'0
Formación Profesional/Anatomía Patológica	7	3'6
Formación Profesional/Radiodiagnóstico	3	1'6
Diplomatura en Educación Infantil	1	0'5
Formación Profesional/Sanitaria	3	1'6
Bachillerato C. Naturales (LOGSE)	1	0'5
Bachillerato C. de la Salud (LOGSE)	2	1'0
Total	192	100

Tabla I-4. Frecuencia y porcentaje de los estudios de Estadística realizados por los sujetos de la muestra en cursos anteriores

Estudios previos de Estadística	Frecuencia	Porcentaje
No ha estudiado Estadística nunca	117	60'6
1º Diplomatura en Ciencias Empresariales	5	2'6
1º Licenciatura en Administración y Dir.	3	1'6
1º de BUP	7	3'6
2º de BUP	5	2'6
3º de BUP	2	1'0
COU	16	8'3
1º Formación Profesional I	5	2'6
3º de Formación Profesional II	15	7'8
1º Licenciatura en Ciencias Matemáticas	1	0'5
4º de Formación Profesional II	5	2'6
Licenciatura en Farmacia	1	0'5
1º Ingeniería T. Industrial (Mecánica)	1	0'5
2º de Formación Profesional I	7	3'6
Diplomatura en Enfermería	2	1'0
Mayores de 25 años	1	0'5
Total	193	100

Anexo II

Libros de texto empleados en la investigación

[77A] AGUSTÍ, J. M. y VILA, A. (1.977). *Matemáticas Tercer Curso.* Barcelona: Vicens-Vives.

[77B] MARTÍNEZ, P., FUERTES, A., GONZÁLEZ, A. y JIMÉNEZ, L. (1.977). *Matemáticas 3º B.U.P.* San Sebastián: Didascalia.

[77 C] BOADAS, J., ROMERO, R. y VILLALBÍ, R. (1.977). *Matemáticas 3er Curso de B.U.P.* Barcelona: Teide.

[77D] GARCÍA, J. y LÓPEZ, M. (1.977). *Matemáticas Tercer Curso de Bachillerato.* Alcoy: Marfil.

[78A] MARCOS DE LANUZA, F. (1.978). *Matemáticas Curso Tercero Bachillerato Unificado Polivalente.* Madrid: G. del Toro.

[81A] VIZMANOS, J. R., ANZOLA, M. y PRIMO, A. (1.981). *Funciones 3. Matemáticas 3º B.U.P. Teoría y Problemas.* Madrid: S.M.

[82A] NEGRO, A. y BENEDICTO, C. (1.982). *Matemática 3º B.U.P.* Madrid: Alhambra.

[86A] CARUNCHO, J., VÁZQUEZ, C. Y GIL, J. (1.986). *Matemáticas B.U.P. 3.* Madrid: Santillana.

[87A] ÁLVAREZ, F., CASTILLO, J. J., CURIEL, F. J., FERNÁNDEZ, T., GARCÍA, L., GARRIDO, L. M., HERNÁNDEZ, R. M., MOLINA, M., MORENO, R. y RUÍZ, A. (1.987). *Matemáticas B.U.P. 3º.* Madrid: Centro de Publicaciones. Secretaría General Técnica Ministerio de Educación y Ciencia.

[88A] GONZÁLEZ, A., GONZÁLEZ, J. y LABORDA, M. (1.988). *Matemáticas 3º B.U.P.* Madrid: Akal.

[90A] BELMONTE, J. M., MONTERO, G., NEGRO, A., PÉREZ, S., SIERRA, T. y SORDO, J. M. (1.990). *Matemáticas 3.* Madrid: Alhambra.

Anexo III

Apuntes de clase del profesor

TEMA 6. REGRESIÓN Y CORRELACIÓN ENTRE DOS VARIABLES ESTADÍSTICAS

1. EL PROBLEMA

Al comenzar el estudio de las variables estadísticas bidimensionales ya comentamos que nuestro objetivo era conocer si habrá algún tipo de dependencia entre ambas variables unidimensionales.

En la práctica es frecuente encontrar fenómenos en los que dos variables están relacionadas. Son, por ejemplo, fenómenos físicos en los que el tipo de relación entre las dos variables es funcional (como el caso de la velocidad y el espacio en el movimiento uniforme). Esta dependencia estaría englobada en lo que llamábamos dependencia funcional.

Pero existen otros fenómenos en donde las variables presentan alguna relación pero en las que es imposible definir sobre ellas una función matemática que verifiquen exactamente. Este tipo de dependencia se le llama dependencia estadística. Es normal en este tipo de dependencia que a cada valor de X correspondan varios valores de Y.

En ese caso la variable X la llamamos variable independiente y a la variable Y dependiente, de manera que estudiaremos de qué forma X condiciona los valores de Y.

La forma en que estudiaremos esta relación es mediante la regresión. La regresión consiste en la búsqueda de una función que exprese lo mejor posible la relación existente entre dos o más variables (en nuestro caso dos variables).

La principal aplicación que puede tener el estudio de la regresión es la de predecir: conociendo el valor de una de las variables, estimar el valor que presentará la otra variable, de forma que ese valor estimado sea lo más exacto posible. O dicho de otro modo que el error cometido sea lo más pequeño posible.

Hay muchos fenómenos en donde tiene sentido hablar de regresión:

* los pesos y alturas de un grupo de personas
* los ingresos y gastos en un grupo de familias
* la tasa de consumo de energía per cápita y la renta per cápita

Así, si quiero predecir los gastos para una familia con unos ingresos dados, es lógico pensar que el error que cometa es mayor si no conozco sus ingresos, que si los conozco. Esto es porque para unos ingresos elevados se esperarán unos gastos elevados, o por lo menos más elevados que si los ingresos fueran pequeños.

Ahora bien, la posible relación de dependencia entre dos variables se pone de manifiesto al representar la nube de puntos correspondiente. Por lo general no se podrá encontrar una función que pase por todos los puntos, pero si nos puede ayudar a determinar una función que exprese dicha relación de la mejor manera. Así

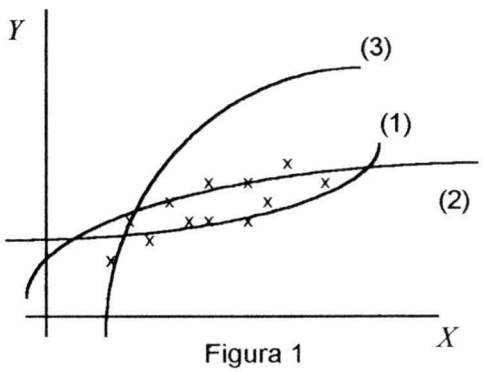


Figura 1

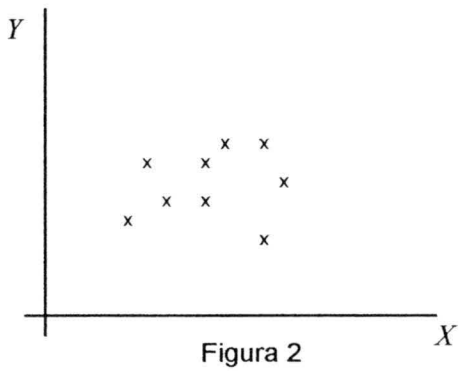


Figura 2

en la figura 1 se observa que al crecer X también crece Y con lo que se observa dependencia entre ambas variables. Además se puede ver que las funciones 1 y 2 expresan bien el fenómeno, pero no sabemos cuál de ellas lo hace mejor, y la función 3 no expresaría bien la dependencia.

En la figura 2 sin embargo no es posible determinar ningún tipo de relación entre X e Y.

Hecha la regresión, nos gustaría saber si es fiable o no. La respuesta nos la dará en gran parte el estudio de la correlación.

Empecemos viendo la regresión tipo I o curvas de regresión.

Se denomina *curva de regresión de Y sobre X* a la representación gráfica del conjunto de puntos $\{(x_i, \bar{y}_i)\}_{i=1,2,\ldots,k}$, que unimos por líneas. La notamos por Y/X.

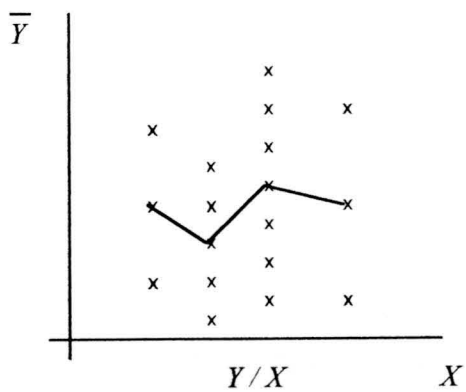


Figura 3

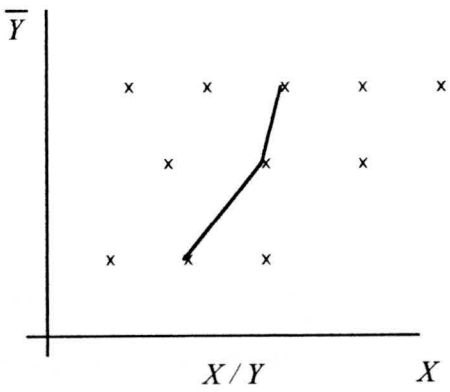


Figura 4

Y *curva de regresión de X sobre Y* a la representación gráfica del conjunto de puntos $\{(\bar{x}_j, y_j)\}_{j=1,2,\ldots,p}$ que unimos por líneas. La notamos por X/Y. En estas curvas sólo conocemos un número finito de puntos.

Veamos que ocurre para los casos en que haya independencia y cuando hay dependencia funcional.

a) Si tenemos independencia de X e Y, las distribuciones condicionadas eran iguales entre sí e igual a la marginal correspondiente. De aquí podemos afirmar que sus medias son las mismas, es decir,

$$\bar{y}_i = \bar{y}, i=1,2,...,k, \bar{x}_j = \bar{x}, j=1,2,...,p$$

La primera igualdad nos dice que la curva de regresión de Y sobre X es una recta paralela al eje de abcisas. Y la segunda igualdad que la curva de regresión de X sobre Y es una recta paralela al eje de ordenadas.

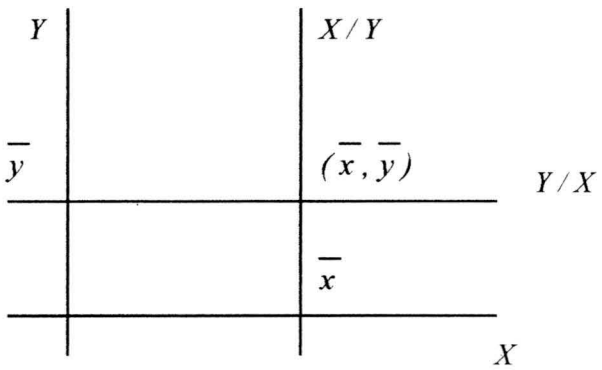


Figura 5

Como se ve cualquiera que sea el comportamiento de X no explica nada del de Y, pues la recta de regresión se mantiene constante. Igualmente Y no explica nada de X.

b) Si hay dependencia funcional de Y respecto de X (no recíproca), teníamos que a cada valor de X le correspondía un único valor de Y, en cuyo caso las distribuciones condicionadas de Y a valores de X estarán formadas por un único valor y_i. De esta forma la media coincidirá con dicho valor $\bar{y}_i = y_i$, i=1,2,...,k. La curva pasa por todos los puntos de la nube.

Es el grado máximo de relación que podemos encontrar. Los valores de X explican perfectamente el comportamiento de Y.

Si la dependencia fuera recíproca las curvas de regresión de *Y/X* y la de *X/Y* coincidirán.

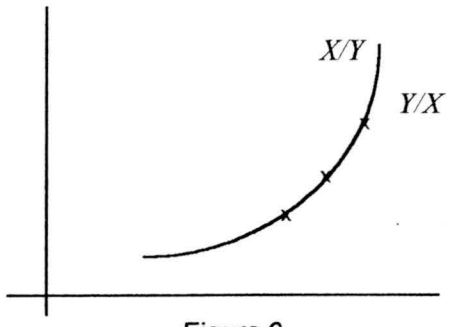


Figura 6

Pero lo usual es un caso intermedio. Entonces, ¿ sería correcto sustituir las distribuciones condicionadas por sus medias ? Esto se verá en el siguiente apartado.

2. AJUSTE MÍNIMO-CUADRÁTICO

Nos proponemos encontrar una función que se " *ajuste* " lo mejor posible a la nube de puntos. Como lo normal es que no se pueda pasar por todos los puntos tenemos que adoptar algún criterio para hallar dicha función.

Dicho criterio va a ser el de *mínimos cuadrados* . Veamos en qué consiste, aunque no el por qué de ser el más adecuado.

Dijimos anteriormente que nuestra función habría de servir para predecir los valores de Y una vez conocidos los de X (suponiendo Y variable explicada, X variable explicativa). La nube de puntos estaba formada por los puntos (x_i, y_j), para $i = 1,2,...,k$, $j = 1,2,...,p$. Así para un valor x_i de X, mediante dicha función que notaremos por $h(x)$ predeciríamos un valor $h(x_i) = y_i^*$ para Y, el cual posiblemente no coincida con el valor o valores y_j con los que x_i presenta frecuencia distinta de cero.

Por lo tanto cometeremos en dicha predicción un error que notaremos mediante $e_{ij} = y_j - y_i^* = y_j - h(x_i)$ y que se llaman *residuos*.

Entonces para hacer un buen ajuste deberíamos minimizar la suma de todos los residuos o errores.

Pero esta opción no es acertada porque los residuos negativos se pueden compensar con los positivos, dándose casos como los siguientes.

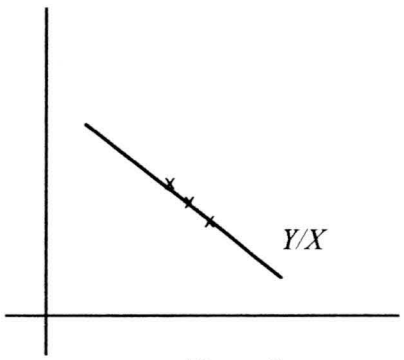


Figura 7

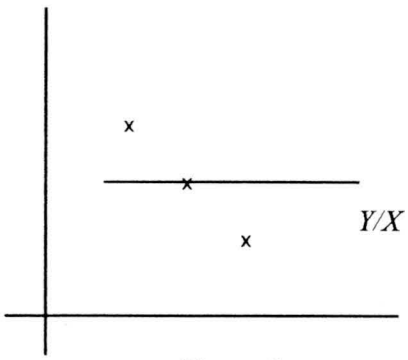


Figura 8

En la Figura 7 los residuos son nulos, luego su suma también es cero. En la Figura 8 los residuos son 1, 0 y -1, por lo tanto su suma también será nula, de forma que en segundo caso, que no es un buen ajuste, parece tan bueno como el primero. Luego interesa que se sumen residuos positivos. Una opinión sería tomar valores absolutos $| e_{ij} |$, pero el valor absoluto no se presta al cálculo algebraico.

Así pues nos queda considerar la suma de los residuos al cuadrado, lo cual tratamos de minimizar:

$$\min \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, e_{ij}^2 = \min \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_j - y_i *)^2 = \min \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_j - h(x_i))^2$$

Ya tenemos el ajuste por mínimos cuadrados. Según este criterio el mínimo se ecuentra para la curva de regresión tipo I, lo cual no vamos a probar aquí.

Este resultado es interesante, pero no es práctico pues con dicha curva disponemos de un número finito de puntos y no podemos predecir. Por tanto vamos a encontrar una función que minimice esa expresión para cualquier valor de X. Tenemos entonces *la regresión tipo II*. Para poder hallar el mínimo, primero tenemos que especificar de qué tipo de función se trata. Vamos a empezar considerando una función lineal (una recta).

Regresión lineal

La función h será de la forma $h(x) = a + bx = y$. Luego para x_i de X se tiene que $h(x_i) = a + bx_i = y_i^*$ y la función a minimizar es

$$\phi(a,b) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} (y_j - a - bx_i)^2$$

Como vemos es una función en los parámetros a y b de la recta, que son los que tenemos que hallar para que $\phi(a,b)$ sea mínima. La condición necesaria para que los valores a y b sean mínimos es que las derivadas parciales se anulen en dichos valores. Dichas derivadas parciales son

$$\left.\begin{aligned} \frac{\partial\phi(a,b)}{\partial a} &= -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_j - a - bx_i) = 0 \\ \frac{\partial\phi(a,b)}{\partial b} &= -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, x_i(y_j - a - bx_i) = 0 \end{aligned}\right\}$$

Tenemos que resolver este sistema que es equivalente a este otro

$$\left.\begin{aligned} \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, y_j &= a + b \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i \\ \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i\, y_j &= a \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i + b \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i^2 \end{aligned}\right\}$$

$$\left.\begin{aligned} m_{0,1} &= a + bm_{1,0} \\ m_{1,1} &= am_{1,0} + bm_{2,0} \end{aligned}\right\}$$

Lo resolvemos despejando a de la primera ecuación

$$a = m_{0,1} - bm_{1,0} = \bar{y} - b\bar{x} = a$$

$$m_{1,1} = (m_{0,1} - bm_{1,0})m_{1,0} + bm_{2,0} \Rightarrow m_{1,1} = m_{0,1}m_{1,0} - bm_{1,0}^2 + bm_{2,0}$$

$$b\left[m_{2,0} - m_{1,0}^2\right] = m_{1,1} - m_{0,1}m_{1,0} \Rightarrow b = \frac{m_{1,1} - m_{0,1}m_{1,0}}{m_{2,0} - m_{1,0}^2} = \frac{\sigma_{XY}}{\sigma_X^2}$$

Luego la recta de regresión de Y sobre X queda

$$y = \bar{y} - b\bar{x} + \frac{\sigma_{XY}}{\sigma_X^2}x = \bar{y} - \frac{\sigma_{XY}}{\sigma_X^2}\bar{x} + \frac{\sigma_{XY}}{\sigma_X^2}x, \; y - \bar{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{x})$$

Si quisiéramos hallar la recta de regresión X/Y, tendríamos que minimizar la función $\phi'(a', b') = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(x_i - a' - b'y_j)^2$. Se resuelve de igual forma que en el caso anterior, obteniéndose los parámetros

$$a' = \bar{x} - b'\bar{y}, \; b' = \frac{\sigma_{XY}}{\sigma_Y^2}$$

de forma que $x = \bar{x} - b'\bar{y} + \frac{\sigma_{XY}}{\sigma_Y^2} = \bar{x} - \frac{\sigma_{XY}}{\sigma_Y^2}\bar{y} + \frac{\sigma_{XY}}{\sigma_Y^2}y$, siendo la recta de regresión de X/Y

$$x - \bar{x} = \frac{\sigma_{XY}}{\sigma_Y^2}(y - \bar{y})$$

Vamos a llamar $b = \frac{\sigma_{XY}}{\sigma_Y^2}$ *coeficiente de regresión de* Y/X. Como b es la pendiente de la recta que hemos hallado, resulta que el coeficiente de regresión nos mide la tasa de incremento de Y para variaciones unitarias de X.

De igual forma, el *coeficiente de regresión de* X/Y será $b' = \frac{\sigma_{XY}}{\sigma_Y^2}$, pendiente de la recta de regresión X/Y. También nos mide la tasa de incremento de X para variaciones unitarias de Y. Los signos de b y b' sólo dependen del signo de la covarianza y, por tanto, son el mismo.

Si $\sigma_{XY} > 0$, ya dijimos que ambas variables variaban con el mismo sentido. Ahora vemos que las dos rectas de regresión son crecientes, lo cual indica tal circunstancia.

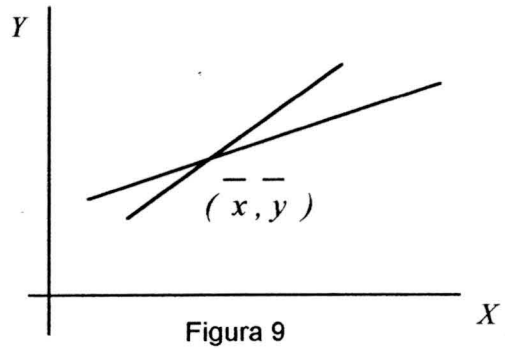


Figura 9

Si $\sigma_{XY} < 0$, las dos rectas de regresión son decrecientes, indicando que las variables varían en sentido opuesto.

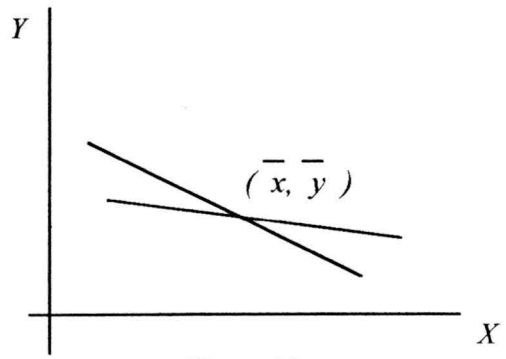


Figura 10

Si $\sigma_{XY} = 0$, las dos rectas de regresión son paralelas a los ejes, y perpendiculares entre sí.

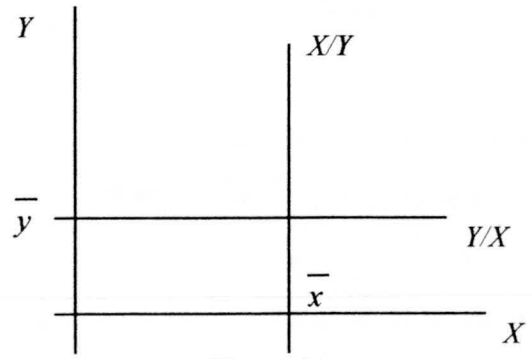


Figura 11

Observemos que si X e Y son independientes, implica que $\sigma_{XY} = 0$, y las rectas de regresión no explican nada de la variable dependiente.

Ejemplo 1 Calcular las rectas de regresión Y/X, X/Y.

x_i	y_j	x_i^2	y_j^2	$x_i\, y_j$
9	8	81	64	72
7	5	49	25	35
3	4	9	16	12
6	2	36	4	12
7	9	49	81	63
5	6	25	36	30
10	10	100	100	100
8	9	64	81	72
2	1	4	1	2
5	5	25	25	25
62	59	442	433	423

$$\bar{x} = \frac{62}{10} = 6'2,\ \bar{y} = \frac{59}{10} = 5'9,\ \sigma_X^2 = \frac{442}{10} - 6'2^2 = 5'76$$

$$\sigma_Y^2 = \frac{433}{10} - 5'9^2 = 8'49,\ \sigma_{XY} = \frac{423}{10} - 6'2 \cdot 5'9 = 5'72$$

$$Y/X:\ y - 5'9 = \frac{5'72}{5'76}(x - 6'2),\ y = 0'993\,x - 0'2566$$

$$X/Y:\ x - 6'2 = \frac{5'72}{8'49}(y - 5'9),\ x = 0'6737\,y + 2'23$$

Veamos como afecta a las rectas de regresión un cambio de origen y escala.

$$Y/X: y = a + bx \qquad X/Y: x = a' + b'y$$

Hacemos un cambio de origen : $X' = X - x_0$, $Y' = Y - y_0$. Nos queda

$$y' = a_1 + b_1 x',\ x' = a_1' + b_1' y'$$

$$\left.\begin{array}{l} b_1 = \frac{\sigma_{X'Y'}}{\sigma^2_{X'}} = \frac{\sigma_{XY}}{\sigma^2_X} = b \\ b'_1 = \frac{\sigma_{X'Y'}}{\sigma^2_{Y'}} = \frac{\sigma_{XY}}{\sigma^2_Y} = b' \end{array}\right\} \Rightarrow b \text{ y } b' \text{ son invariantes ante un cambio de origen}$$

$$a_1 = \bar{y}' - b_1\bar{x}' = y - y_0 - b(x - x_0) = y - y_0 - bx + bx_0 = a - y_0 + bx_0$$

$$a_1 = a' - x_0 + b' y_0$$

Si hacemos un cambio de escala : $X' = \frac{X}{u}$, $Y' = \frac{Y}{v}$

$$b = \frac{\sigma_{XY}}{\sigma^2_X} = \frac{uv\,\sigma_{X'Y'}}{u^2\,\sigma^2_{X'}} = \frac{v}{u}\frac{\sigma_{X'Y'}}{\sigma^2_{X'}} = \frac{v}{u}b' \Rightarrow b_1 = \frac{u}{v}b$$

Igualmente $b_1' = \frac{v}{u}b'$.

$$a_1 = \bar{y}' - \frac{\sigma_{X'Y'}}{\sigma^2_{X'}}\bar{x}' = \frac{\bar{y}}{v} - \frac{u}{v}\,b\,\frac{\bar{x}}{u} = \frac{1}{v}(\bar{y} - b\bar{x}) = \frac{1}{v}\,a$$

Igualmente $a_1' = \frac{1}{u}\,a'$. Por lo tanto con un cambio de origen y escala:

$$X' = \frac{X - x_0}{u}\,,\; Y' = \frac{Y - y_0}{v}$$

$$\begin{cases} a_1 = \frac{1}{v}\,(a - y_0 + \frac{u}{v}\,bx_0) \\ b_1 = \frac{u}{v}\,b \end{cases} \qquad \begin{cases} a'_1 = \frac{1}{u}\,(a' - x_0 + \frac{v}{u}\,b'\,y_0) \\ b'_1 = \frac{v}{u}\,b' \end{cases}$$

3. INTRODUCCIÓN A LA REGRESIÓN NO LINEAL

Hemos estudiado el ajuste de mínimos cuadrados para una función lineal.

Sin embargo, no siempre esta función es la que mejor se ajusta a la nube de puntos. Habrá problemas en los que la función que mejor expresa la relación de dependencia no sea lineal.

Aquí vamos a ver sólo algunos casos, los más simples, que corresponden con funciones que vamos a linealizar.

A. Regresión hiperbólica

Supongamos la siguiente nube de puntos

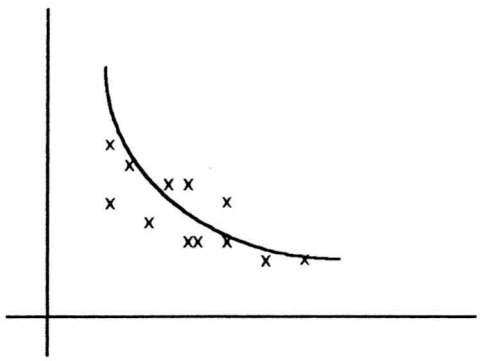


Figura 12

la función que mejor la aproxima es una hipérbola $y = a + \frac{b}{x}$. Efectuando el cambio $z = 1/x$ nos queda una recta $y = a + bz$, por lo que para encontrar a y b utilizamos la recta de regresión para las variables (y,z).

$$\left.\begin{array}{l} b = \frac{\sigma_{ZY}}{\sigma_Z^2} \\ a = \bar{y} - b\bar{z} \end{array}\right\} \Rightarrow Y/Z : y - \bar{y} = \frac{\sigma_{ZY}}{\sigma_Z^2}(z - \bar{z})$$

Posteriormente deshacemos el cambio $z = 1/x$. Esta curva puede ser apropiada cuando se expresa la demanda de un bien en función de la renta.

Ejemplo 2

x_i	y_j	$1/x_i = z_i$	z_i^2	$z_i y_j$
1/2	15	2	4	30
1/4	19	4	16	76
1/6	25	6	36	150
1/7	33	7	49	231
1/8	34	8	64	272
	126	27	169	759

$$\bar{z} = \frac{27}{5} = 5'4, \bar{y} = \frac{126}{5} = 25'2$$

$$\sigma_Z^2 = \frac{169}{5} - 5'4^2 = 4'64, \sigma_{ZY} = \frac{759}{5} - 5'4 \cdot 25'2 = 15'72$$

$$y - 25'2 = \frac{15'72}{4'64}(z - 5'4)$$

$$y = 3'38\, z + 6'9051 \Rightarrow y = \frac{3'38}{x} + 6'9051$$

B. Regresión potencial

Pretendemos ajustar una curva del tipo $y = ax^b$. Para hallar a y b linealizamos dicha función:

$$\ln y = \ln a + b \ln x$$

Si llamamos $\ln Y = v$, $\ln X = u$, $\ln a = A$: $v = A + bu$. Luego hallaremos A y b, calculándose la recta de regresión de V / U, y después se deshace el cambio $\ln a = A$.

Ejemplo 3

x_i	y_j	$u_i = \ln x_i$	$v_j = \ln y_j$	u_i^2	$u_i\, v_j$
1	1	0	0	0	0
2	4	0'6931	1'3862	0'4803	0'9607
3	8	1'0986	2'0794	1'2069	2'2844
4	9	1'3862	2'1972	1'9215	3'0457
		3'178	5'6628	3'6088	6'2909

$$\bar{u} = \frac{3'178}{4} = 0'7945, \bar{v} = \frac{5'6628}{4} = 1'4157$$

$$\sigma_U^2 = \frac{3'6088}{4} - 0'7945^2 = 0'2709, \sigma_{UV}^2 = \frac{6'2909}{4} - 0'7945 \cdot 1'4157 = 0'4479$$

$$v - 1'4157 = \frac{0'4479}{0'2709}\ (u - 0'7945),\ v - 1'4157 = 1'6535\ (u - 0'7945)$$

$$v = 1'6535\ u + 1'1019,\ a = e^{1'1019} = 1'1073 \Rightarrow y = 1'1073x^{1'6535}$$

C. Regresión exponencial

Se ajusta la curva $y = ab^x$ linealizándola: $\ln y = \ln a + x \ln b$. Llamando $\ln y = v$, $\ln a = A$, $\ln b = B$, $v = A + Bx$. Se calcula A y B en la recta de regresión V/X y después se deshacen los cambios: $a = e^A$, $b = e^B$.

Ejemplo 4

x_i	y_j	$\ln y_j = v_j$	x_i^2	$x_i v_j$
1	1	0	1	0
2	4	1'3862	4	2'7728
3	8	2'0794	9	6'2382
4	9	2'1972	16	8'7888
		5'6628	30	17'7998

$$\bar{x} = \frac{10}{4} = 2'5, \bar{v} = \frac{5'6628}{4} = 1'4157$$

$$\sigma_X^2 = \frac{30}{4} - 2'5^2 = 1'25, \sigma_{VX}^2 = \frac{17'7998}{4} - 2'5 \cdot 1'4157 = 0'9107$$

$$v - 1'4157 = \frac{0'9107}{1'25}\ (x - 2'5),\ v = 0'7285x - 0'4057$$

$$a = e^{-0'4057} = 0'66651,\ b = e^{0'7285} = 2'071 \Rightarrow y = 0'66651 \cdot 2'071^x$$

Hay otras funciones que se ajustan como son la curva parabólica, la curva logística, etc. que no vamos a tratar, pero que también se utilizan, sobre todo la curva logística, en problemas socio-económicos (hojas 6.16, 6.18).

4. MEDIDAS DE DEPENDENCIA ESTADÍSTICA

Una vez que hemos determinado la forma en que se relacionan las variables, se plantea el problema de medir el *grado de asociación* de las mismas, o lo que es lo mismo el grado de dependencia mutua. Además podremos comprobar si el ajuste realizado es bueno o no. Esto es lo que se llama *correlación*: el estudio de la bondad del ajuste de una curva.

Lo que realmente hace la correlación es eso, precisar si el ajuste es bueno o no, es decir, si la relación de dependencia es la expresada por la regresión. Pero esto no quiere decir que un mal ajuste implique que no haya asociación entre las variables, sino que éste no es del tipo expresado por la curva ajustada.

Ahora bien, podría hallarse la correlación para la regresión tipo I, que era la mejor, y si nos da que hay poca dependencia, entonces ninguna curva va a ajustarse bien.

De todas formas vamos a empezar con la correlación en el caso lineal.

Coeficiente de determinación. Varianza residual.

Ya comentamos en su momento, que la recta de regresión no pasa por los puntos de la nube, sino que se producían unas diferencias entre los valores observados y_j y los teóricos y_i*, que llamábamos residuos: $e_{ij} = y_j - y_i$*.

$$\bar{e} = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} e_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} (y_j - y_i *) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, y_j -$$

$$- \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, y_i * = \sum_{j=1}^{p} f_{\cdot j}\, y_j - \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, (a + b x_i) = \bar{y} - a \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} -$$

$$- b \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, x_i = \bar{y} - (a + b\bar{x}) = \bar{y} - \bar{y} = 0$$

Es decir, en un ajuste lineal la media de los residuos es cero. Pero veamos su varianza:

$$\sigma_e^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(e_{ij} - \bar{e})^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, e_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_j - y_i\,*\,)$$

A esta varianza la vamos a llamar <u>varianza residual</u>, y se usa como medida de la bondad del ajuste. Cuanto mayor sea la varianza residual peor es el ajuste; pero como tiene medida (la de los datos del problema) no se puede comparar con otras varianzas residuales, ni tampoco podemos determinar a partir de qué valor la varianza residual es lo suficientemente grande como para admitir un mal ajuste. Para resolver estos problemas se utiliza el coeficiente de determinación que aparece al descomponer la varianza de la variable.

$$\sigma_Y^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_j - \bar{y})^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{p} [(y_j - a - bx_i) + (a + bx_i - \bar{y})]^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, e_{ij}^2 + \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_i\,*\,-\bar{y})^2 + 2\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, e_{ij}(y_i\,*\,-\bar{y})$$

Veamos que A vale cero:

$$A = \sum_{i=1}^{k} (y_i\,*\,-\bar{y}) \sum_{j=1}^{p} f_{ij}\, e_{ij} = \sum_{i=1}^{k} (a + bx_i - \bar{y}) \sum_{j=1}^{p} f_{ij}\, e_{ij} =$$

$$= \sum_{i=1}^{k} (\bar{y} - b\bar{x} + bx_i - \bar{y}) \sum_{j=1}^{p} f_{ij}\, e_{ij} = b \sum_{i=1}^{k} (x_i - \bar{x}) \sum_{j=1}^{p} f_j^i f_{i\cdot}\, e_{ij} =$$

$$= b \sum_{i=1}^{k} f_{i\cdot}(x_i - \bar{x}) \sum_{j=1}^{p} f_j^i\, e_{ij} = 0$$

Luego $\sigma_Y^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, e_{ij}^2 + \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}(y_i\,*\,-\bar{y})^2 = \sigma_e^2 + \sigma_{Y*}^2$.

$\sigma^2_{Y^*}$ mide la variación originada por la relación lineal entre X e Y, es decir, en qué medida queda explicada la variable Y (dependiente) mediante la recta de regresión. Esta descomposición, $\sigma^2_Y = \sigma^2_e + \sigma^2_{Y^*}$, es válida para funciones lineales en los parámetros, es decir, para la recta y la hipérbola (y para la parábola que no la hemos estudiado). Por lo tanto, no es válida para las demás curvas - potencial, exponencial, etc. -. Ya vemos aquí que si σ^2_e es pequeña, entonces la varianza explicada por la regresión es mayor y el ajuste es mejor. Pero como seguimos teniendo el problema de la unidad, vamos a definir el *coeficiente de determinación* como la proporción de la varianza total de Y que aparece explicada por la regresión:

$$R^2 = \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = 1 - \frac{\sigma^2_e}{\sigma^2_Y}$$

No tiene unidad.

Campo de variación e interpretación de R^2

Como $\sigma^2_{Y^*} \leq \sigma^2_Y$, y al ser ambos positivos, se tiene que $0 \leq R^2 \leq 1$. Ahora veamos los posibles casos.

- Si $R^2 = 1$, implica que $\sigma^2_e = 0$, es decir, que todos los residuos son nulos, o lo que es igual, la recta pasa por todos los puntos. El ajuste es perfecto y tendríamos una dependencia funcional.

- Si $R^2 = 0$, implica que $\sigma^2_e = \sigma^2_Y$ y $\sigma^2_{Y^*} = 0$, o sea, que el modelo no explica nada de Y a partir de X. X e Y no están asociados por dicha regresión.

- Si $0 < R^2 < 1$, la regresión será mejor cuanto más cercano esté R^2 de 1 (se suele tomar el valor 0'75 como límite para considerar que el ajuste es aceptable).

En definitiva, el coeficiente de determinación mide el grado de ajuste de la recta de regresión.

Cálculo de R^2

$$R^2 = \frac{\sigma^2_{Y^*}}{\sigma^2_Y},\ \sigma^2_{Y^*} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(a + bx_i - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(\bar{y} - b\bar{x} + bx_i - \bar{y})^2 =$$

$$= b^2 \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(x_i - \bar{x})^2 = \frac{\sigma^2_{XY}}{\sigma^4_X}\,\sigma_X{}^2 = \frac{\sigma^2_{XY}}{\sigma^2_X}$$

Luego $R^2 = \dfrac{\sigma^2_{XY}}{\sigma^2_X\,\sigma^2_Y}$.

Definimos el <u>*coeficiente de correlación lineal*</u> $r = \dfrac{\sigma_{XY}}{\sigma_X\,\sigma_Y}$, donde $-1 \le r \le 1$, $(r^2 = R^2)$.

El signo de r depende del signo de la covarianza. Observemos que dicho coeficiente se puede obtener a partir de los coeficientes de regresión lineal de las recta X/Y, Y/X.

$$\left.\begin{aligned} b &= \frac{\sigma_{XY}}{\sigma^2_X} \\ b' &= \frac{\sigma_{XY}}{\sigma^2_Y} \end{aligned}\right\} \Rightarrow bb' = r^2 = R^2$$

Lo cual, también, nos indica que dicho coeficiente es el mismo para la recta de regresión X/Y.

El coeficiente de correlación lineal sirve para medir el grado de asociación lineal entre dos variables, pero no de una forma cuantitativa sino cualitativa.

Vamos a interpretar los valores que toma r :

• Si $r = \pm 1$, entonces hay un <u>*ajuste perfecto*</u>, puesto que $R^2 = 1$ lo que implica que $\sigma^2_e = 0$. La dependencia entre las variables está explicada perfectamente por las rectas de regresión.

Si r = -1 las variables varían en sentido opuesto y si $r = 1$ varían en el mismo sentido.

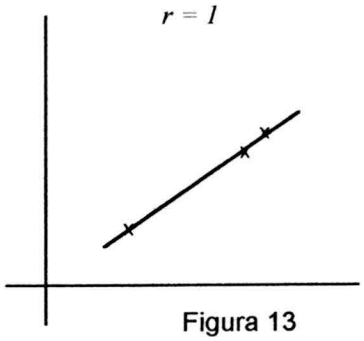


Figura 13

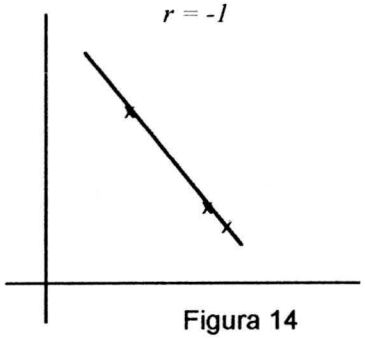


Figura 14

• Si $r = 0$, implica que $\sigma_{XY} = 0$, y las rectas de regresión son $y = \bar{y}, x = \bar{x}$, paralelas a los ejes de coordenadas. Entonces, no hay asociación lineal entre ambas variables, pero esto no quiere decir que no haya otro tipo de dependencia entre ellas.

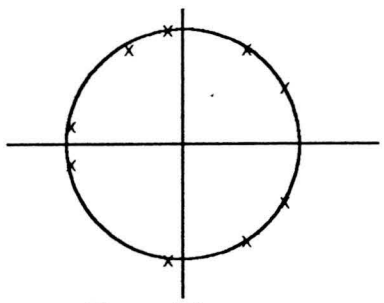


Figura 15

$x^2 + y^2 = 1$

X e Y son dependientes.

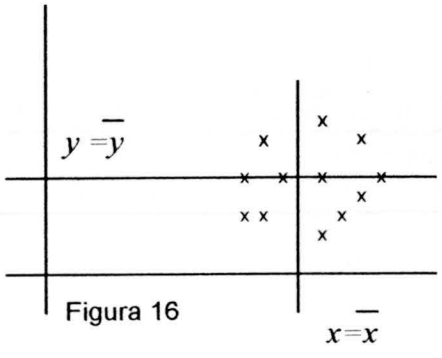


Figura 16

• Si $0 < r < 1$, la regresión lineal será tanto mejor cuanto más próximo esté r de 1 y las variables varían en el mismo sentido.

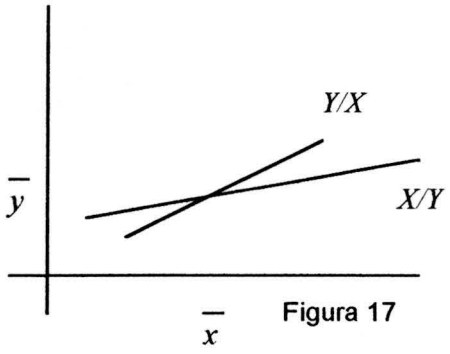


Figura 17

• Si $-1 < r < 0$, la regresión lineal será tanto mejor cuanto más próximo esté r de -1. Las variables varían en sentido opuesto.

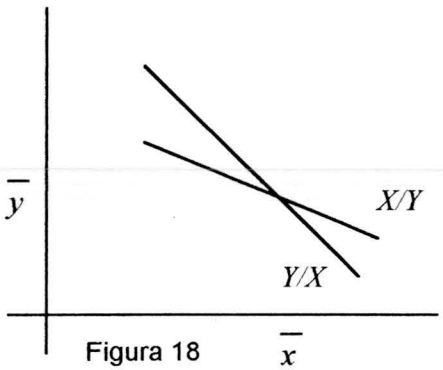


Figura 18

Por último, veamos cómo le afecta un cambio de origen y escala a r. Sean

$$X' = \frac{X - X_0}{c},\ Y' = \frac{Y - Y_0}{d} \Rightarrow \sigma_{XY} = cd\sigma_{X'Y'}, \sigma_X = c\sigma_{X'}, \sigma Y = d\sigma_{Y'}$$

Luego $r = \dfrac{\sigma_{XY}}{\sigma_X\ \sigma_Y} = \dfrac{cd\sigma_{X'Y'}}{c\sigma_{X'}\ d\sigma_{Y'}} = \dfrac{\sigma_{X'Y'}}{\sigma_{X'}\ \sigma_{Y'}} = r'$, lo que implica $r = r'$.

El cambio de origen y escala no le afecta.

Bondad del ajuste para otras funciones

Al hacer el ajuste mediante alguna de las funciones que hemos linealizado (potencial y exponencial), obteníamos una recta de regresión para uno de los transformados.

Si calculamos el coeficiente de determinación R^2 en estas rectas transformadas no tendrá el mismo significado y no será comparable con valores obtenidos sobre la bondad del ajuste de una recta o de una parábola (o de una hipérbola).

Por tanto, si queremos comparar el ajuste entre este tipo de funciones y la recta o parábola, es más adecuado calcular

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2}$$

El problema es que $\sigma_e^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij}\, e_{ij}^2$ no es una varianza y además será $\sigma_Y^2 \neq \sigma_e^2 + \sigma_{Y*}^2$, lo que implica que R^2 puede ser negativo, pues σ_e^2 puede ser mayor que σ_Y^2 .

El ajuste será mejor cuanto más se aproxime R^2 a 1.

Ejemplo 5

Para calcular su coeficiente de correlación lineal sólo basta multiplicar sus coeficientes de regresión lineal que conocíamos ya.

$$bb' = r^2 \text{ , } 0'993 \cdot 0'6737 = 0'6689 = r^2,\ r = 0'8179$$

Esto nos dice que el ajuste no es del todo malo, se podría aceptar incluso. (hoja 6.17.)

Ejemplo 6 Ajustar por una exponencial

x_i	y_j	$v_j = \ln y_j$	x_i^2	$x_i \cdot v_j$	y_j^2	y_i^*	$y_j - y_i^*$	$(y_j - y_i^*)^2$
2	6	1'79	4	3'58	36	3'1202	2'8798	8'2932
4	5	1'61	16	6'44	25	2'9601	2'0399	4'1611
3	1	0	9	0	1	3'0391	-2'0391	4'1579
1	3	1'099	1	1'088	9	3'2034	-0'2039	0'0413
10	15	4'5	30	11'108	71		2'6767	16'6535

$$y = ab^x,\ \ln y = \ln a + x \ln b,\ v = A + Bx$$

$$\bar{x} = \frac{10}{4} = 2'5,\ \bar{v} = \frac{4'5}{4} = 1'125,\ \sigma_X^2 = \frac{30}{4} - 2'5^2 = 1'25$$

$$\sigma_{VX} = \frac{11'119}{4} - 2'5 \cdot 1'125 = -0'03275,\ B = \frac{\sigma_{VX}}{\sigma_X^2} = \frac{-0'03275}{1'25} = -0'0262$$

$$b = e^B = 0'974,\ A = \bar{v} - B\bar{x} = 1'125 - (-0'0262) \cdot 2'5 = 1'1905$$

$$a = e^A = 3'289,$$

luego

$$y = 3'289 \cdot (0'974)^x$$

Estudiamos la bondad del ajuste: $R^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2}$

$$\bar{y} = \frac{15}{4} = 3'75,\ \sigma_Y^2 = \frac{71}{4} - 3'75^2 = 3'6875,\ \sigma_e^2 = \frac{16'6535}{4} = 4'1633$$

$$R^2 = 1 - \frac{4'1633}{3'6875} = -0'12903$$

Se ve que el ajuste no es bueno (la "varianza residual", σ_e^2 , es mayor que la varianza de Y).

Razón de correlación de Pearson

Al principio dijimos que la función que mejor se ajusta a los datos es la curva de regresión de tipo I; por lo tanto su bondad de ajuste nos da el umbral que no podrá ser superado por ningún tipo de función que ajustemos, y también una medida del grado de (relación) asociación.

El coeficiente de determinación para la curva de regresión de tipo I se llama *razón de correlación de Pearson*.

$$\eta^2_{Y/X} = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(y_j - \bar{y}_i\right)^2}{\sigma_Y^2} = 1 - \frac{\sum_{i=1}^{k} f_{i\cdot}\sum_{j=1}^{p} f_j^i\left(y_j - \bar{y}_i\right)^2}{\sigma_Y^2} =$$

$$= 1 - \frac{\sum_{i=1}^{k} f_{i\cdot} v_i(Y)}{\sigma_Y^2} = 1 - \frac{\overline{v_i(Y)}}{\sigma_Y^2} = \frac{v(\bar{y}_i)}{\sigma_Y^2} = \frac{\sum_{i=1}^{k} f_{i\cdot}(\bar{y}_i - \bar{y})^2}{\sigma_Y^2}$$

Luego $\eta^2_{Y/X} = 1 - \frac{\overline{v_i(Y)}}{\sigma_Y^2} = \frac{v(\bar{y}_i)}{\sigma_Y^2}$. Análogamente

$$\eta^2_{X/Y} = 1 - \frac{\overline{v_j(X)}}{\sigma_X^2} = \frac{v\left(\bar{x}_j\right)}{\sigma_X^2}$$

5. APLICACIONES ECONÓMICAS

Nos vamos a centrar, solamente, en la *predicción*. La predicción, dijimos, que consistía en determinar valores de la variable dependiente en función de valores de la variable independiente.

Para efectuar dichas predicciones basta con utilizar la ecuación de la función ajustada: se sustituye un valor de X en la ecuación y obtenemos el correspondiente valor de Y. La predicción será tanto más fiable cuanto menor sea la varianza residual, es decir, cuanto más próximo a 1 esté el coeficiente de determinación.

Además, la fiabilidad de los valores pronosticados disminuye a medida que los valores de X se alejan de su media, o mejor dicho, cuando se alejan del rango de variación de X.

Mirar los ejercicios 2, 3, 4, 5, 6, 9, 10, 11, 12 y 14 del libro *"Problemas de Estadística"* de Muñoz, A., Lozano, E., Rodríguez, J. y Ruíz, J.C.

Regresión parabólica

Queremos ajustar una curva del tipo $y = a + bx + cx^2$. Esta función no se puede linealizar, por tanto, hay que recurrir al criterio de mínimos cuadrados.

Habrá que minimizar

$$\phi(a,b,c) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left[y_j - a - bx_i - cx_i^2 \right]^2$$

La condición necesaria para encontrar un mínimo en a, b y c es que sus derivadas parciales se anulen.

$$\frac{\partial \phi}{\partial a} = -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left[y_j - a - bx_i - cx_i^2 \right] = 0$$

$$\frac{\partial \phi}{\partial b} = -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left[y_j - a - bx_i - cx_i^2 \right] x_i = 0$$

$$\frac{\partial \phi}{\partial c} = -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \left[y_j - a - bx_i - cx_i^2 \right] x_i^2 = 0$$

Se obtiene el sistema

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, y_j = a + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i + c\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i\, y_j = a\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^2 + c\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^3$$

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^2\, y_j = a\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^2 + b\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^3 + c\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\, x_i^4$$

Ejemplo 7 Ajustar una parábola a:

$X \setminus Y$	1	2	3	4	5
1	3	2			
2			4	4	
4					3

x_i	y_j	u_{ij}	$x_i\, u_{ij}$	$y_j\, u_{ij}$	$u_{ij}\, x_i^2$	$u_{ij}\, x_i^3$	$u_{ij}\, x_i^4$	$u_{ij}\, x_i\, y_j$	$u_{ij}\, x_i^2\, y_j$
1	1	3	3	3	3	3	3	3	3
1	2	2	2	4	2	2	2	4	4
2	3	4	8	12	16	3	64	24	48
2	4	4	8	16	16	32	64	32	64
4	5	3	12	15	48	192	768	60	240
		16	33	50	85	261	901	123	359

$$\frac{50}{16} = 16\, a + b\cdot\frac{33}{16} + c\cdot\frac{85}{16}$$

$$\frac{123}{16} = a\cdot\frac{33}{16} + b\cdot\frac{85}{16} + c\cdot\frac{261}{16}$$

$$\frac{359}{16} = a\cdot\frac{85}{16} + b\cdot\frac{261}{16} + c\cdot\frac{901}{16}$$

Entonces a = -1'6, b = 3'45, c = -0'45.

Ejemplo 8 Hallemos la varianza residual:

$$\sigma_e^2 = \frac{1}{u} \sum_{i=1}^{k} \sum_{j=1}^{p} u_{ij}(y_j - y_i *)^2$$

$$y = 6'9051 + \frac{3'39}{x}$$

x_i	y_i*	$y_j - y_i$*	$(y_j - y_i*)^2$
1/2	13'58	1'32	1'7424
1/4	20'46	-1'46	2'1316
1/6	27'24	-2'24	5'0176
1/7	30'63	2'32	5'3824
1/8	34'02	-0'02	0'0004
			14'2744

$$\sigma_e^2 = \frac{14'2744}{5} = 2'8549,\ R^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{2'8569}{56'16} = 0'9491$$

$$(\ R^2 = \frac{15'72^2}{4'64 \cdot 56'16} = 0'9483\ )$$ El ajuste es bueno.

Ejemplo 9 Hallemos su varianza residual: $y = 1'1073\ x^{1'6535}$.

x_i	y_i*	$y_j - y_i$*	$(y_j - y_i*)^2$
1	1'1073	-0'1073	0'0115
2	3'4835	0'5164	0'2667
3	6'8306	1'1894	1'4146
4	10'959	-1'959	3'8377
			5'5306

$$\sigma_e^2 = \frac{5'5306}{4} = 1'3826,\ R^2 = 1 - \frac{1'3826}{10'25} = 0'8651$$

$$\Sigma y_j^2 = 1 + 16 + 64 + 81 = 162,\ \sigma_Y^2 = \frac{162}{4} - 5'25 = 10'25$$

$$\Sigma y_j = 1 + 4 + 8 + 9 = 22,\ \bar{y} = \frac{22}{4} = 5'5$$

El ajuste es aceptable.

Ejemplo 10 $y = 0'66651 \cdot 2'071^x$

x_i	y_i^*	$y_j - y_i^*$	$(y_j - y_i^*)^2$
1	1'3803	-0'3803	0'1446
2	2'8586	1'1413	1'3026
3	5'9203	2'0796	4'3249
4	12'26103	-3'26103	10'6343
			16'4065

$$\sigma_e^2 = \frac{16'4065}{4} = 4'1016$$

Esta varianza residual es mayor que la obtenida con el ajuste potencial, por lo tanto, este es peor ajuste que el potencial.

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{4'1016}{10'25} = 0'5998$$

ya se ve que el ajuste no es del todo bueno.

Anexo IV

Apuntes de las alumnas

TEMA 2. VARIABLES ESTADÍSTICAS UNIDIMENSIONALES.

2.1. INTRODUCCIÓN. DEFINICIONES.

En el tema 1 hemos visto que una variable estadística (X) representa un carácter.

Ahora queremos hacer un estudio conjunto de dos variables x e y. Para ésto se unirán en una (x,y). Ej: Peso y altura de una persona. Esta será una variable estadística bidimensional.

Supongamos que la variable x toma las modalidades $x_1, x_2, \ldots, x_p$ y la variable y, $y_1, y_2, \ldots, y_q$. y los valores de la variable bidimensional (x,y) serán x_i, y_j,

$i = 1, 2, \ldots, p$

$j = 1, 2, \ldots, q$. y

estos serán los valores de dicha variable.

- Frecuencia absoluta (n_{ij}) del par (x_i, y_j): es el nº de individuos que han presentado el valor de x_i de x y y_j de y

- Frecuencia relativa la representamos con una f_{ij} y es $\frac{n_{ij}}{N}$, donde N es el nº total de todos los individuos, es decir, una suma doble $N = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$

- Llamados distribución bidimensional de frecuencia al conjunto de valores $\{(x_i, y_j), n_{ij}) ; i = 1, \ldots, p ; j = 1, \ldots q\}$ Formados por pares de valores y frecuencias absolutas.

2.2 REPRESENTACIONES NUMÉRICAS.

Cuando las variables x e y son discretas o continuas o una discreta y la otra continua. La forma usual de representar su distribución bidimensional de frecuencias es mediante una "tabla de doble entrada".

- Mixta: cuando hay una variable continua y otra discreta. Ej: Sea la distribución dada por la siguiente tabla.

X	Y	n_{ij}
20	50-100	10
20	100-150	3
20	150-200	2
21	50-100	5
21	100-150	15
21	150-200	5
22	50-100	2
22	100-150	20
22	150-200	15
23	100-150	13
23	150-200	10

$N = 100$

X \ Y	50 - 100	100 - 150	150 - 200	$n_{i\cdot}$
	$y_1 = 75$	$y_2 = 125$	$y_3 = 175$	
$x_1 = 20$	$n_{11} = 10$	$n_{21} = 3$	$n_{31} = 2$	$\rightarrow 15 = n_{1\cdot}$
$x_2 = 21$	$n_{21} = 5$	$n_{12} = 15$	$n_{32} = 5$	$\rightarrow 25 = n_{2\cdot}$
$x_3 = 22$	$n_{31} = 2$	$n_{23} = 20$	$n_{33} = 15$	$\rightarrow 37 = n_{3\cdot}$
$x_4 = 23$		$n_{24} = 13$	$n_{34} = 10$	$\rightarrow 23 \quad n_{4\cdot}$
$n_{\cdot j}$	$n_{\cdot 1} = 17$	$n_{\cdot 2} = 51$	$n_{\cdot 3} = 32$	$N = 100$

ESTADÍSTICA

En la última columna aparecen las cantidades $n_{i\cdot}$, que representan el nº de veces que se ha observado el valor x_i de la variable X, sin tener en cuenta el valor presentado por Y.

$n_{i\cdot}$ se obtiene sumando todas las frecuencias de la fila correspondiente al valor x_i de X.

$$n_{i\cdot} = \sum_{j=1}^{q} n_{ij} \quad (i\cdot fijo) \quad \text{(última columna)}$$

$n_{\cdot j}$ se obtiene sumando todas las frecuencias de la fila correspondiente al valor y_j de Y.

$$n_{\cdot j} = \sum_{i=1}^{p} n_{ij} \quad (j\ fijo) \cdot \quad \text{(última fila)}.$$

$$N = \sum_{i=1}^{p} n_{i\cdot} = \sum_{j=1}^{q} n_{\cdot j}$$

2.3. DISTRIBUCIONES MARGINALES Y CONDICIONADAS.

A partir de las frecuencias absolutas anteriores, podemos asignar las <u>frecuencias relativas</u>.

Con la primera y la última de doble entrada que hemos visto en el ejemplo anterior, obtenemos la distribución unidimensional de X donde $f_{i\cdot} = \frac{n_{i\cdot}}{N}$, lo mismo lo podemos hacer con la variable Y.

X	$n_{i\cdot}$	$f_{i\cdot}$
20	15	0'15
21	25	0'25
22	37	0'37
23	23	0'23
	N=100	1

$L_{i-1} - L_i$	Y	$n_{\cdot j}$	$f_{\cdot j}$
50-100	75	17	0'17
100-150	125	51	0'51
150-200	175	32	0'32
		N=100	1

La primera fila y la última nos da la distribución unidimensional de Y. Estas distribuciones se llaman

marginales de X e Y. Este nombre deriva del hecho de que dichas frecuencias se obtiene tomando filas o columnas primeras y últimas

Por ser distri. unidi., a las dist. marginales podemos aplicar todo lo que hemos visto en el T.1.

En el cambio de notación, la notación de los momentos se acompañará de un subíndice u otro elemento diferenciador que nos indique a qué variable se refiere

Ej:

Media marginal de X:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{p} x_i n_{i\cdot} = \sum_{i=1}^{p} x_i f_{i\cdot}$$

Media marginal de Y:

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{q} y_j n_{\cdot j} = \sum_{j=1}^{q} y_j f_{\cdot j}$$

Varianza marginal de X:

$$\sigma_X^2 = \frac{1}{N} \sum_{i=1}^{p} (x_i - \bar{x})^2 n_{i\cdot}$$

$$\sigma^2 x = \sum_{i=1}^{p} (x_i - \bar{x})^2 f_{i\cdot}$$

$$\sigma^2 x = \left(\frac{1}{N} \sum_{i=1}^{p} x_i^2 n_{i\cdot} \right) - \bar{x}^2$$

Varianza marginal de Y

$$\sigma^2 y = \frac{1}{N} \sum_{j=1}^{q} (y_i - \bar{x})^2 n_{\cdot j}$$

$$\sigma^2 y = \sum_{j=1}^{q} (y_i - \bar{y})^2 f_{\cdot j}$$

$$\sigma^2 y = \left(\frac{1}{N} \sum_{j=1}^{q} y^2 n_{\cdot j} \right) - \bar{y}^2$$

2.4. DISTRIBUCIONES CONDICIONADAS.

Las distribuciones condicionadas son distribuciones unidimensionales obtenidas a partir de las bidimensionales manteniendo fijo el valor de una de las variables (que consideramos) los valores de las otras con sus frecuencias

ESTADÍSTICA

la dist. condicionada de X dado que $Y = y_j$ se obtiene a partir de la tabla de doble entrada tomando la 1ª columna de los valores de la variable X y la j^{esima} de la columna de las frecuencias absolutas.

la dist. cond. de Y, dado que $X = x_i$ se obtiene a partir de la tabla de doble entrada tomando la 1ª fila de los valores de la variable Y y la i^{esima} fila de las frecuencias absolutas.

Ejemplos.

*la distribución condicionada para Y, dado que $X = x_2 = 21$.

$y_j / x_2 = 21$	n_{ij}
50 - 100	$n_{21} = 5$
100 - 150	15
150 - 200	5
	$n_{2\cdot} = 25$

las frecuencias relativas para estas distribuciones condicionadas se definen como el cociente entre la frecuencia absoluta y el nº total de observaciones $\begin{matrix} n_{\cdot j} \\ n_{i\cdot} \end{matrix}$ ó

Así las dist. condicionadas Y/x_i notamos $f_{j/i} = \frac{n_{ij}}{n_{i\cdot}}$

Esta frecuencia representa la proporción de individuos de los que cumplen la condición $x = x_i$ y que han presentado el valor $y = y_j$.

De la misma manera para las dist. cond. X/y_j notamos, $f_{i/j} = \frac{n_{ij}}{n_{\cdot j}}$. Tiene un significado análogo al anterior.

$$\sum_{i=1}^{p} f_{i/j} = \sum_{i=1}^{p} \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{\cdot j}}{n_{\cdot j}} = \frac{1}{n_{\cdot j}} \underbrace{\sum_{i=1}^{p} n_{ij}}_{n_{\cdot j}} = \frac{n_{\cdot j}}{n_{\cdot j}} = 1$$

$$\sum_{j=1}^{q} f_{j/i} = \sum_{j=1}^{q} \frac{n_{ij}}{n_{i\cdot}} = \frac{1}{n_{i\cdot}} \underbrace{\sum_{j=1}^{q} n_{ij}}_{n_{i\cdot}} = \frac{n_{i\cdot}}{n_{i\cdot}} = 1.$$

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ij}/N}{n_{\cdot j}/N} = \frac{p_{ij}}{f_{\cdot j}}$$

$$f_{j/i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{ij}/N}{n_{i\cdot}/N} = \frac{f_{ij}}{f_{i\cdot}} \Rightarrow f_{j/i} = \frac{f_{ij}}{f_{c\cdot}}$$

Ejemplo:

Distrib. condicionada de X dados que $Y = y_2 = 125$.

x_i/y_2	$n_{i/2}$	$f_{i/2}$
20	3	0'059
21	15	0'294
22	20	0'392
23	13	0'255
$n_{\cdot 2} =$	51	1

$$f_{1/2} = \frac{n_{12}}{n_{\cdot 2}} = 0'059$$

$$f_{2/2} = \frac{n_{22}}{n_{\cdot 2}} = 0294$$

$$f_{3/2} = \frac{n_{32}}{n_{\cdot 2}} = \frac{13}{51} = 0'255$$

Las distribuciones condicionadas son distribuciones unidimensionales a las cuales se les puede aplicar todo lo conocido para este tipo de distribuciones. A las características calculadas para estas distr. se les añade el calificativo de condicionadas.

- Para las distrib. condicionadas Y/x_i notaremos $f_{j/i}$

$$\boxed{f_{j/i} = \frac{n_{ij}}{n_{i\cdot}}}$$

Representa la proporción de individuos de los que cumplen la condición $x = x_i$ y que han presentado el valor $y_j \neq y_{j\cdot}$

- Para las distrib. condicionadas X/y_j notaremos $f_{i/j}$

$$\boxed{f_{i/j} = \frac{n_{ij}}{n_{\cdot j}}}$$

Tiene un significado análogo al anterior.

- Medias condicionadas para distribuciones de X/y_j:

$$\bar{X}_j = m_{X/y_j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} x_i n_{ij} = \sum_{i=1}^{p} x_i f_{i/j}$$

- Medias condicionadas para distribuciones de Y/x_i:

$$\bar{Y}_i = m_{Y/x_i} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{q} y_j n_{ij} = \sum_{j=1}^{q} y_j f_{j/i}$$

ESTADISTICA

- Varianza condicionada de X/y_j.

$$\sigma_j^2(x) = \sigma^2_{x/y_j} = \frac{1}{n \cdot j} \sum_{i=1}^{p} (x_i - \bar{x}_j)^2 n_{ij}$$

$$\sigma_j^2(x) = \sum_{i=1}^{p} (x_i - \bar{x}_j)^2 f_{i/j}$$

- Varianza condicionada de y/x_i

$$\sigma_i^2(y) = \sigma^2 y/x_i = \frac{1}{n_{i\cdot}} \sum_{j=1}^{q} (y_j - \bar{y}_i)^2 n_{ij}$$

$$\sigma_i^2(y) = \sum_{j=1}^{q} (y_j - \bar{y}_i)^2 f_{j/i}$$

2.5. RELACIÓN ENTRE DISTRIBUCIONES MARGINALES Y CONDICIONADAS

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}} \qquad f_{j/i} = \frac{f_{ij}}{f_{i\cdot}}$$

$$\Rightarrow \boxed{f_{ij} = f_{i/j} \cdot f_{\cdot j}} \qquad \Rightarrow \boxed{f_{ij} = f_{j/i} \cdot f_{i\cdot}}$$

La relación existente entre estas distribuciones puede resumirse en la siguiente relación:

$$\boxed{f_{ij} = f_{i/j} \cdot f_{\cdot j} = f_{j/i} \cdot f_{i\cdot}}$$

De estas relaciones se obtienen las siguientes:

$$\boxed{\bar{x} = \sum_{j=1}^{q} \bar{x}_j f_{\cdot j}} \qquad \boxed{\bar{y} = \sum_{i=1}^{p} \bar{y}_i f_{i\cdot}}$$

$$\boxed{\sigma^2_x = \sum_{j=1}^{q} \sigma_j^2(x) f_{\cdot j} + \sum_{j=1}^{q} (\bar{x}_j - \bar{x})^2 f_{\cdot j}}$$

$$\boxed{\sigma^2_y = \sum_{i=1}^{p} \sigma_i^2(y) f_{i\cdot} + \sum_{i=1}^{p} (\bar{y}_i - \bar{y})^2 f_{i\cdot}}$$

2.6. INDEPENDENCIA ESTADISTICA.

Dos variables X e Y son estadísticamente independientes cuando la variación de una de ellas no influye en la otra, es decir, si la distribución condicionada de Y a $X = x_i$ es indepen-

diente del valor de x_i.

La condición necesaria y suficiente para que dos variables estadísticas sean independientes es la frecuencia relativa conjunta sea igual al producto de los marginales.

① $\boxed{f_{ij} = f_{i\cdot} \cdot f_{\cdot j}}$ $\iff$ X e Y independientes.

Otras condiciones para la independencia:

② X e Y independientes estadísticamente $\iff f_{j/i} = f_{\cdot j}$

$$f_{j/i} = \frac{f_{ij}}{f_{i\cdot}} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{f_{i\cdot}} = f_{\cdot j}$$

③ X e Y independientes estadísticamente $\iff f_{i/j} = f_{i\cdot}$

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{f_{\cdot j}} = f_{i\cdot}$$

EJEMPLO

En la siguiente tabla de doble entrada comprobar si X e Y son independientes.

X \ Y	y_1	y_2	y_3	y_4	$n_{i\cdot}$
x_1	3	5	2	4	14
x_2	6	10	4	8	28
x_3	12	20	8	16	56
$n_{\cdot j}$	21	35	14	28	

Tabla de distribuciones condicionadas a $y = y_j$ $\quad j = 1, 2, 3, 4$.

X	$y = y_1$	$y = y_2$	$y = y_3$	$y = y_4$	$f_{i\cdot}$
x_1	3/21	5/35	2/14	4/28	14/98
x_2	6/21	10/35	4/14	8/28	28/98
x_3	12/21	20/35	8/14	16/28	56/98

ESTADISTICA

$f_{i/j} = f_{i.}$

$\frac{3}{21} = \frac{5}{35} = \frac{2}{14} = \frac{4}{28} = \frac{14}{98} = \boxed{\frac{1}{7}}$ → se cumple.

$\frac{6}{21} = \frac{10}{35} = \frac{4}{14} = \frac{8}{28} = \frac{28}{98} = \boxed{\frac{2}{7}}$ → se cumple.

$\frac{12}{21} = \frac{20}{35} = \frac{8}{14} = \frac{16}{28} = \frac{56}{98} = \boxed{\frac{4}{7}}$ → se cumple.

$\frac{n_{ij}}{n_{.j}} = f_{i/j} = f_{i.}$

X e Y son independientes.

2.7. DEPENDENCIA.

Dada la variable estadística bidimensional (X,Y) diremos que Y depende funcional/ de X si $Y = f(x)$.

Es decir, a cada valor o modalidad de X le corresponde un único valor o modalidad de Y.

Por lo que en cada fila o columna de la tabla hay un solo valor distinto de 0.

Lo normal es que conocidos los valores de una variable sepamos parcialmente qué ocurre con los valores de la otra, surgiendo así el concepto de dependencia estadística.

Dependencia estadística → conocemos los valores parcial/.

Dependencia funcional → conocemos todos los valores de la otra variable mediante una función

2.8. REPRESENTACIONES GRÁFICAS.

A) La representación gráfica más utilizada es la nube de puntos o diagrama por dispersión, que consiste en representar en unos ejes cartesianos los puntos → $\rightarrow \{(x_i, y_j), i = 1, \dots p; j = 1, \dots, q\}$ colocando un punto con coordenadas iguales a cada par (x_i, y_j).

Si ambas variables son discretas se representa un punto para cada par de valores observados; cuando una pareja se ha observado más de una vez junto al punto, se coloca el

valor de la frecuencia o bien se transforman dichos puntos en círculos de tamaños proporcionales a la frecuencia.

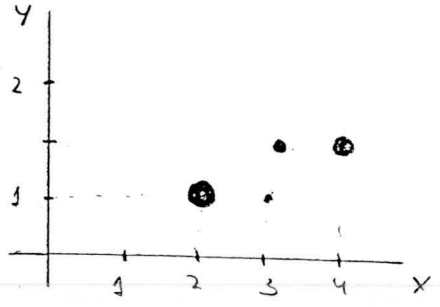
Si ambas variables son continuas buscamos las marcas de clase como representantes de cada intervalo y así estaríamos en la misma situación de las variables discretas.

Otra manera para las continuas consiste en representar en cada rectángulo determinado por los intervalos de (x,y) tantos puntos como indique la frecuencia observada.

EJEMPLO.

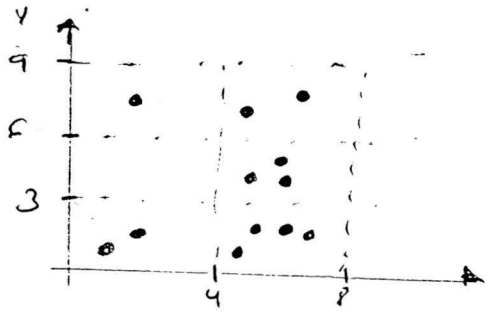
(Caso discreto).

X \ Y	1	2
2	5	0
3	1	2
4	0	3



(Caso continuo)

X \ Y	0-3	3-6	6-9
0-4	2	0	1
4-8	4	3	2

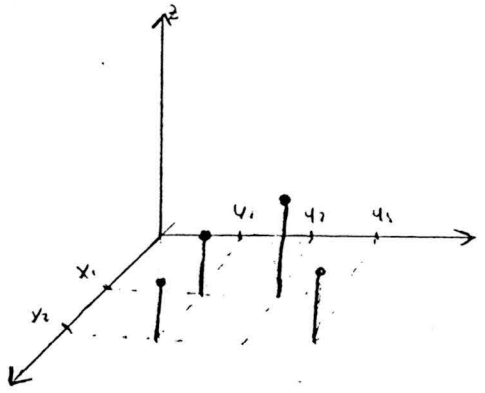


B) Otro tipo de representaciones gráficas es el <u>estereograma o escalograma</u>.

Para variables discretas consiste en representar sobre el plano (x,y) los valores (x_i, y_i) y levantar sobre cada uno de los puntos, un segmento o barra de longitud igual a la frecuencia con que se ha observado dicho par de valores.

ESTADISTICA

X \ Y	y_1	y_2	y_3	
x_1	5	4	0	
x_2	15	0	2	



Para variables continuas se marcan los intervalos sobre X e Y dando lugar a rectangulos sobre el plano (X,Y) y sobre cada uno se construye un paralelepipedo de volumen igual o proporcional a la frecuencia que se quiere representar.

Si $L_i - L_{i-1}$ → del $i^{\text{ésimo}}$ intervalo de X.

Si $L_j - L_{j-1}$ → del $j^{\text{ésimo}}$ intervalo de Y.

h_{ij} es la altura del paralelepípedo construido sobre la clase jésima de X y la clase jésima de Y.

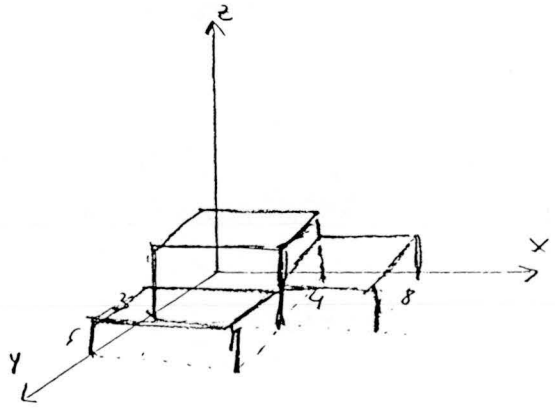
V_{ij} es el volumen del paralelepipedo construido sobre.

Entonces $\boxed{V_{ij} = n_{ij} = (L_i - L_{i-1})(L_j - L_{j-1})\, h_{ij}} \Rightarrow$

$$\boxed{h_{ij} = \frac{n_{ij}}{(L_i - L_{i-1})(L_j - L_{j-1})}}$$

EJEMPLO.

X \ Y	0-4	4-8
0-3	5	3
3-6	4	0



2.9. MOMENTOS DE UNA DISTRIBUCIÓN BIDIMENSIONAL.

El objetivo es análogo al tema 1, construir unos parámetros asociados a la distrib, pero para la distrib bidimensional que resuma diversos aspectos de la misma.

Definición de un momento no centrado (o respecto al origen) de orden r y s de una variable bidimensional:

V. continuas.

$$a_{rs} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_i^r \, y_j^s \, f_{ij}$$

$$a_{rs} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{q} x_i^r \, y_j^s \, n_{ij}$$

V discretas.

$$a_{rs} = \frac{1}{N} \sum_{i=1}^{p} x_i^r \, y_i^s$$

Casos particulares.

$$a_{10} = \sum_{i=1}^{p} \sum_{j=1}^{q} \underbrace{x_i^1}_{x_i} \underbrace{y_i^0}_{1} f_{ij} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_i \, f_{ij} =$$

$$= \sum_{i=1}^{p} x_i \underbrace{\sum_{j=1}^{q} f_{ij}}_{f_{i\cdot}} = \sum_{i=1}^{p} x_i \, f_{i\cdot} = \bar{x}.$$

$$a_{01} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_i^0 \, y_j^1 \, f_{ij} = \bar{y}.$$

Obtenemos también:

$$\left.\begin{array}{l} a_{20} = a_2(x) \\ a_{02} = a_2(y) \end{array}\right\} \Rightarrow$$

Los momentos de 2º orden respecto al origen de las distrib. marginales reciben dados por a_{20} y a_{02} respectivam.

En general, cualquier momento bidimensional en que uno de los subíndices sea 0 se corresponde con un momento unidimensional de la distrib. marginal de X, Y.

El momento no centrado más utilizado es:

V. continuas.

$$a_{11} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{q} x_i \, y_j \, n_{ij}$$

ESTADISTICA

V. discretas.

$$a_{ij} = \frac{1}{n} \sum_{i=1}^{n} x_i y_j$$

Se define el momento centrado con respecto a las medias de orden r y s como:

- Variables continuas: $m_{rs} = \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})^r \rightarrow$

$(y_j - \bar{y})^s \int ij.$

$$m_{rs} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})(y_j - \bar{y})^s n_{ij}.$$

- Variables discretas

$$m_{rs} = \frac{1}{n} \sum_{i=1}^{p} (x_i - \bar{x})^r (y_i - \bar{y})^s.$$

Casos particulares.

$m_{10} = 0$ $\quad m_{01} = 0$

$m_{20} = \sigma_x^2$ varianza marginal de x

$m_{02} = \sigma_y^2$ " " de y.

~~El momento más importante respecto a las x más importante es~~ la covarianza.: es el momento respecto de las medias más importante.

- Para V. Continuas:

$$m_{11} = \sigma_{xy} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})(y_j - \bar{y}) n_{ij}$$

- Para V Discretas

$$m_{11} = \sigma_{xy} = \frac{1}{n} \sum_{i=1}^{p} (x_i - \bar{x})(y_i - \bar{y}).$$

La covarianza es una medida de la variabilidad conjunta de x e y, por tanto de la relación entre las variables x e y. Si la covarianza es positiva, las dos variables varían en el mismo sentido, es decir, si una aumenta la otra también. Si la covarianza es negativa, las dos variables varían en sentido opuesto, es decir, si una aumenta la otra disminuye.

Relación entre momentos centrados (a_{rs}) y no centrados (μ_{rs})

Los d. centrados se pueden expresar en relación de los no centrados.

Para los momentos más utilizados se tienen las siguientes relaciones:

$$\begin{cases} m_{20} = \sigma_x^2 = a_{20} - a_{10}^2 \\ m_{02} = \sigma_y^2 = a_{02} - a_{01}^2 \end{cases} \Rightarrow \text{válidas para distrib. unidimen.}$$

La covarianza: $\boxed{\sigma_{xy} = a_{11} - a_{01}a_{10}}$

• Propiedad

Si x e y son independientes, entonces la covarianza de x e y es cero.

Nota.- La covarianza puede variar sobre todo $\mathbb{R}$, es decir, que no es acotada (desde ∞ hasta $-\infty$). Entonces no podemos saber si el grado de asociación entre x e y es más o menos fuerte.

Por eso se define el <u>coeficiente de correlación lineal</u>, que se define como:

$$r = \frac{m_{11}}{\sqrt{m_{20}m_{02}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

A partir de esta definición vemos que r está dentro del intervalo $(-1 \le r \le 1)$ $(-1,1)$ y que es la covarianza sobre las <u>variables tipificadas</u>, es decir estas variables son:

$$x' = \frac{x - \bar{x}}{\sigma_x} \qquad y' = \frac{y - \bar{y}}{\sigma_y}$$

$$\sigma_{x'y'} = r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Esto nos puede indicar el grado de variación conjunta que tienen las variables x' e y' y por tanto las variables x^* e y^*

• <u>Cambio de origen y escala: en los momentos centrados</u>.

$$\text{Sea } x' = \frac{x-a}{h} \quad y' = \frac{y-b}{e} \Rightarrow \begin{cases} x = hx' + a \\ y = ey' + b \end{cases}$$

ESTADÍSTICA

A partir de ahí, decimos:

$$\begin{cases} \bar{x} = h\,\bar{x}' + a \\ \bar{y} = e\,\bar{y}' + b \end{cases}$$

Por tanto $m_{rs}(x,y) = \sum_{i=1}^{p}\sum_{j=1}^{q}(x_i - \bar{x})^r (y_j - \bar{y})^s f_{ij}$.

obtenemos $\boxed{m_{rs}(x,y) = h^r e^s m_{rs}(x',y')}$.

Casos particulares

$$\boxed{m_{11}(x,y) = he\, m_{11}(x',y')} \Rightarrow \sigma_{xy} = he\sigma_{x'y'}$$

$$\left.\begin{array}{l} m_{20}(x,y) = h^2 m_{20}(x',y') \\ m_{02}(x,y) = e^2 m_{02}(x',y') \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \sigma_x^2 = h^2\sigma_{x'}^2 \\ \sigma_y^2 = e^2\sigma_{y'}^2 \end{array}\right.$$

$$\Rightarrow r_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{he\sigma_{x'y'}}{h\sigma_{x'}e\sigma_{y'}} = \frac{\sigma_{x'y'}}{\sigma_{x'}\sigma_{y'}}$$

$\boxed{r_{xy} = r_{x'y'}}$ ⇒ No le afecta el cambio de origen y escala. Por tanto es un coeficiente adimensional.

2.10. REGRESIÓN Y CORRELACIÓN SIMPLE.

INTRODUCCIÓN.

Existen variables como el consumo y la ~~oferta~~ renta, el consumo y la demanda, etc, entre las que existe una relación, pero es imposible definir sobre ellas una función matemática que refleje esa relación exacta.

Este tipo de dependencia entre variables se llama dependencia estadística, frente a la dependencia funcional, en la que sí hay una función matemática que los satisface de forma exacta.

La regresión pretende poner unas variables en función de otras mediante una ley. Las aplicaciones más interesantes de la regresión son:

a) Crear mejor un fenómeno.

b) Predecir o estimar el valor de una variable conocido el valor de otras, relacionadas con ellas.

Hecha la predicción inmediata surge la duda sobre su fiabilidad. La respuesta a esta duda estará en

gran parte dada por el estudio de la correlación. La variable que se quiere predecir se llama dependiente o a veces endógena. Las variables cuyo conocimiento se usa para la predicción se llaman independientes o exógenas.

Cuando solo/ se usa 1 variable independiente (exógena) estaremos ante la regresión y correlación simple).

Si interviene más de 1 variable independiente la correlación o regresión se llama múltiple.

AJUSTE POR MINIMOS CUADRADOS

Si entre dos variables no existe una dep. funcional es imposible encontrar una función entre ellas cuya representación gráfica pase por todos los puntos del diagrama de dispersión.

El conjunto de puntos se llama nube

Queremos obtener una curva que pase por la mayoría de los puntos.

De ahí buscamos la curva que aunque no pase por todos los puntos de la nube, al menos esté lo más próxima posible a ellos, es decir, se ajuste mejor.

Uno de los criterios más adecuados para encontrar dicha curva es el llamado por ajuste por mínimos cuadrados.

En una variable estadística (x,y) si no hay una ~~relación~~ dependencia funcional entre (x,y), no se puede afirmar que: a cada valor de una de las variables le corresponde de forma unívoca una de la otra., sin embargo es fácil pensar que el comportamiento por ejemplo de la variable condicionada $y/x = x_i$, para cada (x,y) difiera del comportamiento de y.

Son bien conocidas las cualidades de la media aritmética como valor representativo del comportamiento de una variable.

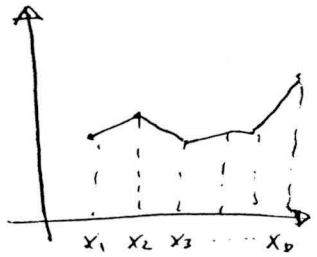
ESTADISTICA

LINEA DE REGRESIÓN

Llamamos línea o curva de regresión de y sobre x a la representación gráfica de $\{(x_i, \bar{y}_i);\ i = 1, \dots, p\}$, que uniéndolas con una línea

$\bar{y}_i$ = media de $y/x = x_i$

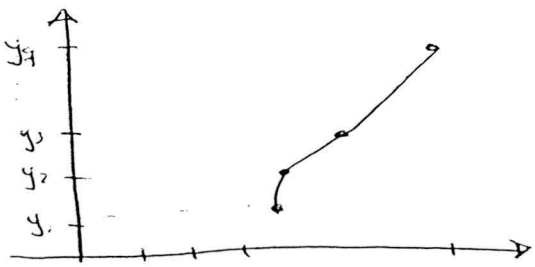
Esta línea nos indica las distribuciones de la variable y condicionada por los valores x_i de la variable x.



De la misma forma se def. la curva de regresión x sobre y como la representación gráfica de:

$\{(\bar{x}_j, y_j);\ j = 1, \dots, q\}$

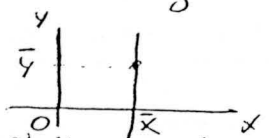
$\bar{x}_j$ ≡ media de $x/y = y_j$



¡Nota.

Aquí trabajamos con un conjunto de puntos pero si x o y fuesen continuas se tendría una verdadera curva de regresión

- Veamos la forma que adopta la curva de regresión según el grado de dep. entre x e y



– Caso de indp. funcional.

Si x e y son ind. las distrib. condicionadas distribución.
serán iguales entre sí e iguales a la ~~distrib~~ marginal correspondiente. Entonces la media condicionada es igual

a la media marginal. $\bar{X}_j = \bar{X}$, $j = 1 \dots, q$. Por lo que este conjunto $\{(\bar{x}_j, y_j); j = 1 \dots q\}$ tiene la misma abscisa.

Por tanto la curva de regresión de x sobre y es paralela al eje de OY pasando por el par $(\bar{x}, \bar{y})$.

Análogamente para estas distrib. condicionadas tenemos todas q.: $\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = \dots = \bar{y}_0 = \bar{y}$.

El conjunto $\{(\bar{x}_i, \bar{y}_i); c = 1 \dots p\}$ tiene en la misma ordenada, por tanto la curva de regresión de y sobre x es paralela al eje OX pasando por $(\bar{x}, \bar{y})$.

- Caso de dep. funcional

Y depende funcionalmente de X si a cada valor x_i de X, le corresponde un único valor y_j de y. Entonces, esta variable condicionada tiene un solo valor, por lo que, la media de esta variable condicionada $\bar{y}_i$ es y_j.

Por lo tanto el conjunto que queremos representar es este $\{(x_i, \bar{y}_i)\} = \{(x_i, y_j)\}$. es decir los valores de x explican perfecta/ los valores de y, o de otra manera $Y = f(x)$.

Análog/ se dice que la variable x dep. funcional/ de Y si a cada valor y_j de y corresponde un único valor posible de x, entonces este conjunto que queremos representar va a ser $\{(\bar{x}_j, y_j)\} = \{(x_i, y_j)\}$.

$$\Rightarrow \boxed{x = g(x)}$$

- Caso general o intermedio

La ind. q dep. funcional son casos extremos, que se encuentran raramente en la práctica

Supongamos que para predecir la variable Y conocemos otra variable relacionada con ella X de alguna forma $y = f(x)$ (no conocemos la relación que hay entre ellas)

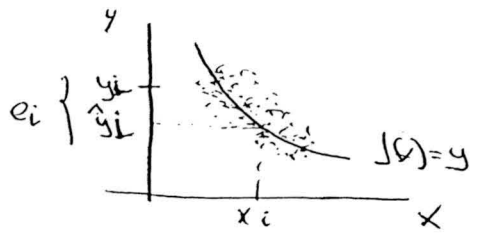
Tendríamos que considerar $x_1, x_2 \dots x_p$ modalidades de X e $y_1, y_2 \dots, y_q$ modalidades de Y con un total de observaciones $n = N$, pero vamos a considerar todos los datos repetidos o no de la forma: $x_1, \dots, x_n$

$y_1, \dots, y_n$

ESTADISTICA

Si $X = x_i$ utilizando la mencionada función, estimamos un valor $\hat{y}_j = f(x_i)$.

$\hat{y}_j$ es un estimador o estimación del valor real y_j



Supongamos en este estudio $i = j = n$.

$\hat{y}_i = f(x_i) \forall y_i$

$\hat{y}_i$ estimador del valor real de y_i.

Posiblemente $\hat{y}_i$ e y_i no sean iguales, se comete un error en la estimación $e_i = y_i - \hat{y}_i$. Este error se llama también residuo.

Un criterio para un buen ajuste es el de minimizar la suma de todos los residuos o errores al cuadrado. $\left(\sum_{i=1}^{n} e_i^2\right)$

Este es el criterio de mínimos cuadrados:

Esto es igual $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$.

El criterio de mínimos cuadrados considera que la función q. más se ajusta a unos datos observados es aquella que minimiza la suma de los cuadrados de los residuos y se comprueba que coincide con la (Esta) curva de regresión.

Las funciones f que se ajustan con más frecuencia son:

- Recta: $y = f(x) = a + bx$
- Parábola: $y = a + bx + cx^2$
- Hipérbola equilátera: $y = a + b\frac{1}{x}$
- Función potencial: $y = ax^b$
- Función exponencial: $y = as^x$

La recta de regresión.

Si la función que ajustamos es una recta la función se llama lineal. De entre todas las rectas buscamos aquella que mejor se ajuste según el criterio de mín. cuadrados, para $X = x_i$; $\hat{y}_i = a + bx_i$, ~~entonces el error~~ y $e_i = y_i - \hat{y}_i = y_i - a - bx_i$.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

El criterio de mínimos cuadrados en este caso de la recta,

trata de encontrar los coeficientes a y b, que hagan mínima la expresión $\sum_{i=1}^{n} (y_i - a - bx)^2 \} = S(a,b)$.

Condición necesaria para la existencia de mínimo en un punto es que las derivadas parciales de 1er orden se anulen en dicho punto $\left.\begin{array}{l} S(a,b) \\ \frac{\partial S(a,b)}{\partial a} = 0 \quad \frac{\partial S(a,b)}{\partial b} = 0 \end{array}\right\} \Rightarrow$

$$\Rightarrow \left\{\begin{array}{ll} 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0 & y_i = k \\ & x_i = k \\ 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0 & \end{array}\right. \quad \boxed{d - a = -1} \quad \boxed{d - bx_i = -x_i}$$

$$\left\{\begin{array}{l} -\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ -\sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{array}\right. \Rightarrow$$

Estas ecuaciones se llama, ecuaciones normales de la recta.

$$\left\{\begin{array}{l} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (a + bx_i) \\ \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} (a + bx_i) x_i \end{array}\right. \Rightarrow \left\{\begin{array}{l} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{array}\right.$$

Dividimos ambas ecuaciones sobre n, obtenemos:

$$\left\{\begin{array}{l} \frac{1}{n}\sum_{i=1}^{n} y_i = a + b\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ \frac{1}{n}\sum_{i=1}^{n} y_i x_i = a\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + b\left(\frac{1}{n}\sum_{i=1}^{n} x_i a\right) \end{array}\right.$$

$$\left\{\begin{array}{l} \bar{y} = a + b\bar{x} \\ a_{11} = a\bar{x} + ba_{20} \end{array}\right. \Rightarrow \left\{\begin{array}{l} a_{01} = a + b\,a_{10} \\ a_{11} = a\,a_{10} + b\,a_{20} \end{array}\right.$$

Este sistema o ecuaciones son válidas para variables discretas y continuas el anterior sistema puede resolverse como sigue. Multiplicamos la 1ª ecuación por (a_{10}) y la sumamos a la segunda ecuación.

ESTADISTICA

$$\Rightarrow \begin{cases} -a_{10}a_{01} = -aa_{10} - ba_{10}^2 \\ + \quad a_{11} = aa_{10} + ba_{10} \end{cases} \Rightarrow \boxed{b = \frac{a_{11} - a_{10}a_{01}}{a_{20} - a_{10}^2}} \Rightarrow$$

$$a_{11} - a_{10}a_{01} = b(a_{20} - a_{10}^2)$$

$$\boxed{b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{m_{11}}{m_{20}}}$$

Sustituimos el valor de b en la 1ª ec. y obtenemos a.

$$a_{01} = a + \frac{m_{11}}{m_{20}} a_{10}$$

$$\boxed{a = a_{01} - \frac{m_{11}}{m_{20}} a_{10}} \Rightarrow \boxed{a = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}}$$

Por lo tanto la recta de regresión de y sobre x queda de esta forma $y = a + bx = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x} + \frac{\sigma_{xy}}{\sigma_x^2}x$.

$$\Leftrightarrow y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

Análogo la recta de regresión de X sobre y sería

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

Recta de regresión de y sobre x:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Recta de regresión de X sobre y.

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Entonces ambas rectas pasan por los puntos $(\bar{x}, \bar{y})$

Esas rectas serán iguales si $r\frac{\sigma_y}{x} = \frac{1}{r\frac{\sigma_x}{\sigma_y}} =$

$$= r\frac{\sigma_y}{\sigma_x} \quad \boxed{\Rightarrow r^2 = 1}$$

y si $r^2 \neq 1$ solo tienen en común un punto = $(\bar{x}, \bar{y})$.

Los ajustes de una hipérbola equilátera se reduce fácilmente al ajuste de una recta.

- Hipérbola: $y = a + \frac{b}{x}$, tomamos $z = \frac{1}{x}$

queda: $y = a + bz$

- F. potencial: $y = ax^b$, tomamos logaritmos,

obtenemos $\ln y = \ln a + b \ln x$ ... $\boxed{y' = a' + bx'}$

- F. exponencial: $y = ab^x$, tomamos logaritmos,

$\ln y = \ln a + x \ln b$ $\boxed{y' = a' + b'x}$

Correlación se ocupa del grado de asociación entre las variables. Este grado de asociación nos indicará en qué medida la expresión encontrada explica una variable en función de otra.

• Varianza residual: el método de mínimos cuadrados toma como medida del error que se comete cuando ajustamos una función la suma de los residuos al cuadrado. (*)

Llamaremos V. residual a la cantidad

$\frac{1}{n} \sum_{i=1}^{n} e_i^2 = \sigma_{ry}^2$ y se usa para medir el grado de la bondad del ajuste.

$$(*) \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Coeficiente de determinación.

¿A partir de qué valores la varianza residual es suficiente/pequeña o grande para medir la bondad del ajuste?

Para responder a esta pregunta definimos el coeficiente de determinación: $R^2 = \frac{\sigma_y^2 - \sigma_{ry}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$.

Por su definición vemos la siguiente igualdad $0 \leq R^2 \leq 1$.

entonces $\frac{\sigma_{ry}^2}{\sigma_y^2} = 1$ ó $\sigma_{ry}^2 = \sigma_y^2$

por tanto el modelo no explica nada de y a partir de x, y decimos que en este caso el ajuste es el peor.

Si $R^2 = 1$. entonces $\frac{\sigma_{ry}^2}{\sigma_y^2} = 0 \Rightarrow \sigma_{ry}^2 = 0$, es decir, todos los residuos serán nulos. ($\sigma_{ry}^2 = \frac{1}{n} \sum e_i^2 \Rightarrow e_i = 0 \ \forall i$ porque todos los residuos son positivos). El ajuste es perfecto.

Para valores intermedios entre 0 y 1, según otros más próximos

ESTADISTICA

a 0 o a 1 nos indicarán un peor o mejor ajuste respectivamente.

Bondad del ajuste de la recta.

Para el ajuste $y = a + bx$ o si estamos ajustando $x = a' + b'y$ el coeficiente de determinación R^2 coincide con el coeficiente de correlación lineal: $R^2 = r^2$.

Además tiene una interpretación particular: $\sigma_Y^2 = \frac{1}{n}$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \boxed{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2} + \frac{2}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

Utilizando $(A+B)^2 = A^2 + B^2 + 2AB$.

Varianza explicada por la regresión.

$$\sigma_{\hat{y}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

Indica en qué medida queda explicada la variable dependiente mediante mediante el modelo estimado

$$\frac{2}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = 0.$$

En el caso de la regresión lineal tenemos.

$$\sum_{i=1}^{n} e_i(\hat{y}_i-\bar{y}) = \sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i.$$

pero $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i-\hat{y}_i) = \sum_{i=1}^{n}(y_i-a-bx_i) = \sum_{i=1}^{n} y_i - an - b\sum_{i=1}^{n} x_i = 0$

es la 1ª ec. normal de la recta.

También $\sum_{i=1}^{n} e_i\hat{y}_i = \sum_{i=1}^{n} e_i(a+bx_i) = a\underbrace{\sum_{i=1}^{n} e_i}_{0} + b\underbrace{\sum_{i=1}^{n} e_i x_i}_{} =$

$= b\left(\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2\right)$ → Es la 2ª ecuación de la recta.

Por lo tanto.

$$\sum_{i=1}^{n} e_i(\hat{y}_i-\bar{y}) = \sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i = 0.$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sigma_{r_y}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

$$\Rightarrow \boxed{\sigma_y^2 = \sigma_{r_y}^2 + \sigma_{\hat{y}}^2}$$ → explicada por la regresión

Coeficiente de determinación (bis).

$$R^2 = \frac{\sigma_y^2 - \sigma_{ry}^2}{\sigma_y^2} = \frac{\sigma_{ey}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$$

$\boxed{y = a + bx}$ → regresión de y sobre x.

El coeficiente de determinación se interpreta como la proporción de la varianza de y que viene explicada por la regresión lineal.

En el caso de la recta de regresión tenemos: $R^2 = r^2$. lo que nos da una mejor interpretación de r. Para ello vemos que: $\sigma_y^2 = \sigma_{ry} + \sigma_{ey}^2 \Rightarrow \sigma_{ry}^2 = \sigma_y^2 - \sigma_{ey}^2 = \sigma_y^2\left(1 - \frac{\sigma_{ey}^2}{\sigma_y^2}\right) = \sigma_y^2(1-r^2)$.

porque $R^2 = r^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$. $\Rightarrow R^2 = 1 - \frac{\sigma_y^2(1-r^2)}{\sigma_y^2} = r^2$.

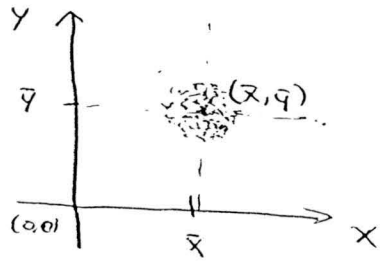
Análogo si se hallara la de esta recta ($x = a' + b'y$ → regresión de x sobre y) obtenemos.

$\sigma_{ex}^2 = \sigma_x^2(1-r^2)$ y también para esta recta puede fácil/ comprobarse que: $R'^2 = r^2$. ▨

Interpretación de r.

De lo anterior es inmediato que el coeficiente de determinación al cuadrado está entre 0 y 1: $0 \leq r^2 \leq 1$ por lo tanto $-1 \leq r \leq 1$ y la interpretación de r tiene relación con la bondad del ajuste de la recta de mínimos cuadrados o recta de regresión.

*) Si $r = 0$ entonces $\sigma_{xy} = 0$, es decir que las 2 rectas de regresión se reducen a: $y = \bar{y}$, $x = \bar{x}$, por tanto ambas son paralelas a los ejes de coordenadas.



$\sigma_{xy} = 0 \Rightarrow$ x e y son independientes.

Si $r = 0$ entonces no existe relación lineal entre x e y, sin embargo x e y pueden estar estrechamente ligadas según

otro tipo de funciones.

*) Si $r=1$, entonces $\sigma_{ry}^2 = 0$.

$\sigma_{ry}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = 0 \Rightarrow e_i = 0 \ \forall i$, esto implica que no hay residuos o los errores son nulos.

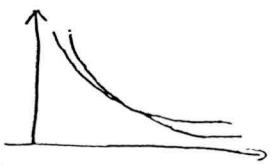
Por tanto la recta de regresión pasa por todos los puntos de la nube. Así las dos rectas coinciden y hay dependencia lineal máxima entre las variables en sentido positivo

*) Si $r=-1$, entonces $\sigma_{ry}^2 = 0$

Es la misma interpretación q el caso de $r=1$, solo que ahora hay dep. lineal máxima en el sentido negativo.

*) Si $0 < r < 1$.

la correlación lineal será mayor cuanto más se aproxime a 1. (positiva)



*) Si $-1 < r < 0$

la correlación lineal será mayor cuanto más se aproxime a -1. (negativa). ▨

AJUSTE ~~PARÁBOLA~~ MINIMOS CUADRADOS DE LA PARABOLA.

Este es un caso de la regresión no lineal, los otros casos son ~~(de la regresión no lineal son)~~ hipérbola, equilátera, f. exponencial, f. potencial que se reducen al ajuste de la recta mediante una transformación adecuada

En este ajuste se consideran las variables discretas, aunque el resultado puede aplicarse para variables continuas, una vez expresado en términos de momento.

Sea $\hat{y}_i = a + bx_i + cx_i^2$, el estimador de y_i ~~deseado~~ y sea $e_i = y_i - \hat{y}_i$ el error que se comete en la estimación.

El método de mínimos cuadrados nos conduce a la parábola que hace mínima la función $S(a,b,c)$ con la suma de los cuadrados de los errores:

$$S(a,b,c) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$S(a,b,c) = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

de la siguiente manera:

SIST. DE ECUACIONES

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)^2(-1) = 0.$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)^2(-x_i) = 0.$$

$$\frac{\partial S}{\partial c} = \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-x_i^2) = 0.$$

Tenemos que despejar a, b, c.

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2. \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3. \\ \sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4. \end{cases}$$

Dividiendo por n obtenemos las anteriores ecuaciones expresadas en momentos.

$$\begin{cases} a_{01} = a + ba_{10} + ca_{20} \\ a_{11} = aa_{10} + ba_{20} + ca_{30} \\ a_{21} = aa_{20} + ba_{30} + ca_{40} \end{cases}$$

Resolviendo el sist. se obtienen los coeficientes de la parábola de regresión de mínimos cuadrados.

Nota.-

Es clave la importancia del ajuste de y parabólicas, por ejemplo: en microeconomía es conocido que la suma de costes marginales en función del volumen de producción es una parábola.

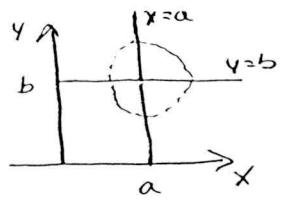
Interpretación de r. (continuación)

* x e y indep $\Rightarrow r_{xy} = 0$.

$r_{xy} = 0 \not\Rightarrow$ x e y indep

Si tiene potencia relación no lineal. $\Rightarrow$ $(x-a)^2 + (y-b)^2 = 1$.

Ejemplo:



ESTADISTICA

BONDAD DEL AJUSTE DE LA PARABOLA.

El método de mín. cuadrados para la parábola nos conducía a las ecuaciones:

$$\begin{cases} \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) = 0 \;=\; \sum_{i=1}^{n} e_i \\ \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)\,x_i = 0 \;=\; \sum_{i=1}^{n} e_i x_i \\ \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)\,x_i^2 = 0 \;=\; \sum_{i=1}^{n} e_i x_i^2 \end{cases}$$

Por tanto:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} e_i \cdot e_i = \sum_{i=1}^{n} e_i (y_i - a - bx_i - cx_i^2).$$

$$\boxed{\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} e_i y_i} - a\overbrace{\sum_{i=1}^{n} e_i}^{0} - b\overbrace{\sum_{i=1}^{n} e_i x_i}^{0} - c\overbrace{\sum_{i=1}^{n} e_i x_i^2}^{0} = 0.$$

$$\sigma_{ry}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} e_i y_i.$$

$$\sigma_{ry}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i (y_i - a - bx_i - cx_i^2)\right) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{a}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} y_i x_i - c\sum_{i=1}^{n} y_i x_i^2.$$

$$\sigma_{ry}^2 = a_{02} - aa_{01} - ba_{11} - ca_{21}$$

El coeficiente de determinación será:

$$R_p^2 = 1 - \frac{a_{02} - aa_{01} - ba_{11} - ca_{21}}{m_{02}}$$

$$\boxed{R_p^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}}$$

Notas:

* Si ajustamos una parábola de la forma $x = \hat{a} + \hat{b}y + \hat{c}y^2$ para expresar el comportamiento de x en función de y, la varianza residual y el coeficiente de determinación serán:

$$\sigma_{ry}^2 = a_{20} - \hat{a}a_{10} - \hat{b}a_{11} - \hat{c}a_{12}$$

$$R_p^2 = \frac{1 - a_{20} - \hat{a}a_{10} - \hat{b}a_{11} - \hat{c}a_{12}}{m_{20}}$$

* Debido a que la recta es un caso particular de parábola degenerada para $c = 0$, se obtendrán siempre mejores ajustes mediante parábolas que usando funciones lineales.

siendo el coeficiente de det. de la parábola siempre mayor q. el de la recta. $R_p^2 > R^2 = r^2$

$y = a + bx + cx^2 \rightarrow$ parábola.

$y = a + bx \Rightarrow$ recta (caso particular de la parábola).

* Como sabemos en el caso de la recta el coeficiente de determinación r^2 indica qué parte de la variancia de la variable independiente, es explicada por la regresión. Si calculamos este coeficiente sobre las rectas obtenidas, después de transformar las funciones potencial, exponencial y hipérbola (parábolas) equiláteras en rectas, no tendrán el mismo significado y no (tendrá) será comparable con valores obtenidos sobre la obtención del ajuste de una recta o parábola.

EJERCICIOS.

Nº 1

a)

X \ Y	1	2	3	4	5	6	n_i
1	0	1	0	0	0	3	4
2	1	0	1	0	1	0	3
3	1	0	0	2	0	0	3
4	2	3	0	0	1	1	7
5	2	1	1	0	0	1	5
6	0	1	0	0	1	0	2
n_j	6	6	2	2	3	5	24 = N

b) Medias marginales:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{6} n_i \cdot x_i = \frac{1}{24} \left[1(4) + 2(3) + 3(3) + 4(7) + 5(5) + 6(2) \right]$$

$= \boxed{3{,}5}$

ESTADISTICA

$\bar{Y} = \frac{1}{N}\sum_{j=1}^{6} n_{\cdot j}\, y_j = \boxed{3'21}$ Varianzas marginales

b) $\sigma_X^2 = \left(\frac{1}{N}\sum_{i=1}^{6} x_i^2 n_{i\cdot}\right) - \bar{X}^2$

$\sigma_X^2 = \frac{1}{24}\left(4(4) + 4(3) + 9(3) + 16(7) + 25(5) + 36(2)\right) - (2'5)^2$

$\boxed{\sigma_X^2 = 2'42}$

$\sigma_Y^2 = \left(\frac{1}{N}\sum_{i=1}^{6} y_i^2 \cdot n_{\cdot j}\right) - \bar{y}^2 = \boxed{3'65}$

c) Se han obtenido sólo dos pares (3,4) dos veces.

$\left.\begin{array}{l} 24 - 2 \\ 100 - x \end{array}\right\}\ x = \boxed{8'33\,\%}$

$\boxed{N=2}$

a) $Y \equiv$ salario (miles de ptas)

$X \equiv$ % en gastos de alimentación

$Y/X \geq 25$

$Y/X=25$	$n_{i/x_i \geq 25}$	$N_{i/x \geq 25}$
100 - 150	18	18
150 - 200	10	28 ←
200 - 250	12	40
250 - ∞	5	45
	$N = 45$	

$\alpha = \frac{45}{2} = 22'5$

$N = 22'5$. El próximo es el 28.

$Me \in [150, 200]$

$Me = L_{i-1} + \frac{N/2 - N_{i-1}}{10} \cdot 50$

$Me = 150 + \frac{22'5 - 18}{10} \times 50 =$

$\boxed{Me = 172'5 \text{ miles de ptas.}}$

b) La distribución marginal de Y es:

Y	$n_{\cdot j}$	$N_{\cdot j}$
100 - 150	18	18
150 - 200	12	30
200 - 250	15	45
250 - ∞	10	55
	55.	

P_r 175 $\in [150, 200)$

$175 = 150 + \frac{55c - 18}{12} \cdot 50$

$P_r = L_{i-1} + \frac{\frac{Nc}{100} - N_{i-1}}{n_i} a_i$

$\Rightarrow r = 43'6363\%$

$P_5 = 225 \in [200 - 250)$

$225 = 200 + \dfrac{\frac{5s}{100} - 30}{15} \cdot 50$

$\Rightarrow s = 68'1818\%$

La campaña de publicidad ha dirigido a $s - r =$ $24'54\%$ de familias

c) La distribución de $x/4 < 175$.

$x/4 < 175$	$n_i/x \leq 175$
$10 - 25$	1
$25 - 65$	14
$65 - 80$	9

$Y/x \in [10, 25)$	$n_j/4$	$N_j/4$
$100 - 150$	0	0
$\rightarrow$ $150 - 200$	2	2
$200 - 250$	3	5
$250 - \infty$	5	10
	10	

$Pr = 175 \in [150, 200)$

$175 = 150 + \dfrac{\frac{10}{100} - 0}{2} \cdot 50 \Rightarrow r = 10\%$ de $10 = \boxed{1}$

$Y/x \in [25, 65)$	$n_j/4$	$N_j/4$
$100 - 150$	10	10
$150 - 200$	8	18
$200 - 250$	10	28
$250 - \infty$	4	32
	32	

$Pr = 175 \in [150, 200)$

$175 = 150 + \dfrac{\frac{32s}{100} - 10}{8} \cdot 50$

$\Rightarrow r = 43'75\%$

$43'75\%$ de 32 es 14

$Y/x \in [65 - 80)$	$n_j/3$	$N_j/3$
$100 - 150$	8	8
$150 - 200$	2	10
$200 - 250$	2	12
$250 - \infty$	1	13
	13	

$Pr = 175 \in [150 - 200)$

$175 = 150 + \dfrac{\frac{12s}{100} + 8}{2} \cdot 50$

$\Rightarrow r = 69'23\%$

$69'23\%$ de $13 = 9$

ESTADISTICA

$Y/x<64$	$n_j/x_i<68$
100 - 150	• 11'6
150 - 200	• 10'4
200 - 250	• 13'4
250 - ∞	• 9'2

$Y/y \in [100,150)$	$n_i/_4$	$N_i/_4$
10-25	0	0
25-65	10	10
65-80	5	15
	15	

$P_3 = 68 \in [65,80)$

$68 = 65 + \frac{\frac{15}{100} - 10}{5} \cdot 15$

$\Rightarrow S = 64'4\%$

64'4% de 15 = 11'6

$Y/y \in [150-200)$	$n_i/_2$	$N_i/_2$
10-25	2	2
25-65	8	10
65-80	2	12
	12	

$P_2 = 68 \in [65,80)$

$68 = 65 + \frac{\frac{12}{100} - 10}{2} \cdot 15$

$\Rightarrow S = 86'6\%$

86'6% de 15 = 13

$Y/y \in [200-250)$	$n_i/_2$	$N_i/_2$
10-25	3	3
25-65	10	13
65-80	2	15
	15	

$P_2 = 68 \in [65,80)$

$68 = 65 + \frac{\frac{15}{100} - 13}{2} \cdot 15$

$\Rightarrow S = 89'3\%$

89'3% de 15 = 13'4

Y / X	mayor de 250
65-80	1

(64)

A calcular:

Uno repartido entre 15u. $[(80-65)=15]$ desde el 65 hasta el 80, entonces a cada unidad le corresponde $1/15$ luego en 3 que va desde el 65 al 68, le corresponde $3 \cdot \frac{1}{15} = \frac{1}{5} = 0'2$, por tanto para abajo del 68, hasta el principio, tenemos $5+4+0'2 = 9'2$

Aquí tenemos unas distribuciones condicionadas que son unidimensionales y ahora vamos a calcular su varianza y sus coeficientes de variación.

Y/X<60	n_j	y_j	$n_j \cdot y_j$	$n_j y_j^2$
100 - 150	11'6	125	1450	181250
150 - 200	10'4	175	1820	318.500
200 - 250	13'4	225	3015	678 375
250 - 300	9'2	275	2530	695 150
	44'6		8815	1873275

$\bar{Y} = 197.6454$

$\sum y^2$ 2942'260154

$\Rightarrow \sigma_y$ 54'251

$Cv = \frac{\sigma_y}{\bar{Y}} = 0'27467$

$\bar{Y} = \frac{8815}{44'6}$ $\bar{Y}^2 = 39063'82213$

X/Y<175	n_i	x_i	$n_i \cdot x_i$	$n_i \cdot x_i^2$
10 - 25	1	17'5	17'5	306' 25
25 - 65	14	45	630	28350
65 - 80	9	72'5	652'5	47306'25
			1300	75962'5

$\bar{X} = 54'16$

σ_x^2 281'046

$\Rightarrow \sigma_x$ 15'201

$Cv_x = \frac{\sigma_x}{\bar{x}} = 0'2806$

$Cv_y = 0'27467$

- Más homogénea la de y para gastos de alimentación que no superan el 68% del sueldo

Más cerca de 0

$0'27 < 0'28$

$0'27$

Relación 2.

(2.d)

x_i	y_i	n_{ij}	$n_{ij}x_i$	$n_{ij}y_j$	$n_{ij}x_iy_j$	$n_{ij}x_i^2$	$n_{ij}y_j^2$
17'5	175	2	35	350	6125	612'5	61250
17'5	225	3	52'5	675	11812'5	918'75	918'75
17'5	275	5	87'5	1375	240625	1531'25	378125
45	125	10	450	1250	56.250	20250	156.250
45	175	8	360	1400	63000	16.200	245.000
45	225	10	450	2250	101.250	20250	506.250
45	275	4	180	1100	495000	8.100	302.500
72'5	125	8	580	1000	72.500	42050	125.000
72'5	175	2	145	350	25.375	10512'5	61.250
72'5	225	2	145	450	32.625	10.512'5	101.250
72'5	275	1	72'5	275	19937'5	5256'25	75.650
		55	2557'5	104.575	462.707'5	136.193'75	2164.375

$\bar{X} = 46'5$.

$\sigma_x^2 = 314 \Rightarrow \sigma_x = 17'72$

$\bar{Y} = 190'45$

$\sigma_y^2 = 3079'338845 \Rightarrow \sigma_y = 55'49179$

$\sigma_{xy} = -443'27$.

$\sigma_{xy} = a_{11} - a_{01}a_{10}$

$\sigma_{xy} = \frac{1}{N} \sum n_{ij} x_i y_j - \bar{x}\bar{y}$

$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = -0'45 \rightarrow$ No es fiable.

Conclusión, la recta de regresión se hallá para los valores entre 100 y 250.

TEMA 2: VARIABLES ESTADÍSTICAS BIDIMENSIONALES. REGRESIÓN Y CORRELACIÓN.

2.1.- INTRODUCCIÓN

X.- Variable estadística unidemensional q. representa un caracter.

X.- } dos caracteres (x,y)
Y.- }

Vamos a estudiar variables estadísticas q. representan 2 caracteres x e y.
por ej: peso (x) e (y) altura.

El par (x,y) es una variable estadística bidimensional.

$$\left.\begin{array}{l} X: x_1, x_2, x_3, \ldots x_p \\ Y: y_1, y_2, y_3, \ldots y_q \end{array}\right\} (x,y) \rightarrow (x_i, y_j) \quad \begin{array}{l} i=1,2,\ldots p \\ j=1,2,\ldots q \end{array}$$

- Frecuencia absoluta: n_{ij} del par (x_i, y_j): es el nº de individuos q. han presentado el valor x_i de X e y_j de Y
- Frecuencia relativa: f_{ij}: es $= \frac{n_{ij}}{N}$

$$f_{ij} = \frac{n_{ij}}{N}$$

N = nº total de individuos

$$N = \sum_i^p \sum_j^q n_{ij}$$

- Distribución bidimensional de frecuencias: conjunto de valores

$$\{((x_i, y_j), n_{ij}); i=1,\ldots p; j=1,\ldots q\}$$

2.2.- REPRESENTACIONES NUMERICAS

* Cuando las variables x e y son discretas, continuas o una discreta y la otra continua la forma usual de representar su distribución bidimensional de frecuencias es mediante una tabla de doble entrada.

$(x, y) \Longrightarrow$ Variable bidimensional mixta.

(x: continua, y: discreta)

Ej: Sea la distribución dada x la siguiente tabla:

X	Y	n_{ij}
20	50-100	10
20	100-150	3
20	150-200	2
21	50-100	5
21	100-150	15
21	150-200	5
22	50-100	2
22	100-150	10
22	150-200	15
23	50-100	13
23	100-150	10
		$N = 100$

X \ Y	50-100 ($y_1 = 75$)	100-150 ($y_2 = 125$)	150-200 ($y_3 = 175$)	$n_{i\bullet}$
$x_1 = 20$	$n_{11} = 10$	$n_{12} = 3$	$n_{13} = 2$	$15 = n_{1\bullet}$
$x_2 = 21$	$n_{21} = 5$	$n_{22} = 15$	$n_{23} = 5$	$25 = n_{2\bullet}$
$x_3 = 22$	$n_{31} = 2$	$n_{32} = 20$	$n_{33} = 15$	$37 = n_{3\bullet}$
$x_4 = 23$	$n_{41} = 0$	$n_{42} = 13$	$n_{43} = 10$	$23 = n_{4\bullet}$
$n_{\bullet j}$	$n_{\bullet 1} = 17$	$n_{\bullet 2} = 51$	$n_{\bullet 3} = 32$	$N = 100$

- En la columna $n_{i\bullet}$ se representa el nº de veces q. se ha observado el valor x_i de la variable X. sin tener en cuenta el valor presentado por Y.

  - $n_{i\bullet}$: se obtiene sumando todas las frecuencias de la fila correspondiente al valor x_i de X.

$$n_{i\bullet} = \sum_{j=1}^{q.} n_{ij} \quad i \text{ fijo.}$$

Lo mismo para $n_{\bullet j} = \sum_{i=1}^{p.} n_{ij}$ (j fijo)

$$- N = \sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j}$$

23.- DISTRIBUCIONES MARGINALES.

Ejemplo:

X	$n_{i\bullet}$	$f_{i\bullet}$
20	15	0'15
21	25	0'25
22	37	0'37
23	23	0'23
	N=100	1

$$f_{i\bullet} = \frac{n_{i\bullet}}{N}$$

Con la 1ª y la última columna de la tabla de doble entrada del ej. anterior obtenemos la distribución unidimensional.

Ejemplo:

$L_{i+1} - L_i$	Y	$n_{\bullet j}$	$f_{\bullet j}$
50-100	75	17	0'17
100-150	125	51	0'51
150-200	175	32	0'32
			1

$$f_{\bullet j} = \frac{n_{\bullet j}}{N}$$

La 1ª fila y la última de la tabla de doble entrada del ej. anterior nos da la distribución unidimensional

* Estas distribuciones se llaman distribuciones marginales de X e Y.

* $f_{i\bullet}$ y $f_{\bullet j}$ → frecuencias relativas marginales

Por ser distribuciones unidimensionales, a las distribuciones marginales podemos aplicar todo lo visto en el tema 1.

Con este cambio de notaciones los momentos se notan con su notación del tema 1 acompañada de un subíndice indicando a q. variable se refiere.

* Media marginal de X:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} x_i n_{i\bullet} = \sum_{i=1}^{p} x_i f_{i\bullet}$$

* Media marginal de Y

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{q} y_j n_{\bullet j} = \sum_{j=1}^{q} y_j f_{\bullet j}$$

* Varianza marginal de X:

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^{p} (x_i - \bar{x})^2 n_{i\bullet} = \sum_{i=1}^{p} (x_i - \bar{x})^2 \cdot f_{i\bullet} = \left(\frac{1}{N} \sum_{i=1}^{p} x_i^2 n_{i\bullet} \right) - \bar{x}^2$$

* Varianza marginal de Y:

$$\sigma_y^2 = \frac{1}{N} \sum_{j=1}^{q} (y_j - \bar{y})^2 n_{\bullet j} = \sum_{j=1}^{q} (y_j - \bar{y}) f_{\bullet j} = \frac{1}{N} \sum_{j=1}^{q} y_j^2 n_{\bullet j} - \bar{y}^2$$

2.4.- DISTRIBUCIONES CONDICIONADAS

Estas distribuciones condicionadas son: distribuciones unidimensionales obtenidas a partir de las bidimensionales manteniendo fijo el valor de una de la variables y considerando los valores de la otra con sus respectivas frecuencias

- La distribución condicionada de X dado q. $Y = y_j$ se obtiene a partir de la tabla de doble entrada tomando la 1ª columna de los valores de la variable X y la jésima columna de las frecuencias absolutas.

- La distribución condicionada de Y dado q. $X = x_i$ se obtiene a partir de la tabla de doble entrada tomando la 1ª fila de los valores de i y la iésima fila de las frecuencias absolutas.

Ejemplo: Queremos obtener la distribución condicionada para Y dado q. $X = x_2 = 21$. (Según el ej. de la tabla de doble entrada).

$y_j / x_2 = 21$	$n_{j/2}$
50 - 100	$n_{21} = 5$
100 - 150	15
150 - 200	3
	$n_{2\bullet} = 23$

$\Rightarrow$ Representa la distribución condicionada de Y dado q $X = x_2 = 21$

Las frecuencias relativas en estas distribuciones condicionadas se definen como el cociente entre la frecuencia absoluta y el nº total de observaciones.

Para las distribuciones condicionadas Y/x_i notaremos las frecuencias relativas $f_{j/i}$ "Y condicionado por"

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}}$$

Esta frecuencia representa la proporción de individuos q. cumplen la condición $X = x_i$ y q. han presentado el valor $Y = y_j$

- La distribución condicionada de X dado q. $Y = y_j$ su frecuencia relativa será $f_{i/j}$

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$$

Representa la proporción de individuos q. cumplen la condición $Y = y_j$ y han presentado el valor $X = x_i$

$$\sum_{i=1}^{p} f_{i/j} = \sum_{i=1}^{p} \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{\cdot j}}{n_{\cdot j}} = \frac{1}{n_{\cdot j}} \underbrace{\sum_{i=1}^{p} n_{ij}}_{n_{\cdot j}} = \frac{n_{\cdot j}}{n_{\cdot j}} = 1$$

$$\sum_{j=1}^{q} f_{j/i} = \sum_{j=1}^{q} \frac{n_{ij}}{n_{i\cdot}} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{q} n_{ij} = \frac{n_{i\cdot}}{n_{i\cdot}} = 1$$

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{ij}/N}{n_{\cdot j}/N} = \frac{f_{ij}}{f_{\cdot j}}$$

$$f_{j/i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{ij}/N}{n_{i\cdot}/N} = \frac{f_{ij}}{f_{i\cdot}}$$

Ejemplo: A partir de la tabla de doble entrada del ej: 1º vamos a obtener la distribución condicionada de X dado q. $Y = y_2 = 125$

	n_{i2}	$f_{i/2}$
20	3	[illegible]
21	15	[illegible]
22	10	[illegible]
23	13	[illegible]
	51	1
	$n_{\cdot 2}$	

$$f_{1/2} = \frac{n_{12}}{n_{\cdot 2}} = \frac{3}{51} = 0,058$$

$$f_{21/2} = \frac{n_{22}}{n_{\cdot 2}} = \frac{15}{51} = 0,29$$

Las distribuciones condicionadas son distribuciones unidimensional a las cuales se les puede aplicar todo lo conocido en este tipo de distribuciones. A las características de estas distribuciones se les añade el calificativo "condicionadas".

* Medias condicionadas xa distribuciones de x/y_j

$$\bar{x}_j = m_{x/y_j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} x_i n_{ij} = \sum_{i=1}^{p} x_i \cdot f_{i/j}$$

* Medias condicionadas xa la distribución de Y/x_i

$$\bar{y}_i = m_{y/x_i} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{q} y_j n_{ij} = \sum_{j=1}^{q} y_j f_{j/i}$$

* Variança condicionada de x/y_j

$$\sigma_j^2(x) = \sigma_{x/y_j}^2 = \frac{1}{n_{\cdot j}} \sum_{i=1}^{p} (x_i - \bar{x}_j)^2 n_{ij} = \sum_{i=1}^{p} (x_i - \bar{x}_j) f_{i/j}$$

* Variança condicionada de Y/x_i

$$\sigma_i^2(y) = \sigma_{y/x_i}^2 = \frac{1}{n_{i\cdot}} \sum_{j=1}^{q} (y_j - \bar{y}_i)^2 n_{ij} = \sum_{j=1}^{q} (y_j - \bar{y}_i) f_{j/i}$$

2.5.- Relación entre Distribuciones Marginales y Condicionadas:

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}} \Rightarrow f_{i/j} = f_{i/j} \cdot f_{\cdot j}$$

$$f_{j/i} = \frac{f_{ji}}{f_{i\cdot}} \Rightarrow f_{ij} = f_{j/i} \cdot f_{i\cdot}$$

La relación existent en estas dos distribuciones puede resumirse en la siguiente relación:

$$\boxed{f_{ij} = f_{i/j} \cdot f_{\cdot j} = f_{j/i} \cdot f_{i\cdot}}$$

De estas relaciones se obtienen las siguientes:

$$\bar{x} = \sum_{j=1}^{q} \bar{x}_j f_{\bullet j} \qquad \bar{y} = \sum_{i=1}^{p} \bar{y}_i f_{i\bullet}$$

$$\sigma_x^2 = \sum_{j=1}^{q} \sigma_j^2(x) f_{\bullet j} + \sum_{j=1}^{q} (\bar{x}_j - \bar{x})^2 f_{\bullet j}$$

$$\sigma_y^2 = \sum_{i=1}^{p} \sigma_i^2(y) f_{i\bullet} + \sum_{i=1}^{p} (\bar{y}_i - \bar{y})^2 f_{i\bullet}$$

2.6. – INDEPENDENCIA ESTADÍSTICA

- Dos variables x e y son estadística/ independientes cuando la variación de una de ellas no influye en la otra; es decir si la distribución condicionada de $y_{\downarrow}$ es independiente del valor de x_i

  a $x = x_i$

- La condición necesaria y suficiente para q. dos variables estadísticas sean independientes es q. la frecuencia relativa conjunta sea igual al producto de las marginales.

$$\boxed{f_{ij} = f_{i\bullet} \cdot f_{\bullet j}}$$

- x e y son independientes estadísticamente si y solo si la $f_{i/j} = f_{i\bullet}$

$$f_{i/j} = \frac{f_{ij}}{f_{\bullet j}} = \frac{f_{i\bullet} f_{\bullet j}}{f_{\bullet j}} = f_{i\bullet}$$

x e y son independientes estadística/ $\Leftrightarrow f_{j/i} = f_{\bullet j}$

Ejemplo: en la siguiente tabla de doble entrada comprobar si x e y son independientes.

x \ y	y_1	y_2	y_3	y_4	$n_{i\bullet}$
x_1	3	5	2	4	14
x_2	6	10	4	8	28
x_3	12	20	8	16	56
$n_{\bullet j}$	21	35	14	28	98

Tabla de distribuciones condicionadas a $y=y_j$

x	$y=y_1$	$y=y_2$	$y=y_3$	$y=y_4$	$p_{i\bullet}$
x_1	3/21	5/35	2/14	4/28	14/98
x_2	6/21	10/35	4/14	8/28	28/98
x_3	12/21	20/35	8/14	16/28	56/98

$\boxed{f_{i/j} = p_{i\bullet}}$ $\quad \frac{3}{21} = \frac{5}{35} = \frac{2}{14} = \frac{4}{28} = \frac{14}{98} \left(\frac{1}{7}\right)$ cumple

$\frac{6}{21} = \frac{10}{35} = \frac{4}{14} = \frac{8}{28} = \frac{28}{98} \left(\frac{2}{7}\right)$ cumple

$\frac{12}{21} = \frac{20}{35} = \frac{8}{14} = \frac{16}{28} = \frac{56}{98} \left(\frac{4}{7}\right)$ cumple

x e y son independientes

2.7.- DEPENDENCIA FUNCIONAL.

Dada la variable estadística bidimensional (x,y) diremos q. y depende funcional/ de x si $y = F(x)$ es decir a cada valor o modalidad de x le corresponde un único valor o modalidad de y. ; por lo q. en cada fila o columna de la tabla hay un solo valor $\neq$ de **0**.

Lo normal es q. conocidos los valores de una variable sepamos parcial/ q. ocurre con los valores de la otra surge así el concepto de dependencia estadística

Dependencia estadística ⟶ conocemos los valores parcial/
Dependencia funcional ⟶ conocemos todos los valores de la otra variable mediante una función.

2.8.- REPRESENTACIONES GRÁFICAS

A) LA NUBE DE PUNTOS.

Consiste en representar en unos ejes cartesianos los puntos : $\{(x_i, y_j), i = 1,2 \ldots p ; j = 1, \ldots q\}$ colocando un punto con coordenadas = a (x_i, y_j)

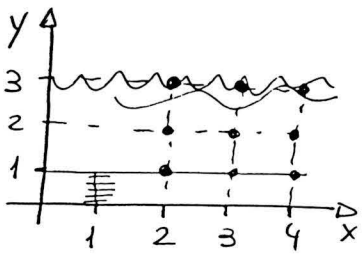
Si ambas variables son discretas se representa un punto xa cada par de valores observados, cuando una pareja se ha observado + de una vez junto al punto se coloca el valor de la frecuencia. ó bien se transforman dichos puntos en círculos de tamaños proporcionales a la frecuencia.

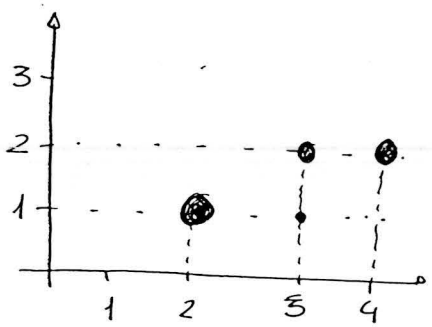
Si ambas variables & son continuas tomamos las marcas de clase como representantes de cada intervalo, y así estaríamos en la = situación q. las variables discretas.

Otra manera de representar las variables continuas es representar en cada rectángulo determinado x los intervalos de "x" e "y" tantos puntos como indiq. la frecuencia observada

Ejemplo: (caso discreto)

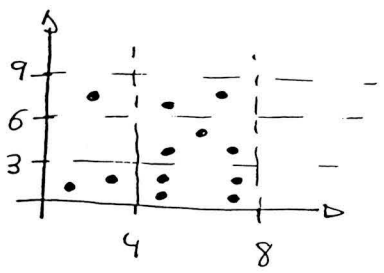
x \ y	1	2
2	5	0
3	1	2
4	0	3





(caso continuo)

x \ y	0-3	3-6	6-9
0-4	2	0	1
4-8	4	3	2.

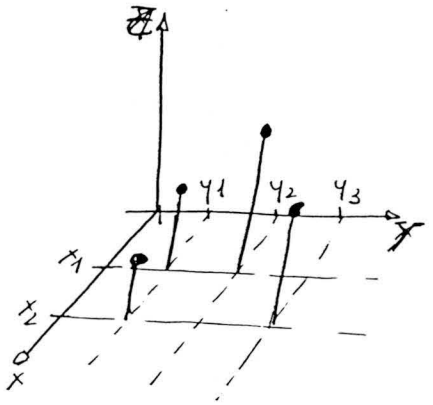


B) ESTEREOGRAMA ó ESCALOGRAMA

Variables discretas: consiste en representar sobre el plano (x,y) los valores (x_i, y_j) y levantar sobre cada uno de estos puntos un segmento o barra de longitud = a la frecuencia con q. se ha observado dicho par de valores.

Ejemplo:

X \ Y	y_1	y_2	y_3
x_1	3	4	0
x_2	1'5	0	2



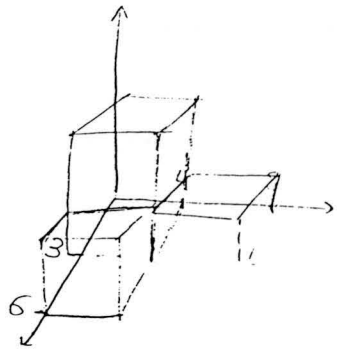
Variables continuas: se marcan los intervalos sobre los ejes x e y dando lugar a rectángulos sobre el plano (x,y) y sobre cada uno se construye un paralelepípedo de volumen = ó proporcional a la frecuencia q. se quiere representar.

- Si $L_i - L_{i-1}$ → amplitud del iésimo intervalo de x
- Si $L_j - L_{j-1}$ → amplitud del jésimo intervalo de y
- h_{ij} → altura del paralelepípido construido sobre la clase iésima de x y la clase jésima de y.
- V_{ij} → volumen del paralelepípido construido sobre la clase iésima de x y la clase jésima de y.

$$V_{ij} = n_{ij} = (L_i - L_{i-1})(L_j - L_{j-1})\, h_{ij} \Longrightarrow h_{ij} = \frac{n_{ij}}{(L_i - L_{i-1})(L_j - L_{j-1})}$$

Ejemplo:

x \ y	0-4	4-8
0-3	5	3
3-6	4	0



2.9.- MOMENTOS DE UNA DISTRIBUCIÓN BIDIMENSIONAL.

A) Definición del momento no centrado: (o respecto al origen) de órdenes r y s de una variable bidimensional:

- Variables continuas:

$$a_{rs} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i^r y_i^s \cdot f_{ij} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} x_i^r y_j^s n_{ij}$$

- Variables discretas:

$$a_{rs} = \frac{1}{N}\sum_{i=1}^{\hat{}} x_i^r y_i^s$$

Casos particulares:

$$a_{10} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i^{1} y_j^{0} f_{ij} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i f_{ij} = \sum_{i=1}^{p} x_i \underbrace{\sum_{j=1}^{q} f_{ij}}_{f_{i\bullet}} = \sum_{i=1}^{p} x_i \cdot f_{i\bullet} = \bar{x}$$

$$a_{01} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i^0 y_j f_{ij} = \bar{y}$$

$$a_{20} = a_2(x) \qquad a_{02} = a_2(y)$$

Los momentos de segundo orden respecto al origen de las distribuciones marginales de x e y vendrán dados por a_{20} y a_{02}; en general cualq. momento bidimensional en q uno de los subíndices sea 0 se corresponde

con un momento unidimensional de la distribución marginal de "x" e "y".

- Momento + utilizado:

$$a_{11} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} x_i y_j n_{ij} \longrightarrow \text{V. continuas}$$

$$a_{11} = \frac{1}{n} \sum_{i=1}^{n} x_i y_j \longrightarrow \text{V. discretas}$$

B-) Definición de momento centrado con respecto a las medias de orden r y s

- Variables continuas: $m_{rs} = \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})^r (y_i - \bar{y})^s f_{ij}$.

$$m_{rs} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})^r (y_j - \bar{y})^s n_{ij}.$$

- Variables discretas: $m_{rs} = \frac{1}{n} \sum_{i=1}^{p} (x_i - \bar{x})^r (y_i - \bar{y})^s$

Casos particulares:

$m_{10} = 0 \qquad m_{01} = 0$

$m_{20} = \sigma_x^2$ varianza marginal de x

$m_{02} = \sigma_y^2$ varianza marginal de y

Covarianza: el momento respecto a las medias + importante

- Variables continuas: $m_{11} = \sigma_{xy} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})(y_i - \bar{y}) n_{ij}$

- Variables discretas: $m_{11} = \sigma_{xy} = \frac{1}{n} \sum_{i=1}^{p} (x_i - \bar{x})(y_i - \bar{y})$

La covarianza es una medida de la variabilidad conjunta de "x" e "y" y a tanto de la relación entre las variables "x" e "y".

Covarianza — positiva: las dos variable varían en el mismo sentido si una ↑ la otra ↑

Covariancia negativa → las variables varían en sentido opuesto
Si una ↑ la otra ↓

RELACIÓN ENTRE LOS a_{rs} Y m_{rs}

Los momentos centrados se pueden expresar en función de los momentos no centrados.

$$m_{20} = \sigma_x^2 = a_{20} - a_{10}^2$$

$$m_{02} = \sigma_y^2 = a_{02} - a_{01}^2$$

Son válidas tb. xa distribuciones unidimensionales

RELACIÓN ENTRE LA σ_{xy}, a_{rs} Y m_{rs}

$$\sigma_{xy} = a_{11} - a_{01}\,a_{10}$$

PROPIEDAD DE LA COVARIANZA

* Si "x" e "y" son independientes entonces la covariancia de x e y es cero

$$x \text{ e } y \text{ indepad} \Longrightarrow \sigma_{xy} = 0$$

$$\sigma_{xy} = 0 \not\Longrightarrow x \text{ e } y \text{ independientes}$$ no siempre se cumple

* La covariancia puede variar sobre todo $\mathbb{R}$ desde $-\infty$ hasta $+\infty$, no está acotada, entonces no sabemos si el grado de asociación entre "x" e "y" es + o − fuerte.

Coeficiente de correlación lineal: $r_{xy} = \dfrac{m_{11}}{\sqrt{m_{20}\,m_{02}}} = \dfrac{\sigma_{xy}}{\sigma_x\,\sigma_y}$

Varía: $-1 < r < 1$.

r: es la covariancia sobre las variables tipificadas

$$x' = \frac{x - \bar{x}}{\sigma_x} \qquad y' = \frac{y - \bar{y}}{\sigma_y} \qquad \sigma_{x'y'} = r_{xy} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$$

Como está acotada nos puede decir el grado de variación conjunta de las variables x' e y' y a tanto las variables "x" e "y".

Cambio de origen y escala de los momentos centrados.

$$\left.\begin{aligned} x' &= \frac{x-a}{h} \\ y' &= \frac{y-b}{e} \end{aligned}\right\} \begin{array}{l} a.- \text{ cambio de origen } .-b \\ b.- \quad " \quad " \text{ escala } .-e \end{array} \Rightarrow \begin{aligned} x &= hx' + a \\ y &= ey' + b \end{aligned} \Rightarrow \begin{aligned} \bar{x} &= h\bar{x}' + a \\ \bar{y} &= e\bar{y}' + b \end{aligned} \Rightarrow$$

$\Rightarrow$ Por tanto $m_{rs}(x,y) = \sum_{i=1}^{p}\sum_{j=1}^{q} (x_i - \bar{x})^r (y_j - \bar{y})^s f_{ij}$

Sustituyendo $x_i = hx_i' + a$

$y_j = ey_j' + b$

Obtenemos $m_{rs}(x,y) = h^r e^s m_{rs}(x',y')$

Casos particulares: $\boxed{m_{11}(x,y) = he\, m_{11}(x',y')} \Rightarrow \sigma_{xy} = he\,\sigma_{x'y'}$

$$\left.\begin{aligned} m_{20}(x,y) &= h^2 e\, m_{20}(x',y') \\ m_{02}(x,y) &= e^2 m_{02}(x',y') \end{aligned}\right\} \begin{array}{l} \boxed{\sigma_x^2 = h^2 \sigma_{x'}^2} \\ \boxed{\sigma_y^2 = e^2 \sigma_{y'}^2} \end{array} \text{ Sacando la Raíz cuadrada}$$

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{he\,\sigma_{x'y'}}{h\,\sigma_{x'}\, e\,\sigma_{y'}} = r_{x'y'}$$

Por tanto el coeficiente de correlación lineal no varía ante un cambio de origen y escala.

2.10.- REGRESIÓN y CORRELACIÓN SIMPLE.

INTRODUCCIÓN.

∃ variables como la oferta y la Demanda, el consumo y la renta, ... entre las q. ∃ una relación no es imposible definir sobre ellas una función matemática q verifiq. exactal.

Este tipo de dependencia entre variables se denomina dependencia estadística frente a la depedencia funcional en la q. si hay una función matemática q. lo satisface de forma exacta.

La Regresión pretende poner unas variables en función de otras mediante una ley.

Las aplicaciones + interesantes de la regresión son:

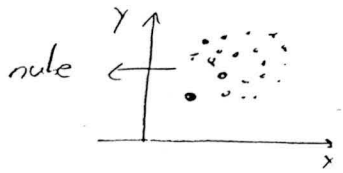
- conocer mejor un fenómeno
- predecir o estimar el valor de una variable conociendo el valor de otras relacionadas con ella.

Hecha la predicción inmediatal surge la duda sobre su fiabilidad. La respuesta a esa duda estará en gran parte dada x el estudio de la correlación. La variable q. se quiere predecir se llama dependiente o endógena

Las variables cuyo conocimiento se usa xa la predicción se llaman independientes o exógenas. Cuando sólo se usa una variable independiente (exógena) estaremos ante la regresión o correlación simple. Si interviene + de una variable independiente lo relación y correlación se llama múltiple.

AJUSTE POR MÍNIMOS CUADRADOS.

Si entre dos variables no ∃ una dependencia funcional es imposible encontrar una función entre ellas cuya representación gráfica pase x todos los puntos del diagrama de dispersión.



Es imposible encontrar una curva q pase x todos los puntos

Hay q. buscar la curva q pase x el nº máximo de puntos

Del diagrama de dispersión buscamos la curva q aunq. no pase x todos los puntos de la nube sí menos esté lo + próxima posible a ellos, es decir se ajuste mejor.

Uno de los criterios + adecuados xa encontrar dicha curva es el llamado "Ajuste x mínimos cuadrados"

En una variable estadística (x,y) si no hay una dependencia funcional entre "x" e "y" no se puede afirmar q. a cada valor de una de las variables le corresponda unívoca uno de la otra, sin embargo es fácil pensar q. el comportamiento x ejemplo de la variable condicionada $Y/X = x_i$ xa cada "x_i" difiera del comportamiento de y. sobre

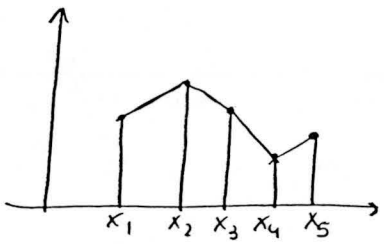
Son bien conocidas las cualidades de la media aritmética como el valor representativo del comportamiento de una variable.

LÍNEA de REGRESIÓN Y sobre X:

- Es la representación gráfica de $\{(x_i, \bar{y}_i)\ ;\ i = 1, \ldots, p\}$ unidos los x una línea.

$\bar{y}_i \equiv$ media de $Y/X = x_i$

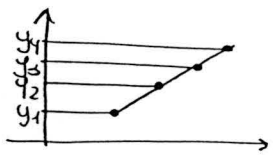
Esta línea nos indica las distribuciones de la variable "y" condicionada x los valores x_i de la variable "x"



Línea de regresión de X sobre Y: representación gráfica de

$\{(\bar{x}_j, y_j)\ ;\ j=1,\ldots q\}$

$\bar{x}_j$ → media de $X/y = y_j$



Nota: Aquí trabajamos con un conjunto de puntos no si "x" o "y" fuesen continuas se habría una verdadera curva de regresión.

Forma q. adopta la curva de regresión según el grado de dependencia entre "x" e "y".

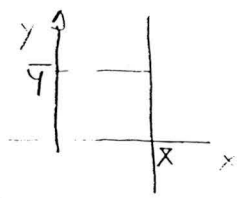
1º.-) "x" "y" independientes funcional:

- las distribuciones condicionadas $X/y = y_j$ serán = entre sí e = a la media marginal correspondiente.

$\bar{x}_j = \bar{x}$; $j=1,\ldots q$ $\quad x_1 = x_2 = x_3 = \ldots x_q = \bar{x}$

Por lo q el conjunto $\{(\bar{x}_j, y_j),\ j=1,\ldots q\}$ tienen la misma abcisa

Por tanto la curva de regresión de "x" e "y" es paralela al eje OY pasando por el par $(\bar{x}, \bar{y})$



- Análogamente las distribuciones condicionadas $Y/x = x_i$

tenemos q. $\bar{y} = \bar{y}_i$ $\quad y_1 = y_2 = y_3 \ldots y_p = \bar{y}$

entonces todos los puntos del conjunto $\{(x_i, \bar{y}_i),\ i=1\ldots p\}$ tienen la misma ordenada.

- Por tanto, la curva de regresión de Y sobre X es paralela al eje OX pasando por $(\bar{x}, \bar{y})$.

3.º) Dependencia funcional:

- "y" depende funcional de "x" si a cada valor x_i de "X" le corresponde un único valor ~~de~~ y_j de "y".

Entonces $Y/X = x_i$ tiene un solo valor x lo q. la media de esta variable $\bar{y}_i$ es = a $y_j \Longrightarrow$ el conjunto q. queremos representar es $\{(x_i, \bar{y}_i)\} = \{(x_i, y_j\}$ es decir los valores de x explican perfectamente los valores de y. $\Rightarrow y = F(x)$

- Análoga/ se dice q. la variable "x" depende funcional de "y" si a cada valor y_j de "y" le corresponde un único valor posible de "x". Entonces este conjunto q. queremos representar va a ser $\{(\bar{x}_j, y_j)\} = \{(x_i, y_j\}$ entonces a $X/Y = y_j$ le corresponde un único valor.

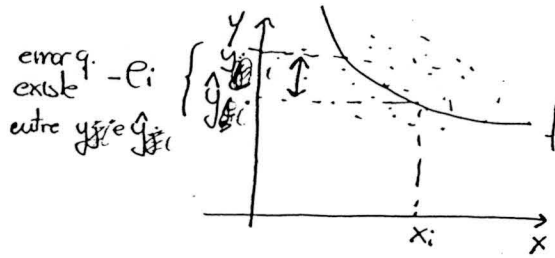
$$\bar{x}_j = x_i \qquad X = g(Y)$$ - va a ser una curva.

3º.-) Caso general: Caso intermedio: la dependencia y la independencia funcional son casos extremos q. se encuentran raral en la práctica. En la práctica encontramos el caso intermedio de una información parcial.

Supongamos q. para predecir la variable "y" conocemos otra variable relacionada con ella "x" de alguna forma, x ejemplo de la forma $Y = f(x)$ en la q.ro conocemos la función "f", tendríamos q. considerar $x_1, x_2, \ldots x_p$ modalidades de x, e, $y_1, y_2, y_3, \ldots y_q$ modalidades de y con un total de observaciones $n \equiv N$, no vamos a considerar todos los datos repetidos o no de la forma $x_1, \ldots x_n$

$y_1, \ldots y_n$

Si $x = x_i$ utilizando la mencionada función estimamos un valor $\hat{y}_i = f(x_i)$, $\hat{y}_i$ es una estimación del valor real y_i



Posible/ $\hat{y}_i$, y_i no sean =, se comete un error en la estimación $e_i = y_i - \hat{y}_i$. A este error se denomina residuo. Un criterio xa un buen ajuste es el de minimizar la suma de todos los residuos o errores al cuadrado

$$\text{minimizar} \Rightarrow \sum_{i=1}^{n} e_i^2 \longrightarrow \text{método de mínimos cuadrados}$$

$$\text{ll}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Mínimos cuadrados

Definición: el criterio de mínimos cuadrados considera q. la función q. mejor se ajusta a unos datos observados es aq.lla q. minimiza la suma de los cuadrados de los residuos y se comprueba q. coincide con la curva de regresión de Y sobre X.

Las funciones f q. se ajustan con + frecuencia son:

- f es una recta: $y = f(x) = a + bx$
- f es una parábola: $a + bx + cx^2 = y$
- f es una hipérbola equilátera: $y = a + b\frac{1}{x}$
- f es una función potencial: $y = ax^b$
- f es una función exponencial: $y = a \cdot b^x$

LA RECTA DE REGRESIÓN

Si la función q. ajustamos es una recta, la regresión se llam lineal, de entre todas las rectas $y=a+bx$ buscamos aq.lla q. mejor se ajuste según el criterio de mínimos cuadrados; para $x=x_i$, $\hat{y}_i = a+bx_i$

y $e_i = y_i - \hat{y}_i = y_i - a - bx_i$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

El criterio de mínimos cuadrados en este caso trata de encontrar los coeficientes a y b, q. hagan mínima la expresión $\underbrace{\sum_{i=1}^{n} (y_i - a - bx_i)^2}_{S''(a,b)}$

Condición necesaria xa la Exist de mínimo en un punto es q. las derivables parciales de primer orden se anulen en dicho punto.

$$S(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \left.\begin{array}{l} \frac{\partial S(a,b)}{\partial a} = 0 \\ \frac{\partial (a,b)}{\partial b} = 0 \end{array}\right\} \Longrightarrow$$

derivada con respecto a "a" considerando b constante

$$\left.\begin{array}{l} 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0 \\ 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0 \end{array}\right\} \Longrightarrow \left.\begin{array}{l} -\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ -\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 \end{array}\right\} \text{Ecuaciones normales de la recta.}$$

$$\left.\begin{array}{l} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (a + bx_i) \\ \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} (a + bx_i)x_i \end{array}\right\} \left.\begin{array}{l} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{array}\right\} \Longrightarrow$$

$\Longrightarrow$ Dividimos ambas ecuaciones por n obtenemos.

$$\left.\begin{array}{l} \frac{1}{n}\sum_{i=1}^{n} y_i = a + b\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ \frac{1}{n}\sum_{i=1}^{n} x_i y_i = a\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + b\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) \end{array}\right\} \left.\begin{array}{l} \bar{y} = a + b\bar{x} \\ a_{11} = a\bar{x} + b\,a_{20} \end{array}\right\} \Longrightarrow$$

$a_{01} = a + b a_{10}$ Estas ecuaciones son válidas sea variable discreta y continua. El sistema anterior

$\Rightarrow$ $a_{11} = a \cdot a_{10} + b a_{20}$

#3 puede resolverse: multiplicamos la 1ª ecuación por $(-a_{10})$ y la sumamos a la 2ª ecuación

$$-a_{10} a_{01} = -a a_{10} - b a_{10}^2$$
$$+ \quad a_{11} = a a_{10} + b a_{20}$$

$$a_{11} - a_{10} a_{01} = b (a_{20} - a_{10}^2) \Rightarrow \boxed{b = \frac{a_{11} - a_{10} a_{01}}{a_{20} - a_{10}^2}}$$

$\boxed{b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{m_{11}}{m_{20}}}$ Sustituimos el valor de "b" en la 1ª ecuación y obtenemos "a"

$$a_{01} = a + \frac{m_{11}}{m_{20}} a_{10}$$

$\boxed{a = a_{01} - \frac{m_{11}}{m_{20}} a_{10}}$ $\boxed{a = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x}}$

La Recta de Regresión de Y sobre X sea

$$y = a + bx = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x} + \frac{\sigma_{xy}}{\sigma_x^2} \cdot x \Rightarrow \boxed{y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} (x - \bar{x})} = x \text{ atrás}$$

Análogo la Recta de Regresión de X sobre Y sería

$\boxed{x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2} (y - \bar{y})}$ = x atrás

Sustituimos en R.d.R. de Y sobre X

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Sustituimos en R.d.R de X sobre Y

$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$ → Recta de Regresión de Y sobre X

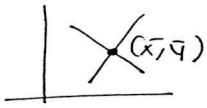
$x - \bar{x} = r \frac{\sigma_x}{\sigma_y}(y - \bar{y})$ → Recta de Regresión de X sobre Y

Entonces ambas rectas pasan por el punto $(\bar{x}, \bar{y})$

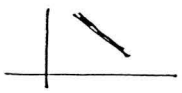
Las dos rectas serán iguales si $r \frac{\sigma_y}{\sigma_x} = r \frac{\sigma_x}{\sigma_y} = \frac{1}{r} \frac{\sigma_y}{\sigma_x} \Rightarrow$

$$\Rightarrow \boxed{r^2 = 1} = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right)^2$$

Si $r^2 \neq 1$ solo tienen en común el punto $(\bar{x}, \bar{y})$



Si $r^2 = 1$ son =



NOTA: Los ajustes de una hipérbola equilátera, una función potencial o exponencial se reducen fácil/ el ajuste de una recta.

- Hipérbola: $y = a + \frac{b}{x}$
  tomamos $z = \frac{1}{x}$ $\Rightarrow \boxed{y = a + bz}$
- Función potencial: $y = ax^b$
  tomamos logaritmos. $\ln y = \ln a + b \ln x$
  $\boxed{y' = a' + bx'}$
- Función exponencial: $y = ab^x$
  tomamos logaritmos $\ln y = \ln a + x \ln b$
  $\boxed{y' = a' + b'x}$

CORRELACIÓN. - Correlación

- Se ocupa del grado de asociación entre las variables, este grado de asociación nos indicará en q. medida la expresión encontrada explica una variable en función de otra

- Varianza residual: el método de mínimos cuadrados toma como medida del error q. se comete cuando ajustamos una función la suma de los residuos al cuadrado.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

Llamaremos varianza residual a la cantidad $\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \sigma_{ry}^2$

Se usa para medir la bondad del ajuste

~~Determinación~~ - Coeficiente de Determinación

¿A partir de q valores la varianza residual es suficiente/pequeña o grande como para medir la bondad del ajuste?

Para responder a esta pregunta definimos el coeficiente de determinación (R^2).

$$R^2 = \frac{\sigma_y^2 - \sigma_{xy}^2}{\sigma_y^2} = 1 - \frac{\sigma_{xy}^2}{\sigma_y^2} \qquad 0 \leq R^2 \leq 1$$

Si $R^2 = 0 \longrightarrow \frac{\sigma_{xy}^2}{\sigma_y^2} = 1$ ó $\sigma_{xy}^2 = \sigma_y^2 \Rightarrow$ en este caso el modelo no explica nada de "y" a partir de "x". → → ajuste es el peor.

Si $R^2 = 1$ $\frac{\sigma_{xy}^2}{\sigma_y^2} = 0$ $\sigma_{xy}^2 = 0 \Longrightarrow$ Todos los residuos son nulos. → → el ajuste es perfecto.

Para valores intermedios entre 0 y 1 → Según estén + próximos a:
- 0 → indicará un peor ajuste
- 1 → indicará un ajuste perfecto.

- <u>Bondad del ajuste de la recta.</u>

$y = a + bx$
$x = a' + b'y$

Para ambas R^2 coincide con el coeficiente de correlación lineal (r^2) y además tiene una interpretación particular:

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \frac{\not{A}}{\not{n}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \underbrace{\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{0}$$

- <u>Varianza explicada por la regresión:</u>

$$\sigma_{\hat{y}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Nos indica en q. medida q.da explicada la variable dependiente mediante el modelo estimado.

$\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ porque: en el caso de la regresión lineal, tenemos:

$$\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i \quad \text{pero}$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - a - bx_i) = \underbrace{\sum_{i=1}^{n} y_i - an - b\sum_{i=1}^{n} x_i}_{} = 0$$

es la 1ª ecuación normal de la recta. y por tanto es = a 0

$$\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} e_i (a + bx_i) = a \sum_{i=1}^{n} e_i + b \sum_{i=1}^{n} e_i x_i =$$

$$= b \underbrace{\left(\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} x_i^2 \right)}_{\text{es la 2ª ecuación normal de la recta y x tanto es } = 0} = 0$$

Por lo tanto:

$$\sum_{i=1}^{n} e_i (\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i \hat{y}_i - \bar{y} \sum_{i=1}^{n} e_i = 0$$

$$\sigma_y^2 = \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\sigma_{ey}^2} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\sigma_{ry}^2} + \underbrace{\frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{e_i = 0}$$

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sigma_{ry}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

$$\boxed{\sigma_y^2 = \sigma_{ry}^2 + \sigma_{ey}^2}$$

↓ residual — ↓ explicada por la regresión.

$$R^2 = \frac{\sigma_y^2 - \sigma_{ry}^2}{\sigma_y^2} = \frac{\sigma_{ey}^2}{\sigma_y^2}$$ considerando el ajuste $\boxed{y = a + bx}$

El coeficiente de determinación se interpreta como la proporción de la varianza de "y" q viene explicada x la regresión lineal.

En el caso de la recta de tendencias: $R^2 = r^2$ lo q. nos da una mejor interpretación de r^2.

Para ello vemos q $\sigma_y^2 = \sigma_{ry}^2 + \sigma_{ey}^2$

$$\sigma_{ry}^2 = \sigma_y^2 - \sigma_{ey}^2 = \sigma_y^2 \left(1 - \frac{\sigma_{ey}^2}{\sigma_y^2} \right) \Longrightarrow \text{continúa}$$

continuación $\Longrightarrow$ $\sigma_y^2(1-r^2)$ (x.q. $R^2=r^2$)

x.q. $R^2=r^2=1-\frac{\sigma_{ry}}{\sigma_y^2}=1-\frac{\sigma_y^2(1-r^2)}{\sigma_y^2}$

Análogo si se tratara ahora de la recta $x=a'+b'y$ regresión de x sobre y obtenemos:

$\sigma_{rx}^2=\sigma_x^2(1-r^2)$

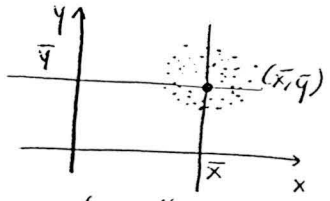
y tb. para esta recta (prueba fácil comprobarse q. $R^2=r^2$

Como $R^2=r^2$ → Posibles interpretaciones del coeficiente de correlación (r)

de lo anterior es inmediato q. $0\le r^2\le 1$ x.q. $0\le R^2\le 1$

por tanto $-1\le r\le 1$ y la interpretación de r tiene relación con la bondad del ajuste de la recta de mínimos cuadrados o de regresión.

*) Si $r=0$, entonces $\sigma_{xy}=0$ Es decir las dos rectas de regresión se reducen a $\begin{cases} y=\bar{y} \\ x=\bar{x} \end{cases}$ por tanto las dos rectas son paralelas a los ejes de coordenadas.



Si $r=0 \Rightarrow \sigma_{xy}=0 \Longrightarrow$ x e y independientes.

"x" e "y" no tienen relación entre ellas. pero pueden estar estrechamente ligadas según otro tipo lineal de función

*) Si $r=1$, entonces $\sigma_{ry}^2=0$

$r^2=R^2=1-\frac{\sigma_{ry}^2}{\sigma_y^2}=1 \Longrightarrow \frac{\sigma_{ry}^2}{\sigma_y^2}=0 \Longrightarrow \sigma_{ry}^2=0$

$\sigma_{ry}^2=\frac{1}{n}\sum_{i=1}^{n} e_i^2=0 \Longrightarrow e_i=0$ No hay residuos, los errores son nulos; y tanto las rectas de regresión pasan x todos puntos de la nube. y así

las rectas de regresión coinciden, entonces hay dependencia lineal máxima entre las variables en sentido positivo $r = 1 > 0$

*) Si $r = -1$, entonces $\sigma_{rY}^2 = 0$ y tenemos la misma interpretación q. para $r = 1$ sólo q. ahora hay dependencia lineal máxima en sentido positivo.

*) Si $0 < r < 1$ La correlación lineal será > cuanto + se aproxima a 1. → → correlación positiva

*) Si $-1 < r < 0$ La correlación. lineal será > cuanto + se aproxima a -1 → → correlación negativa o inversa.

— AJUSTE POR MÍNIMOS CUADRADOS DE LA PARÁBOLA:

Este es un caso de la regresión no lineal los otros casos son de hipérbola equilatera, función potencial y exponencial q. se reducen al ajuste de la recta mediante una transformación adecuada. En este ajuste se consideran las variables discretas aunq. el resultado puede aplicarse a variables continuas una vez expresado en términos de momentos.

Sea $\hat{y}_i = a + bx_i + cx_i^2$, el estimador de y_i y sea $e_i = y_i - \hat{y}_i$ el error q. se comete en la estimación. El método de mínimos cuadrados nos conduce a la parábola q. hace mínima la función $S(a,b,c)$ con: $S(a,b,c) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$

$$= \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

de la siguiente manera:

$$\left.\begin{aligned} \frac{\partial S}{\partial a} &= \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-1) = 0 \\ \frac{\partial S}{\partial b} &= \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-x_i) = 0 \\ \frac{\partial S}{\partial c} &= \sum_{i=1}^{n} 2(y_i - a - bx_i - cx_i^2)(-x_i^2) = 0 \end{aligned}\right\} \leftarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \end{cases}$$

dividiendo por n obtenemos las anteriores ecuaciones expresadas en momentos.

$$\begin{cases} a_{01} = a + b a_{10} + c a_{20} \\ a_{11} = a a_{10} + b a_{20} + c a_{30} \\ a_{21} = a a_{20} + b a_{30} + c a_{40} \end{cases}$$

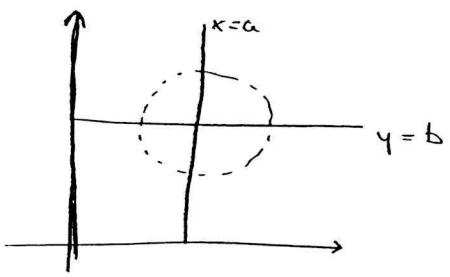
es válido para variables cualesquiera

Resolviendo el sistema se obtienen los coeficientes de la parábola de regresión de mínimos cuadrados.

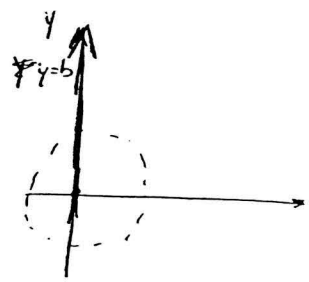
NOTA: Es clara la importancia del ajuste de funciones parabólicas a ej: en microeconomía es conocido q. la curva de costs marginales en función de volumen de producción es una parábola.

Ejemplo: x e y Relación no lineales

$(x-a)^2 + (y-b)^2 = 1$



$x^2 + y^2 = 1$ $a=b=0$



* Bondad del ajuste de la parábola:

El método de mínimos cuadrados para la parábola conduce a las ecuaciones:

$$\begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2) = 0 = \sum_{i=1}^{n} e_i \\ \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2) = 0 = \sum_{i=1}^{n} e_i x_i \\ \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2) = 0 = \sum_{i=1}^{n} e_i^2 x_i \end{cases}$$

Y también:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} e_i e_i = \sum_{i=1}^{n} e_i (y_i - a - bx_i - cx_i^2) = \sum_{i=1}^{n} e_i y_i - \underbrace{a \sum_{i=1}^{n} e_i}_{=0} - \underbrace{b \sum_{i=1}^{n} e_i x_i}_{=0} - \underbrace{c \sum_{i=1}^{n} e_i x_i^2}_{=0}$$

La varianza residual para este ajuste de la parábola es:

$$\sigma_{ry}^2 = \frac{1}{n} \sum e_i^2 = \frac{1}{n} \sum_{i=1}^{n} e_i y_i$$

$$\boxed{\sigma_{ry}^2} = \frac{1}{n} \left(\sum_{i=1}^{n} y_i (y_i - a - bx_i - cx_i^2) \right) = \frac{1}{n} \sum y_i^2 - \frac{a}{n} \sum_{i=1}^{n} y_i - \frac{b}{n} \sum_{i=1}^{n} y_i x_i -$$

$$- \frac{c}{n} \sum_{i=1}^{n} y_i x_i^2 = \boxed{a_{02} - a a_{01} - b a_{11} - c a_{21}}$$

Se completa este estudio con el cálculo del

Coeficiente de determinación.

$$R_p^2 = 1 - \frac{a_{02} - a a_{01} - b a_{11} - c a_{21}}{m_{02}} = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$$

NOTAS:

* Si ajustamos una parábola de la forma: $x = \hat{a} + \hat{b}y + \hat{c}y^2$ para expresar el comportamiento de x en función de y. La varianza residual y el coeficiente de determinación serán:

$$\sigma_{rx}^2 = a_{20} - \hat{a}a_{10} - \hat{b}a_{11} - \hat{c}a_{12}$$

$$R_p^2 = 1 - \frac{a_{20} - \hat{a}a_{10} - \hat{b}a_{11} - \hat{c}a_{12}}{n_{20}}$$

* Debido a q. la recta es un caso particular de parábola degenerada xa $c = 0$ se obtendrán siempre mejores ajustes mediante parábolas q. usando funciones lineales siendo el coeficiente de determinación xa la parábola siempre > q el de la recta.

coeficiente de determinación xa la parábola

$$y = a + bx + cx^2$$
$$\vec{0}$$
$$y = a + bx$$

$R_p^2 > R^2$

↓

Coeficiente de determinación xa la parábola

*- Como sabemos en el caso de la recta el coeficiente de determinación r^2 indica q. parte de la varianza de la variable independiente es explicada x la regresión. Si calculamos est coeficiente sobre los ctas obtenidos después de transformar las funciones potencial, exponencial y la hipérbola equilátra en rectas no tendrán el mismo significado y no será comparable con valores obtenidos sobre la bondad del ajuste de una recta o de una parábola.

TEMA 2 : Esquema.

→ (x,y) → variable estadística bidimensional : representa 2 caracteres (x,y)

$$\left.\begin{array}{l} X: x_1, x_2, \ldots x_p \\ Y: y_1, y_2, \ldots y_q \end{array}\right\} (x,y) = (x_i, y_j)$$

n_{ij} del par (x_i, y_j) → Frecuencia absoluta : nº de individuos q. han presentado el valor x_i de X e y_j de Y.

f_{ij} → Frecuencia relativa

$$\boxed{f_{ij} = \frac{n_{ij}}{N}}$$

→ REPRESENTACIÓN GRÁFICA → Tabla de doble entrada.

X \ Y	$L_1 - L_2$ $y_1 = \frac{L_1+L_2}{2}$	$L_2 - L_3$ $y_2 = \frac{L_2+L_3}{2}$	$L_{q-1} - L_q$ $y_q = \frac{L_{q-1}+L_q}{2}$	$n_{i\bullet}$
x_1	n_{11}	n_{12}	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	n_{2q}	$n_{2\bullet}$
x_p	n_{p1}	n_{p2}	n_{pq}	$n_{p\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet q}$	N

$$\boxed{n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \rightarrow i \text{ fijo}}$$

$$\boxed{n_{\bullet j} = \sum_{i=1}^{p} n_{ij} \rightarrow j \text{ fijo}}$$

$$\boxed{N = \sum_{i=1}^{p} n_{i\bullet} = \sum_{i=1}^{q} n_{\bullet j}}$$

→ DISTRIBUCIONES MARGINALES.

* 1º y última columna de la tabla de doble entrada.

X	$n_{i\bullet}$	$f_{i\bullet}$
x_1	$n_{1\bullet}$	$f_{1\bullet} = \frac{n_{1\bullet}}{N}$
x_2	$n_{2\bullet}$	$f_{2\bullet} = \frac{n_{2\bullet}}{N}$
$\vdots$	$\vdots$	$\vdots$
x_p	$n_{p\bullet}$	$f_{p\bullet} = \frac{n_{p\bullet}}{N}$
	N	

frecuencia relativa marginal → $\boxed{f_{i\bullet} = \frac{n_{i\bullet}}{N}}$

* 1ª y última fila de la tabla de doble entrada

$l_{q-1} - l_q$	Y	$n_{\bullet j}$	$f_{\bullet j}$
$l_1 - l_2$	y_1	$n_{\bullet 1}$	$f_{\bullet 1} = \frac{n_{\bullet 1}}{N}$
$l_2 - l_3$	y_2	$n_{\bullet 2}$	$f_{\bullet 2} = \frac{n_{\bullet 2}}{N}$
⋮	⋮	⋮	⋮
$l_{q-1} - l_q$	y_q	$n_{\bullet q}$	$f_{\bullet q} = \frac{n_{\bullet q}}{N}$
		N	

frecuencia relativa marginal →

$$f_{\bullet j} = \frac{n_{\bullet j}}{N}$$

Media marginal

de X: $$\bar{x} = \frac{1}{N} \sum_{i=1}^{p} x_i n_{i\bullet} = \sum_{i=1}^{p} x_i f_{i\bullet}$$

de Y: $$\bar{y} = \frac{1}{N} \sum_{j=1}^{q} y_j n_{\bullet j} = \sum_{j=1}^{q} y_j f_{\bullet j}$$

Varianza marginal

de X: $$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^{p} (x_i - \bar{x})^2 n_{i\bullet} = \sum_{i=1}^{p} (x_i - \bar{x})^2 f_{i\bullet} = \left(\frac{1}{N} \sum_{i=1}^{p} x_i^2 n_{i\bullet} \right) - \bar{x}^2$$

de Y: $$\sigma_y^2 = \frac{1}{N} \sum_{j=1}^{q} (y_j - \bar{y})^2 n_{\bullet j} = \sum_{j=1}^{q} (y_j - \bar{y})^2 f_{\bullet j} = \left(\frac{1}{N} \sum_{j=1}^{q} y_j^2 n_{\bullet j} \right) - \bar{y}^2$$

→ DISTRIBUCIONES CONDICIONADAS.

- Son: distribuciones unidimensionales obtenidas a partir de las bidimensionales manteniendo fijo el valor de una de las variables y considerando los valores de la otra con sus respectivas frecuencias.
- Distribución condicionada de X dado q. $Y = y_j$ se obtiene a partir de la tabla de doble entrada tomando la 1ª columna de los valores de X y la j-ésima columna de las frecuencias absolutas.
- Distribución condicionada de Y dado q. $X = x_i$ se obtiene a partir de la tabla de doble entrada tomando la 1ª fila de los valores de Y y la i-ésima fila de las frecuencias absolutas.

- frecuencia relativa de la distribución Y/x_i (de Y condicionada por $x=x_i$) → → proporción de individuos q cumplen la condición $X=x_i$ y q han presentado el valor $Y=y_j$

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{\bullet j}}$$

- frecuencia relativa de la distribución X/y_j (de X condicionada por $y=y_j$) → → proporción de individuos q. cumplen la condición $Y=y_j$ y han presentado el valor $X=x_i$.

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{i\bullet}}$$

- Media condicionada de la distribución
  - de X/y_j: $\bar{x}_j = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} x_i n_{ij} = \sum_{i=1}^{p} x_i f_{i/j}$
  - de Y/x_i: $\bar{y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} y_j n_{ij} = \sum_{j=1}^{q} y_j f_{j/i}$

- Varianza condicionada
  - de X/y_j: $\sigma_j^2(x) = \sigma_{x/y_j}^2 = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} (x_i - \bar{x}_j)^2 n_{ij} = \sum_{i=1}^{p} (x_i - \bar{x}_j)^2 f_{i/j}$
  - de Y/x_i: $\sigma_i^2(y) = \sigma_{y/x_i}^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} (y_j - \bar{y}_i)^2 n_{ij} = \sum_{j=1}^{q} (y_j - \bar{y}_i)^2 f_{j/i}$

→ RELACION ENTRE DISTRIBUCIONES MARGINALES y CONDICIONADAS.

$$f_{ij} = f_{i/j} \cdot f_{\bullet j} = f_{j/i} \cdot f_{i\bullet}$$

Relación entre media marginal y media condicionada $\Longrightarrow$

$$\bar{x} = \sum_{j=1}^{p} \bar{x}_j f_{\bullet j} = \frac{1}{N} \sum_{j=1}^{p} \bar{x}_j n_{\bullet j}$$

$$\bar{y} = \sum_{i=1}^{q} \bar{y}_i f_{i\bullet} = \frac{1}{N} \sum_{i=1}^{q} \bar{y}_i n_{i\bullet}$$

Relación entre varianza marginal y varianza condicionada $\Rightarrow$

$$\sigma_x^2 = \sum_{j=1}^{q} \sigma_j^2(x) f_{\bullet j} + \sum_{i=1}^{q} (\bar{x}_j - \bar{x})^2 f_{\bullet j}$$

$$\sigma_y^2 = \sum_{i=1}^{p} \sigma^2(y) f_{i\bullet} + \sum_{i=1}^{p} (\bar{y}_i - \bar{y})^2 f_{i\bullet}$$

→ INDEPENDENCIA ESTADÍSTICA.

- Dos variables "x" e "y" son estadísticamente independientes cuando la variación de una de ellas no influye en la otra.
- Dos variables estadísticas son independientes si:

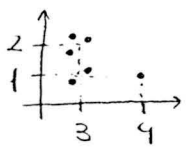
$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \qquad f_{i/j} = f_{i\bullet} \qquad f_{j/i} = f_{\bullet j}$$

- Dependencia estadística → conocemos los valores parcial/
- Dependencia funcional → conocemos todos los valores de la otra variable mediante una función.

→ REPRESENTACIONES GRÁFICAS.

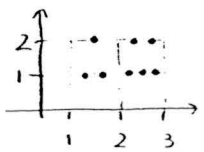
Nube de puntos.

- 2 variables discretas



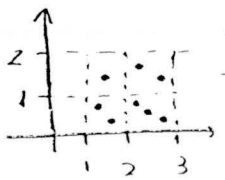
X \ Y	1	2
3	2	3
4	1	0

- 1 variable discreta y otra continua



X \ Y	1	2
(1,2]	2	1
(2,3]	3	2

- 2 variables continuas

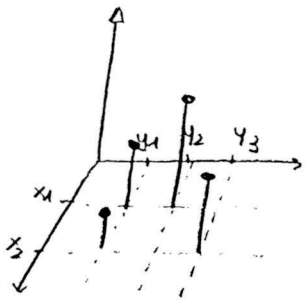


X \ Y	(0,1]	(1,2]
(1,2]	2	1
(2,3]	3	2

Estereograma ó Escalograma

Variables discretas: consiste en representar sobre el plano (x,y) los valores (x_i, y_j) y levantar sobre cada uno de estos puntos un segmento de longitud = a la frecuencia con q. se ha observado dicho par de valores.

x \ y	y_1	y_2	y_3
x_1	3	4	0
x_2	1'5	0	2

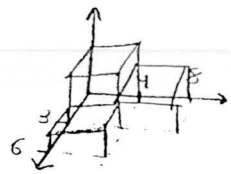


Variables continuas: se marcan los intervalos sobre los ejes "x" e "y" dando lugar a rectángulos sobre el plano (x,y) y sobre cada uno se construye un paralelepípedo de volumen = ó proporcional a la frecuencia q. se quiere representar.

- $h_{ij} \Rightarrow$ altura del paralelepípedo
- $V_{ij} \Rightarrow$ volumen del paralelepípedo

x \ y	0-4	4-8
0-3	5	3
3-6	4	0

$$V_{ij} = n_{ij} = (L_i - L_{i-1})(L_j - L_{j-1})\, h_{ij}$$



→

MOMENTOS DE UNA DISTRIBUCIÓN BIDIMENSIONAL

momento no centrado de órdenes r y s.

Variables continuas

$$a_{rs} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i^r y_i^s f_{ij} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} x_i^r y_j^s n_{ij}$$

Variables discretas.

$$a_{rs} = \frac{1}{N}\sum_{i=1}^{n} x_i^r y_i^s$$

Casos particulares

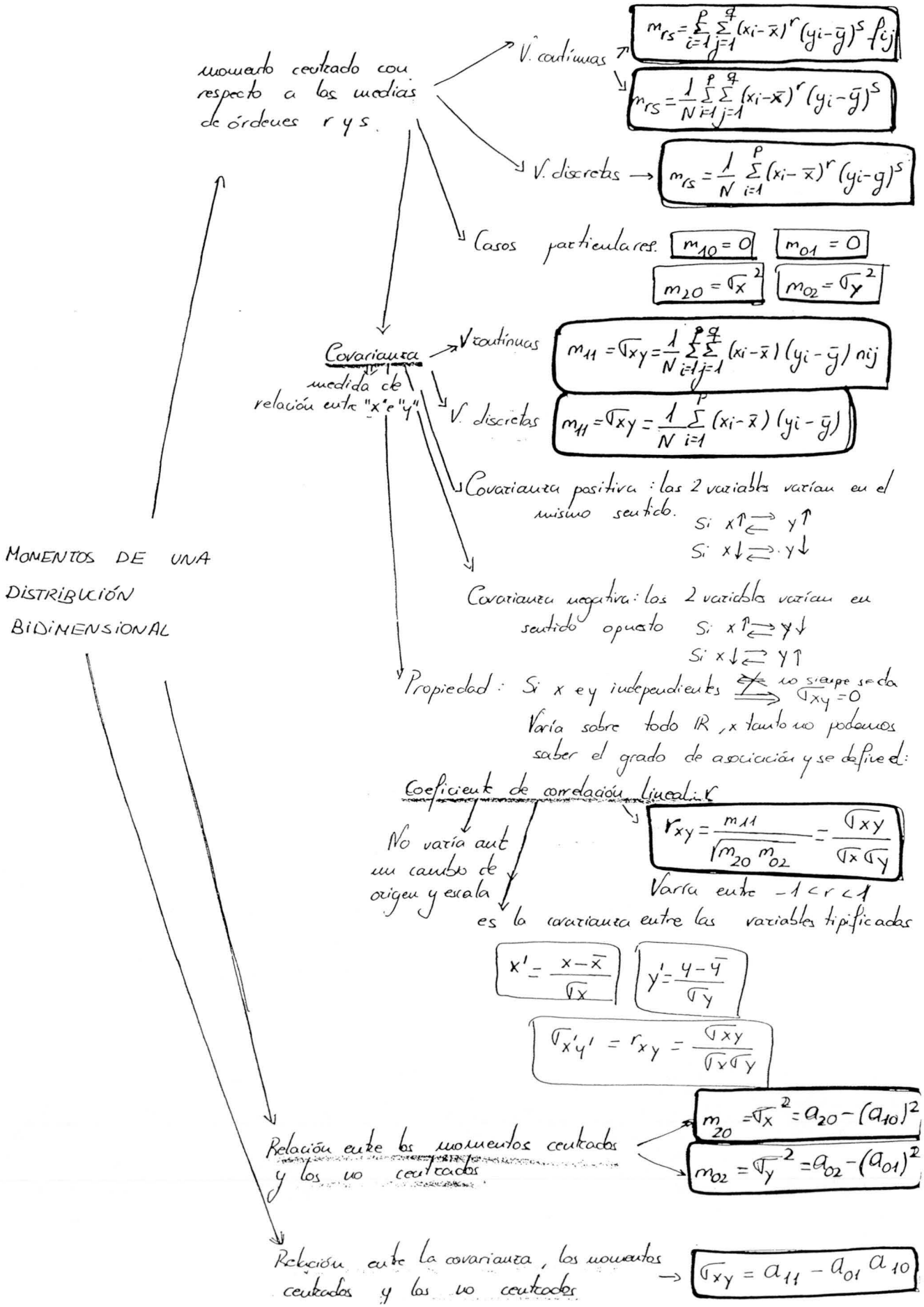
$a_{10} = \bar{x}$ $\quad a_{01} = \bar{y}$ $\quad a_{20} = a_2(x)$ $\quad a_{02} = a_2(y)$

a_{11}

V. continuas $\Rightarrow$ $a_{11} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} x_i y_j n_{ij}$

V. discretas $\Rightarrow$ $a_{11} = \frac{1}{N}\sum_{i=1}^{n} x_i y_i$

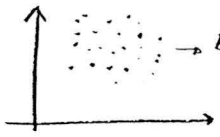
momento centrado con las medias de órdenes r y s. continúa x otras


MOMENTOS DE UNA DISTRIBUCIÓN BIDIMENSIONAL
momento centrado con respecto a las medias de órdenes r y s.
V. continuas
$m_{rs}=\sum_{i=1}^{p}\sum_{j=1}^{q}(x_i-\bar{x})^r(y_i-\bar{y})^s f_{ij}$
$m_{rs}=\frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q}(x_i-\bar{x})^r(y_i-\bar{y})^s$
V. discretas
$m_{rs}=\frac{1}{N}\sum_{i=1}^{p}(x_i-\bar{x})^r(y_i-\bar{y})^s$
Casos particulares:
$m_{10}=0$
$m_{01}=0$
$m_{20}=\sigma_x^2$
$m_{02}=\sigma_y^2$
Covariancia
medida de relación entre "x" e "y"
V. continuas
$m_{11}=\sigma_{xy}=\frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q}(x_i-\bar{x})(y_i-\bar{y})n_{ij}$
V. discretas
$m_{11}=\sigma_{xy}=\frac{1}{N}\sum_{i=1}^{p}(x_i-\bar{x})(y_i-\bar{y})$
Covariancia positiva: las 2 variables varían en el mismo sentido.
Si $x\uparrow \rightleftarrows y\uparrow$
Si $x\downarrow \rightleftarrows y\downarrow$
Covariancia negativa: las 2 variable varían en sentido opuesto
Si $x\uparrow \rightleftarrows y\downarrow$
Si $x\downarrow \rightleftarrows y\uparrow$
Propiedad: Si x e y independientes $\not\Longrightarrow$ no siempre se da $\sigma_{xy}=0$
Varía sobre todo $\mathbb{R}$, x tanto no podemos saber el grado de asociación y se define el:
Coeficiente de correlación lineal: r
$r_{xy}=\frac{m_{11}}{\sqrt{m_{20}\,m_{02}}}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$
Varía entre $-1<r<1$
No varía ante un cambio de origen y escala
es la covariancia entre las variables tipificadas
$x'=\frac{x-\bar{x}}{\sigma_x}$
$y'=\frac{y-\bar{y}}{\sigma_y}$
$\sigma_{x'y'}=r_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$
Relación entre los momentos centrados y los no centrados
$m_{20}=\sigma_x^2=a_{20}-(a_{10})^2$
$m_{02}=\sigma_y^2=a_{02}-(a_{01})^2$
Relación entre la covariancia, los momentos centrados y los no centrados
$\sigma_{xy}=a_{11}-a_{01}a_{10}$

→ REGRESIÓN y CORRELACIÓN SIMPLE.

→ La Regresión pretende poner unas variables en función de otras mediante una ley.

- Variable q. se quiere predecir → dependiente ó endógena.
- Variable q. se utiliza xa la predicción → independiente ó exógena.
  - Sólo utilizamos una variable independiente → Regresión o correlación simple
  - Si utilizamos + de una variable independiente → Regresión o correlación múltiple



→ Es imposible encontrar una curva q. pase x todos los puntos Hay q buscar la curva q. pase x el nº máximo de puntos

→ Del diagrama de dispersión buscamos la curva q. aunq. no pase x todos los puntos de la nube al menos esté lo + próxima.

- criterio + adecuado: "Ajuste x mínimos cuadrados"

→ Línea de regresión Y sobre X: es la representación gráfica de $\{(x_i, \bar{y}_i);\ i=1,\dots p\}$ y uniéndolos x una línea.

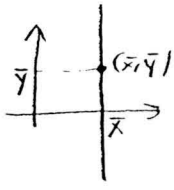
→ Línea de regresión de X sobre Y: es la representación gráfica de $\{(\bar{x}_j, y_j);\ j=1,\dots q\}$

→ Forma q. adopta la curva de regresión según el grado de dependencia entre "x" e "y":

1.-) Dependencia funcional; "x" e "y" dependientes funcional!

→ las distribuciones condicionadas X/y_j serán = entre sí e= a la media marginal correspondiente. $x_1 = x_2 = x_3 = \dots x_q = \bar{x}$ $\bar{x}_j = \bar{x}$

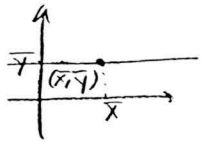
- Conjunto $\{(\bar{x}_j, y_j);\ j=1\dots q\}$ tiene la misma abcisa



⇐ Curva de regresión de "x" e "y" es paralela al eje OY pasando x $(\bar{x}, \bar{y})$

→ las distribuciones condicionadas Y/x_i : $\bar{y} = \bar{y}_i$ $y_1 = y_2 = \dots y_p = \bar{y}$

- Conjunto $\{(x_i, \bar{y}_i);\ i=1\dots p\}$ todos los puntos tienen la misma ordenada.



⇐ La curva de regresión de Y sobre X es paralela al eje OX pasando x $(\bar{x}, \bar{y})$

2º-) Dependencia funcional

→ "y" depende funcional/ de "x" si a cada valor de x_i de "x" le corresponde un único valor y_j de "y":

- $Y/_{X=x_i}$ tiene un solo valor $\Rightarrow \bar{y}_i = y_j \Rightarrow$ conjunto q. queremos representar $\{(x_i, \bar{y}_i)\} = \{(x_i, y_i)\} \Rightarrow y = f(x)$

→ "x" depende funcional/ de "y" si a cada valor y_j de "y" le corresponde un único valor posible de "x":

- conjunto q. queremos representar $\{(\bar{x}_j, y_j)\} = \{(x_i, y_j)\}$

$\bar{x}_j = x_i \qquad X = g(y)$

3º-) Caso general o intermedio:

→ Supongamos: - Para predecir la variable "y" conocemos otra variable relacionada con ella "x" $\Rightarrow y = f(x)$; f no la conocemos.

- Consideramos $n \equiv N$ observaciones totales.

→ Utilizando la función $x = x_i$

- estimamos un valor $\hat{y}_i = f(x_i)$

estimación del valor real y_i.

- $\hat{y}_i \neq y_i \Rightarrow$ Se comete un error en la estimación.

- Error = residuo = $e_i = y_i - \hat{y}_i$

- criterio para un buen ajuste: minimizar la suma de todos los residuos.

$$\text{minimizar} \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{j=1}^{n} (y_i - f(x_i))^2$$

→ Método de mínimos cuadrados.

- Considera q. la función q. mejor se ajusta a unos datos observados es aq.lla q minimiza la suma de los cuadrados de los residuos

continúa →

- las funciones f q. se ajustan con + frecuencia son:
  - f una recta: $y = f(x) = a + bx$
  - f una parábola: $y = a + bx + cx^2$
  - f una hipérbola equilátera: $y = a + b\frac{1}{x}$
  - f una función potencial: $y = ax^b$
  - f una función exponencial: $y = a \cdot b^x$

- **RECTA DE REGRESIÓN**

  - Si la función f q. ajustamos es una recta → → regresión lineal
  - $S(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$: hay q. buscar los a y b q. hagan mínima $S(a,b)$
  - Para q. haya mínimo en un punto → las derivadas parciales de 1er orden se tienen q anular en ese punto.

$$S(a,b) = \sum_{i=1}^{n} (y_i - a - bx)^2 \qquad \left.\begin{array}{l} \frac{\partial S(a,b)}{\partial a} = 0 \\ \frac{\partial S(a,b)}{\partial b} = 0 \end{array}\right\} \cdots \left.\begin{array}{l} \bar{y} = a + b\bar{x} \\ a_{11} = a\bar{x} + b a_{20} \end{array}\right\} \cdots$$

$$\Rightarrow \left.\begin{array}{l} a_{01} = a + b a_{10} \\ a_{11} = a \cdot a_{10} + b a_{20} \end{array}\right\} \begin{array}{l} \rightarrow \boxed{b = \frac{a_{11} - a_{10} a_{01}}{a_{20} - a_{10}^2} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{m_{11}}{m_{20}}} \\ \rightarrow \boxed{a = a_{01} - \frac{m_{11}}{m_{20}} a_{10} = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2} \bar{x}} \end{array}$$

Ambas pasan x $(\bar{x}, \bar{y})$

Serán = si $r\frac{\sigma_y}{\sigma_x} = r\frac{\sigma_x}{\sigma_y} \Rightarrow$

$\Rightarrow \boxed{r^2 = 1}$

$r^2 \neq 1$ sólo tienen en común el punto $(\bar{x}, \bar{y})$ | $\times(\bar{x}, \bar{y})$

$r^2 = 1$ ambas rectas son =

- Recta de regresión de Y sobre X

$y = a + bx \Rightarrow$ sustituyendo a y b $\Rightarrow \boxed{y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})}$

$$\boxed{y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})}$$

- Recta de regresión de X sobre Y

$$\boxed{x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})} \Longrightarrow \boxed{x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})}$$

- los ajustes de una hipérbola equilátera, una función potencial o exponencial se reducen al ajuste de una recta.

  • Hipérbola: $y = a + \frac{b}{x}$ $\quad z = \frac{1}{x} \Rightarrow y = a + bz$

  • Función potencial: $y = ax^b$ $\quad \ln y = \ln a + b \ln x$

  $y' = a' + bx'$

  • Función exponencial: $y = ab^x$ $\quad \ln y = \ln a + x \ln b$

  $y' = a' + b'x$

* CORRELACIÓN: Nos indicará en q medida la expresión encontrada explica una variable en función de otra.

- Varianza residual: Se usa xa medir la bondad del ajuste.

$$\sigma_{ry}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

- Coeficiente de determinación:

$$R^2 = \frac{\sigma_y^2 - \sigma_{xy}^2}{\sigma_y^2} = 1 - \frac{\sigma_{xy}^2}{\sigma_y^2}$$

$R^2 = 0 \longrightarrow \frac{\sigma_{xy}^2}{\sigma_y^2} = 1 \longrightarrow$ el modelo no explica nada de "y" a partir de "x" $\longrightarrow$ es el peor ajuste.

$R^2 = 1 \longrightarrow \frac{\sigma_{xy}^2}{\sigma_y^2} = 0 \longrightarrow$ todos los residuos son nulos $\longrightarrow$ el ajuste es perfecto.

Para valores próximos a: $\begin{cases} 0 \longrightarrow \text{peor ajuste} \\ 1 \longrightarrow \text{mejor ajuste} \end{cases}$

- Bondad del ajuste de la recta.

$R^2 = r^2$

$y = a + bx$

$x = a' + b'y$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$\overset{\text{"}}{0}$

- Variauza explicada x la regresión: nos indica en q. medida queda explicada la variable dependiente mediante el modelo estimado.

$$\sigma_{ey}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \sigma_y^2 = \sigma_{ry}^2 + \sigma_{ey}^2$$

$$R^2 = \frac{\sigma_y^2 - \sigma_{ry}^2}{\sigma_y^2} = \frac{\sigma_{ey}^2}{\sigma_y^2} \qquad \sigma_{ry}^2 = \sigma_y^2(1-r^2)$$

considerando el ajuste $y = a + bx$

$\sigma_{rx}^2 = \sigma_x^2(1-r^2)$ considerando el ajuste $x = a' + b'y$

- Posibles interpretaciones del coeficiente de correlación:

$0 \leq r^2 \leq 1$; $0 \leq R^2 \leq 1$; $-1 \leq r \leq 1$.

A-) Si $r=0 \Rightarrow \sigma_{xy}=0$: ⟶ las dos rectas de regresión se reducen a: $x = \bar{x}$, $y = \bar{y}$

- las dos rectas son paralelas al eje de coordenadas.

⟶ x e y son independientes: no tienen relación lineal.

B-) Si $r=1 \Rightarrow \sigma_{ry}^2 = 0 \Rightarrow \sigma_{ry}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = 0 \Rightarrow e_i = 0 \Rightarrow$ No hay residuos, los errores son nulos, y tanto las rectas de regresión pasan x todos los puntos de la nube.

- las dos rectas de regresión coinciden → hay dependencia lineal máxima entre las variables en sentido positivo $r = 1 > 0$

C-) Si $r = -1$ → hay dependencia lineal máxima entre las variables en sentido negativo.

D-) Si $0 < r < 1$ → correlación positiva → la correlación es > cuanto + se aproxime a 1

E-) Si $-1 < r < 0$ → correlación negativa → la correlación es > cuanto + se aproxima a -1.

– AJUSTE x MÍNIMOS CUADRADOS

- Es un caso de regresión no lineal
- Método de mínimos cuadrados ⟶ parábola q. hace mínima la función $S(a,b,c)$

$\hat{y}_i = a + bx_i + cx_i^2 \qquad e_i = y_i - \hat{y}_i$

$$S(a,b,c) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

$$\left.\begin{array}{l} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \\ \frac{\partial S}{\partial c} = 0 \end{array}\right\} \begin{array}{l} a_{01} = a + b a_{10} + c a_{20} \\ a_{11} = a a_{10} + b a_{20} + c a_{30} \\ a_{21} = a a_{20} + b a_{30} + c a_{40} \end{array}$$

- Bondad del ajuste de la parábola

$$\sigma_{ry}^2 = a_{02} - a a_{01} - b a_{11} - c a_{21}$$

- Coeficiente de determinación.

$$R_p^2 = 1 - \frac{a_{02} - a a_{01} - b a_{11} - c a_{21}}{m_{02}} = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$$

- Nota: Si ajustamos una parábola de la forma: $x = \hat{a} + \hat{b}y + \hat{c}y^2$

$$\sigma_{rx}^2 = a_{20} - \hat{a} a_{10} - \hat{b} a_{11} - \hat{c} a_{12}$$

$$R_p^2 = 1 - \frac{a_{20} - a a_{10} - b a_{11} - c a_{12}}{m_{20}} = 1 - \frac{\sigma_{rx}^2}{\sigma_x^2}$$

TEMA 2.

$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$ i-fijo

$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$ j-fijo

$N = \sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j}$

Frecuencias relativas marginales

$f_{i\bullet} = \frac{n_{i\bullet}}{N}$

$f_{\bullet j} = \frac{n_{\bullet j}}{N}$

Media marginal

de $X \Rightarrow \bar{x} = \frac{1}{N}\sum_{i=1}^{p} x_i \cdot n_{i\bullet} = \sum_{i=1}^{p} x_i \cdot f_{i\bullet}$

de $Y \Rightarrow \bar{y} = \frac{1}{N}\sum_{j=1}^{q} y_j \, n_{\bullet j} = \sum_{j=1}^{q} y_j \cdot f_{\bullet j}$

Varianza marginal

de $X \Rightarrow \sigma_x^2 = \sum_{i=1}^{p} (x_i - \bar{x})^2 \cdot f_{i\bullet}$

de $Y \Rightarrow \sigma_y^2 = \sum_{j=1}^{q} (y_j - \bar{y})^2 \cdot f_{\bullet j}$

Frecuencia relativa condicionada de la distribución Y/x_i $\Rightarrow f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{ij}}$

Frecuencia relativa condicionada de la distribución X/y_j $\Rightarrow f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{i\bullet}}$

Media condicionada xa la distribución

de $X/y_j \Rightarrow \bar{x}_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{p} x_i n_{ij} = \sum_{i=1}^{p} x_i \cdot f_{i/j}$

de $Y/x_i \Rightarrow \bar{y}_i = \frac{1}{n_{i\bullet}}\sum_{j=1}^{q} y_j n_{ij} = \sum_{j=1}^{q} y_j \cdot f_{j/i}$

Varianza condicionada

de $X/y_j \Rightarrow \sigma_j^2(x) = \sigma^2_{x/y_j} = \sum_{i=1}^{p} (x_i - \bar{x}_j) f_{i/j}$

de $Y/x_i \Rightarrow \sigma_i^2(y) = \sigma^2_{y/x_i} = \sum_{j=1}^{q} (y_j - \bar{y}_i)^2 f_{j/i}$

Relación entre distribuciones marginales y condicionadas $\Rightarrow$ $f_{ij} = f_{i/j} \cdot f_{\bullet j} = f_{j/i} \cdot f_{i\bullet}$

Relación entre media marginal y media condicionada $\Rightarrow$

$$\bar{x} = \sum_{j=1}^{q} \bar{x}_j \cdot f_{\bullet j} = \frac{1}{N} \sum_{j=1}^{q} \bar{x}_j \, n_{\bullet j}$$

$$\bar{y} = \sum_{i=1}^{p} \bar{y}_i \cdot f_{i\bullet} = \frac{1}{N} \sum_{i=1}^{p} \bar{y}_i \cdot n_{i\bullet}$$

Relación entre varianza marginal y varianza condicionada $\Rightarrow$

$$\sigma_x^2 = \sum_{j=1}^{q} \sigma_j^2(x) \cdot f_{\bullet j} + \sum_{j=1}^{q} (\bar{x}_j - \bar{x})^2 f_{\bullet j}$$

$$\sigma_y^2 = \sum_{i=1}^{p} \sigma_i^2(y) \cdot f_{i\bullet} + \sum_{i=1}^{p} (\bar{y}_i - \bar{y})^2 f_{i\bullet}$$

Dos variables estadísticas son independientes si:

- $f_{ij} = f_{i\bullet} \cdot f_{\bullet j}$
- $f_{i/j} = f_{i\bullet}$
- $f_{j/i} = f_{\bullet j}$

momento no centrado de orden r y s

- V. c. $\Rightarrow$ $a_{rs} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_i^r y_j^s f_{ij} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{q} x_i^r y_j^s n_{ij}$
- * V. d. $\Rightarrow$ $a_{rs} = \frac{1}{N} \sum_{i=1}^{n} x_i^r y_i^s$
- * Casos particulares $\Rightarrow$ $a_{10} = \bar{x}$; $a_{01} = \bar{y}$; $a_{20} = a_2(x)$; $a_{02} = a_2(y)$

a_{11}:

- V. c. $\rightarrow$ $a_{11} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{q} x_i y_j n_{ij}$
- V. d. $\Rightarrow$ $a_{11} = \frac{1}{N} \sum_{i=1}^{n} x_i y_i$

momento centrado con respecto a las medias de orden r y s

- V. c. $\Rightarrow$ $m_{rs} = \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})^r (y_i - \bar{y})^s f_{ij}$
- $\Rightarrow$ $m_{rs} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})^r (y_i - \bar{y})^s n_{ij}$
- * V. d $\Rightarrow$ $m_{rs} = \frac{1}{N} \sum_{i=1}^{p} (x_i - \bar{x})^r (y_i - \bar{y})^s$
- * Casos particulares $\Rightarrow$ $m_{10} = 0$; $m_{01} = 0$; $m_{20} = \sigma_x^2$; $m_{02} = \sigma_y^2$

* Covarianza

- V. c. $\Rightarrow$ $\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})(y_i - \bar{y}) n_{ij}$
- V. d. $\Rightarrow$ $\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{p} (x_i - \bar{x})(y_i - \bar{y})$

* $r_{xy} = \frac{m_{11}}{\sqrt{m_{02} \, m_{20}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Relación entre los momentos centrados y los no centrados

$$m_{20} = \sigma_x^2 = a_{20} - (a_{10})^2$$

V.d. $= \frac{1}{N}\Sigma x_i^2 - \bar{x}^2$ (2.)

V.c. $= \frac{1}{N}\Sigma\Sigma x_i^2 n_{ij} - \bar{x}^2$

$$m_{02} = \sigma_y^2 = a_{02} - (a_{01})^2$$

V.d. $= \frac{1}{N}\Sigma y_j^2 - \bar{y}^2$

V.c. $= \frac{1}{N}\Sigma\Sigma y_j^2 n_{ij} - \bar{y}^2$

Relación entre la covariauza, los momentos centrados y los no centrados

$$\sigma_{xy} = a_{11} - a_{01}a_{10}$$

V.c. $= \frac{1}{N}\Sigma n_{ij} x_i y_j - \bar{x}\,\bar{y}$

V.d. $= \frac{1}{N}\Sigma x_i y_j - \bar{y}\,\bar{x}$

$$a = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\,\bar{x}$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

Recta de regresión de Y sobre X $\Rightarrow$ $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

Recta de regresión de X sobre Y $\Rightarrow$ $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Cuando vamos la doble [illegible] doble [illegible]

Hipérbola: $y = a + \frac{b}{x}$; $z = \frac{1}{x}$ $\Rightarrow$ $y = a + bz$

$r = \sqrt{bb'}$

$\sigma_{xy} = \frac{1}{n}x_i\Sigma n_{ij} - \bar{x}\bar{y}$

Función potencial: $y = ax^b$ $\quad \ln y = \ln a + b\ln x$

$y' = a' + bx'$

Función exponencial: $y = ab^x$ $\quad \ln y = \ln a + x\ln b$

$y' = a' + b'x$

Varianza residual: $\sigma_{ry}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_j - \hat{y}_j)^2$

Coeficiente de determinación: $R^2 = 1 - \frac{\sigma_{xy}^2}{\sigma_y^2}$

$R^2 = 0 \rightarrow$ peor ajuste

$R^2 = 1 \rightarrow$ ajuste perfecto.

Bondad del ajuste de la recta: $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y})(\hat{y}_i - y_i)$

Varianza explicada x la regresión: $\sigma_{ey}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ $\quad \sigma_y^2 = \sigma_{ry}^2 + \sigma_{ey}^2$

$R^2 = \frac{\sigma_{ey}^2}{\sigma_y}$ $\quad \sigma_{ry}^2 = \sigma_y^2(1-r^2) \longrightarrow y = a + bx$

$\sigma_{rx}^2 = \sigma_x^2(1-r^2) \longrightarrow x = a' + b'y$

Bondad del ajuste de la parábola: $\sigma_{ry}^2 = a_{02} - a a_{01} - b a_{11} - c a_{21}$

Coeficiente de determinación: $R_p^2 = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$

$x = \hat{a} + \hat{b}y + \hat{c}y^2$ $\quad \sigma_{rx}^2 = a_{20} - a a_{10} - b a_{11} - c a_{12}$

$R_p^2 = 1 - \frac{\sigma_{rx}^2}{\sigma_x^2}$

Anexo V

Cuestionario utilizado en la investigación

UNIVERSIDAD DE GRANADA
DEPARTAMENTO DE DIDÁCTICA DE LAS MATEMÁTICAS

PRUEBA DE COMPRENSIÓN DE LAS NOCIONES DE CORRELACIÓN Y REGRESIÓN

Alumna/o _____ **Curso** _____

Edad _____ **Titulación** _____ **Fecha** _____

CUESTIONARIO

INSTRUCCIONES

Estamos interesados en mejorar la enseñanza de algunos temas de Estadística. Para ello necesitamos conocer como responden los alumnos a estas preguntas. Por favor, lea con atención los enunciados y responda todas las preguntas que se proponen.

PREGUNTAS PRELIMINARES

A. ¿Qué forma de acceso a la universidad ha tenido?

C.O.U. _____ ¿Qué modalidad? _____

F.P. _____ ¿Qué especialidad? _____

Otras formas de acceso (indicarla) _____

B. ¿Ha estudiado en cursos anteriores nociones de Estadística?

NO _____

SI _____ ¿En qué cursos? _____

C. ¿Qué interés tiene la asignatura de Estadística para su formación en la carrera que está estudiando? Marcar con una cruz (X)

a. Mucho
b. Bastante
c. Suficiente
d. Poco
e. Muy poco

D. ¿Qué interés tiene el tema de Correlación y Regresión para su formación en la carrera que está estudiando? Marcar con una cruz (X)

a. Mucho
b. Bastante
c. Suficiente
d. Poco
e. Muy poco

A continuación presentamos una serie de preguntas que ofrecen varias respuestas. Deberá poner una cruz (X) *en* todas las contestaciones *que sean correctas.*

Por ejemplo:

¿Qué es un triángulo rectángulo?

a. Un triángulo cuyos ángulos son todos agudos
X b. Un triángulo que tiene un ángulo recto
X c. Un triángulo en el que se verifica $a^2 = b^2 + c^2$, donde a, b y c representan las longitudes de los lados del mismo
d. Un triángulo con un ángulo obtuso

---

1. Si la covarianza de las variables X e Y es mayor que 0, las variables X e Y presentan

a. Correlación positiva
b. La regresión podría ser no lineal
c. Las variables podrían estar no correlacionadas
d. La pendiente de la recta de regresión tiene un signo positivo
e. El coeficiente de correlación es positivo

2. Juan correlaciona alturas y pesos de los estudiantes varones de 1º de la Diplomatura en Empresariales utilizando como unidades el metro y el kilogramo. Ángela registra las alturas y los pesos empleando centímetros y gramos como medida. Ambos calculan la correlación entre sus dos conjuntos de medidas.

a. La correlación de Ángela será mayor que la de Juan
b. Los dos coeficientes de correlación serán aproximadamente iguales
c. El coeficiente de Juan tenderá a ser mayor que el de Ángela
d. El valor del coeficiente depende de la dispersión de los datos

3. Cuando la intensidad de la relación entre dos variables decrece

a. La pendiente de la recta de regresión de Y sobre X crece
b. La pendiente de la recta de regresión de X sobre Y crece
c. Hay mayor dispersión en la nube de puntos
d. La covarianza aumenta de valor absoluto

4. Si las dos variables están correlacionadas positivamente

a. Cuando una aumenta, la otra también aumenta
b. Cuando una disminuye, la otra aumenta
c. Cuando una disminuye, la otra disminuye
d. La relación entre las dos variables es de tipo lineal

5. Ordene los siguientes valores según expresen mayor correlación entre dos variables: 0'5, -0'8, 0'2, -0'4,0.

_____ mayor valor de la correlación
_____
_____
_____
_____ no existe correlación

6. Si el coeficiente de correlación entre dos variables es nulo

a. Ambas rectas de regresión de Y sobre X y de X sobre Y son paralelas
b. La covarianza también es nula
c. Ambas rectas de regresión de Y sobre X y de X sobre Y coinciden
d. Las variables pueden tener una relación no lineal
e. Ambas rectas de regresión de Y sobre X y de X sobre Y son perpendiculares

7. Si r es el coeficiente de correlación de dos variables, indique qué afirmaciones son correctas

a. r = 0 indica que las variables son independientes
b. Si r = 0'6 la correlación entre las variables X e Y es doble que cuando r = 0'3
c. Una relación funcional entre variables se corresponde con un valor de r de +1 ó -1
d. El coeficiente de correlación puede interpretarse como un porcentaje de la varianza

8. Al estudiar las superficies sembradas de trigo en miles de hectáreas y las cosechas obtenidas en millones de quintales métricos, en cinco años consecutivos, el coeficiente de correlación obtenido fue 0'91. Luego

a. Podría haber otros factores que hagan variar los resultados
b. Deberíamos tomar una muestra más grande para poder expresar la relación entre la superficie plantada y la cosecha obtenida
c. La cosecha obtenida presenta una alta correlación con la superficie plantada
d. Si plantamos doble superficie, obtendremos con seguridad doble cosecha

9. ¿En qué predicción tendría más confianza?

a. La predicción de la estatura de un hombre a partir de su peso
b. La predicción del peso de un hombre a partir de su estatura
c. Las dos me dan la misma confianza

10. Las rentas se usan para predecir los ahorros, ambos medidos en miles de pesetas. Para la ecuación de regresión y = 1000 + 0'1x, ¿cuál de las siguentes afirmaciones es verdadera?

a. Y es la renta, X es el ahorro, la renta es la variable independiente
b. Y es la renta, X es el ahorro, el ahorro es la variable independiente
c. Y es el ahorro, X es la renta, el ahorro es la variable independiente
d. Y es el ahorro, X es la renta, la renta es la variable independiente

11. ¿Qué valor ha de tener r si las dos rectas de regresión tienen una pendiente idéntica?

a. 0
b. 1
c. -1
d. 0'5

12. Si la correlación entre X e Y es perfecta, el ángulo que forman las rectas de regresión es de

a. 120º
b. 90º
c. 45º
d. 0º

A continuación planteamos una serie de preguntas en que se le solicita una estimación, que dibuje un diagrama de dispersión, etc. Deberá razonar *la respuesta que dé.* NO DEBE EFECTUAR CÁLCULOS NUMÉRICOS *en estas cuestiones.*

1. Dadas las siguientes parejas de variables, dibuje un diagrama de dispersión que contenga 10 puntos y que muestre razonablemente su variación conjunta.

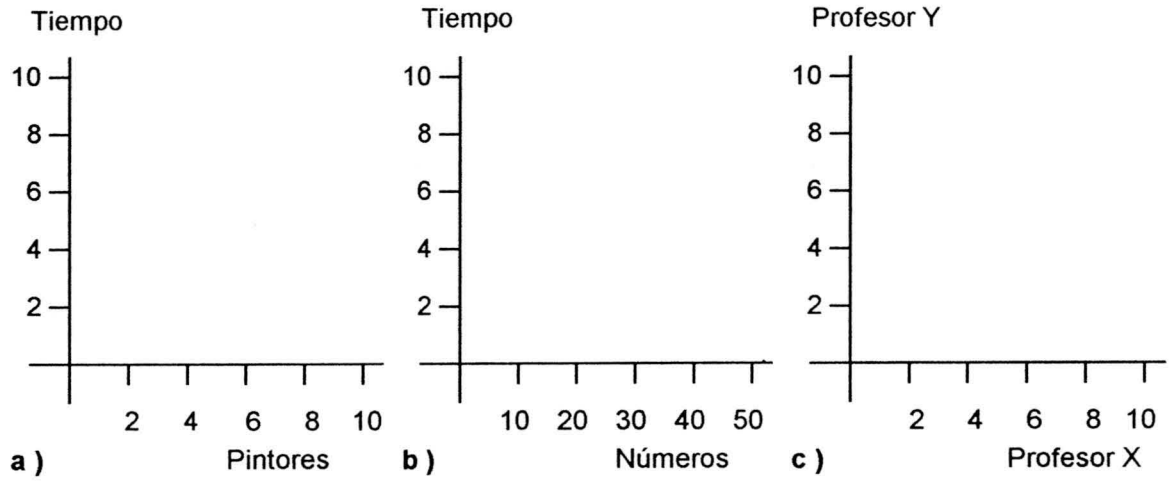
a) Número de pintores pintando una habitación y tiempo en horas para acabar el trabajo

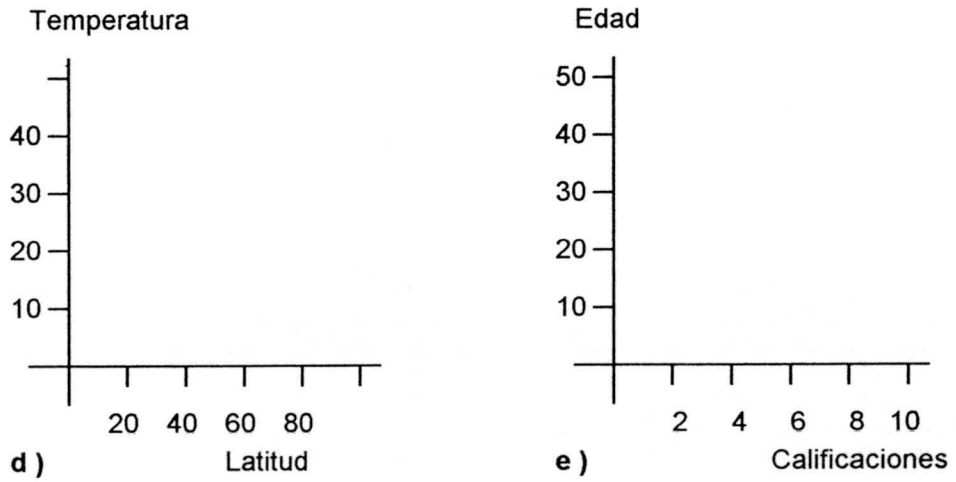
b) Longitud de una lista de números y tiempo empleado por una persona en memorizarla

c) Calificaciones de un mismo examen por dos profesores de un tribunal de oposiciones

d) La latitud de capitales europeas y la temperatura que hace en ellas un día determinado

e) Las calificaciones en Estadística y la edad del alumno





2. Dadas las siguientes parejas de variables, indique un valor razonable para el coeficiente de correlación que expresaría el tipo de relación entre las variables (directa, inversa o independencia) y su intensidad (fuerte o débil).

a) Altura y envergadura -distancia entre los extremos de los brazos puestos en cruz- de los estudiantes de la Diplomatura en Empresariales

r =

b) Ordenación por grado de timidez de los estudiantes de la Universidad de Jaén y ordenación por número de ciudades diferentes que han visitado

r =

c) Grado de ambición y estatura de los estudiantes de la Universidad de Jaén

r =

d) Tiempo semanal dedicado por los estudiantes a actividades atléticas y la posición que obtienen en una prueba de rendimiento físico

r =

e) Número de días de lluvia y número de horas de sol registradas durante un año por un observatorio de las diferentes comunidades autónomas

r =

3. Para cada una de las siguientes tablas de datos, estimar un valor razonable del coeficiente de correlación que muestre el tipo de relación entre las variables (directa, inversa o independencia) y su intensidad (fuerte o débil). **[No efectuar cálculos numéricos]**

a) Tiempo en meses desde que se prepara un medicamento y porcentaje de efectividad para una cierta enfermedad

Tiempo en meses	1	2	3	4	5
% de efectividad	90	75	42	30	21

r =

b) Calificaciones de 10 alumnos de COU en los exámenes de Matemáticas y Física

Matemáticas	1	2	2	3	4	4	5	6	7	7
Física	3	7	6	2	2	7	4	5	3	5

r =

c) Calificaciones de 10 alumnos de COU en los exámenes de Matemáticas y Educación Física

Matemáticas	2	3	4	4	5	6	7	8	9	10
Educación Física	2	5	7	8	5	4	6	5	5	9

r =

d) Valores de la presión sanguínea antes y después de haber efectuado un cierto tratamiento médico a un grupo de 10 mujeres

	Presión sanguínea en cada mujer									
Mujer	Sra A	Sra B	Sra C	Sra D	Sra E	Sra F	Sra G	Sra H	Sra I	Sra J
Antes tratamiento	115	112	107	119	115	138	126	105	104	115
Después tratamiento	128	115	106	128	122	145	132	109	102	117

r =

e) Ordenación dada por dos entrenadores a 10 atletas según su estado físico

	A	B	C	D	E	F	G	H	I	J
Entrenador A	1	2	3	4	5	6	7	8	9	10
Entrenador B	2	8	1	3	9	10	4	5	7	6

r =

4. Dadas las siguientes gráficas que representan la variación conjunta de variables, estimar el valor de su coeficiente de correlación teniendo en cuenta el tipo de relación (directa, inversa o independencia) y la intensidad de la misma. **[No efectuar cálculos numéricos]**

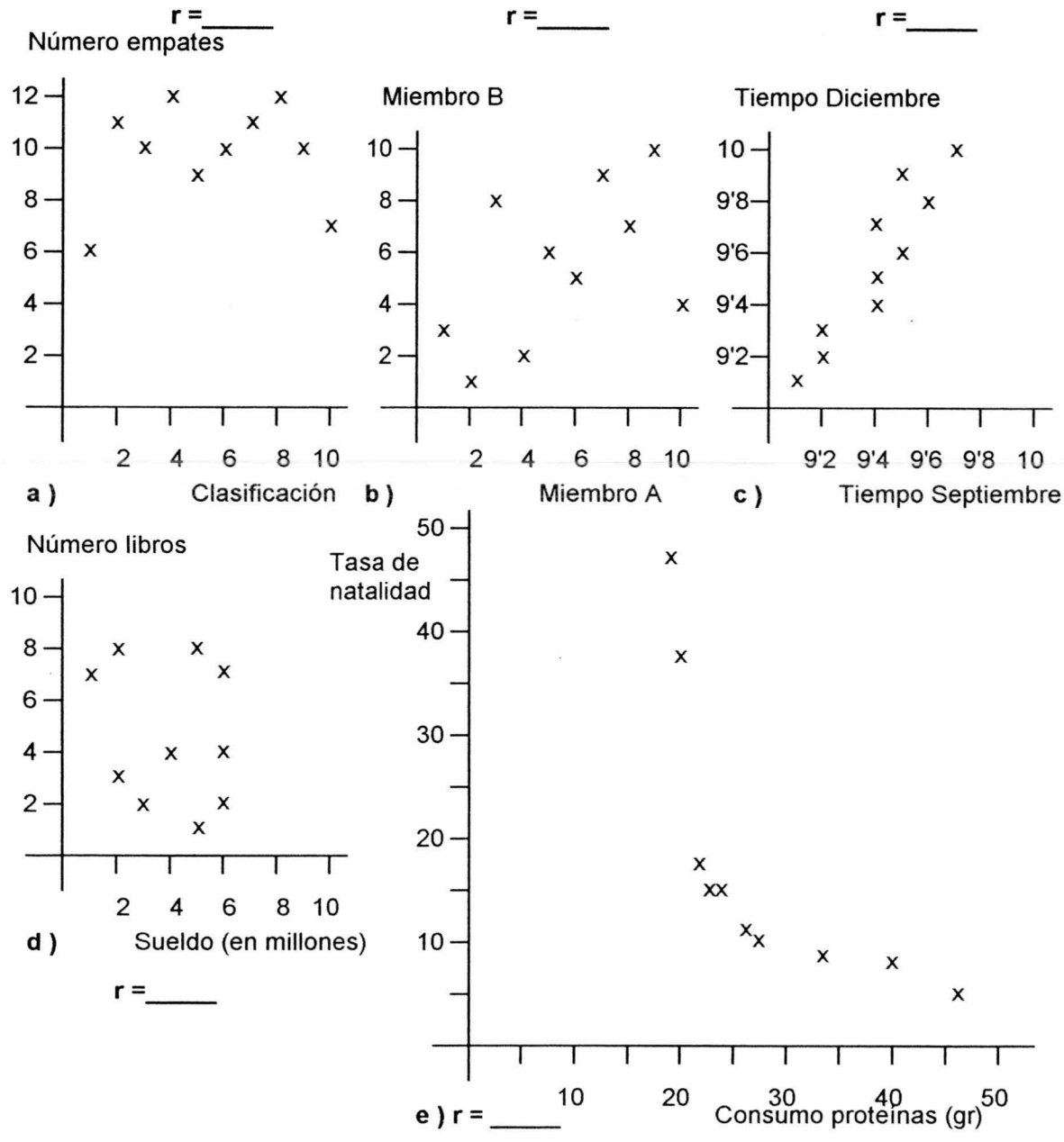
a) Puesto ocupado por los 10 primeros equipos de 1ª División de la liga de futbol en la temporada 1.987-88 y los partidos empatados

b) Puntuaciones otorgadas por los miembros A y B de un tribunal a 10 proyectos presentados

c) Tiempo, en segundos, de 10 atletas en correr 100 m lisos en septiembre y diciembre

d) Sueldo, en millones de pesetas, de los empleados de una empresa y número de libros que leen al cabo de un año

e) Tasa de natalidad y consumo diario de proteínas animales en 10 países



5. Dados los siguientes valores del coeficiente de correlación lineal, describir dos variables para las cuales fuese razonable obtener este coeficiente de correlación en función del tipo de dependencia entre las variables (directa o inversa) y la intensidad de la misma.

a) r = 1

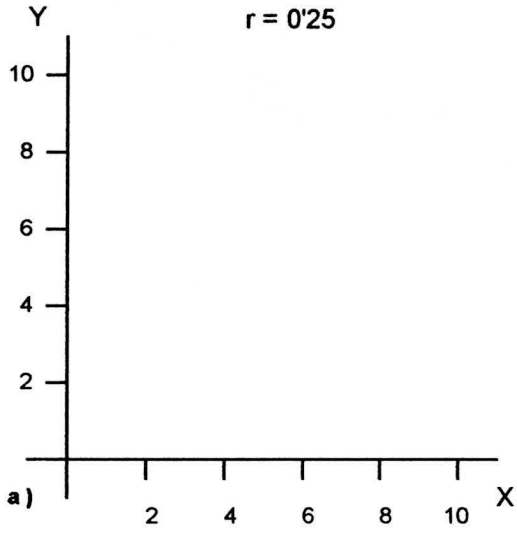
b) r = -0'3

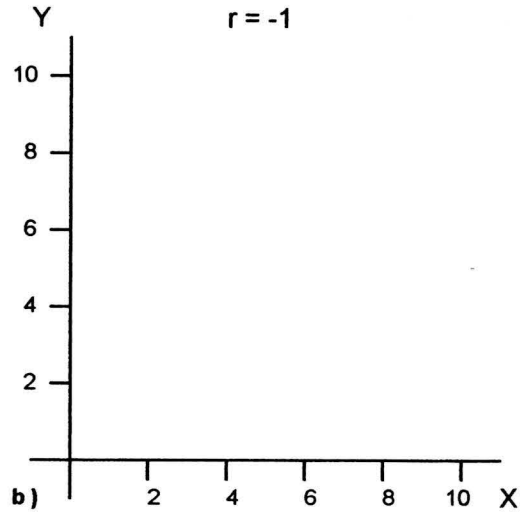
c) r = 0'05

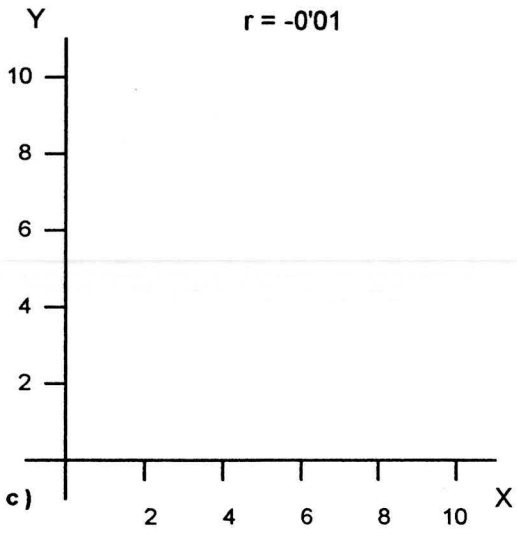
d) r = -0'8

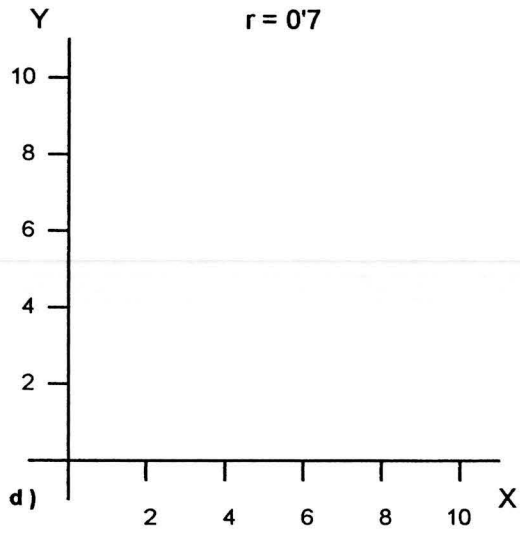
e) r = 0'5

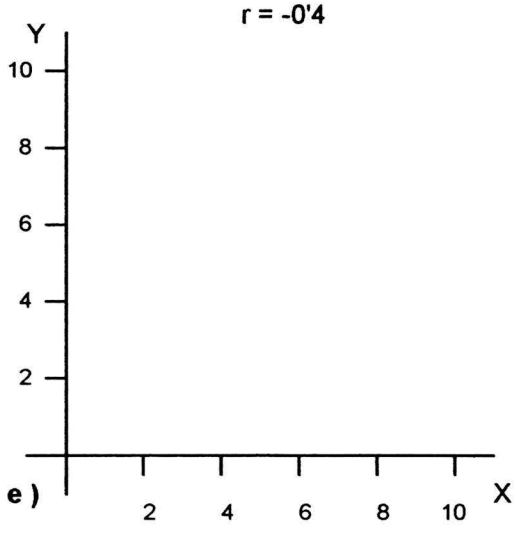
6. Dados los siguientes valores del coeficiente de correlación entre dos variables X e Y, dibujar un diagrama de dispersión, con 10 puntos, que se adapte razonablemente a ellos.











Debe responder a los siguientes problemas realizando los cálculos numéricos que considere oportunos.

Problema 1. Una recta de regresión tiene una pendiente de 16 y corta al eje de ordenadas en el punto y = 4. Si la media de la variable independiente es 8, ¿cuál es la media de la variable dependiente?

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que están presentes en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

a) Calcule el coeficiente de correlación lineal

b) Decir qué tipo de relación (directa, inversa o independencia) existe entre ambas variables

c) Determine la recta de regresión de y, número de bacterias por unidad de volumen, sobre x, número de horas

d) ¿Qué número de bacterias cabe esperar que habrá, transcurridas 2'5 horas?¿Y cuando pasen 6 horas?

e) ¿Qué tiempo deberá pasar para que el número de bacterias del cultivo sea de 27?

Anexo VI

Tablas de respuestas de los alumnos a los items de opciones múltiples

Nota. En todo el Anexo VI las respuestas simbolizadas con una letra en minúscula (mayúscula) representan a una respuesta correcta (incorrecta).

ÍTEM 1. Si la covarianza de las variables X e Y es mayor que 0, las variables X e Y presentan

a) Correlación positiva

b) La correlación podría ser no lineal

c) Las variables podrían estar no correlacionadas

d) La pendiente de la recta de regresión tiene un signo positivo

e) El coeficiente de correlación es positivo

Tabla VI-1. Frecuencia y porcentaje de las respuestas al ítem 1

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
abde	2	1'0	a)	127	65'8
abd	2	1'0	e)	115	59'6
ade	32	16'6	d)	87	45'1
bde	1	0'5	b)	22	11'4
ab	7	3'6	C)	17	8'8
ad	16	8'3			
ae	44	22'8			
bd	4	2'1			
be	1	0'5			
de	16	8'3			
a	20	10'4			
b	1	0'5			
d	8	4'1			
e	12	6'2			
abCde	1	0'5			
aCde	2	1'0			
aCe	1	0'5			
Cde	1	0'5			
bC	3	1'6			
Cd	2	1'0			
Ce	2	1'0			
C	5	2'6			
No responde	10	5'2			
Total	193	100			

ÍTEM 2. Juan correlaciona alturas y pesos de los estudiantes varones de 1º de la Diplomatura en Empresariales utilizando como unidades el metro y el kilogramo. Ángela resgistra las alturas y los pesos empleando centímetros y gramos como medida. Ambos calculan la correlación entre sus dos conjuntos de medidas.

a) La correlación de Ángela será mayor que la de Juan

b) Los dos coeficientes de correlación serán aproximadamente iguales

c) El coeficiente de Juan tenderá a ser mayor que el de Ángela

d) El valor del coeficiente depende de la dispersión de los datos

Tabla VI-2. Frecuencia y porcentaje de las respuestas al ítem 2

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
bd	44	22'8	b)	107	55'4
b	61	31'6	d)	102	52'9
d	44	22'8	A)	23	11'9
Cd	9	4'7	C)	17	8'8
Ad	5	2'6			
Ab	2	1'0			
C	6	3'1			
A	14	7'3			
AC	2	1'0			
No responde	6	3'1			
Total	193	100			

ÍTEM 3. Cuando la intensidad de la relación entre dos variables decrece

a) La pendiente de la recta de regresión de Y sobre X crece

b) La pendiente de la recta de regresión de X sobre Y crece

c) Hay mayor dispersión en la nube de puntos

d) La covarianza aumenta de valor absoluto

Tabla VI-3. Frecuencia y porcentaje de las respuestas al ítem 3

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
c	100	51'8	c)	124	64'3
cD	17	8'8	D)	33	17'1
Bc	3	1'6	B)	32	16'6
Ac	1	0'5	A)	23	12'0
AcD	1	0'5			
ABc	2	1'0			
B	17	8'8			
D	12	6'2			
A	10	5'2			
AB	8	4'1			
BD	2	1'0			
AD	1	0'5			
No responde	19	9'8			
Total	193	100			

ÍTEM 4. Si las dos variables están correlacionadas positivamente

a) Cuando una aumenta, la otra también aumenta

b) Cuando una disminuye, la otra aumenta

c) Cuando una disminuye, la otra disminuye

d) La relación entre las dos variables es de tipo lineal

Tabla VI-4. Frecuencia y porcentaje de las respuestas al ítem 4

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
ac	57	29'5	a)	152	78'7
a	29	15'0	D)	91	47'2
c	3	1'6	c)	90	46'6
aBc	1	0'5	B)	8	4'1
aD	37	19'2			
aB	1	0'5			
cD	1	0'5			
Bc	1	0'5			
acD	27	14'0			
D	23	11'9			
B	2	1'0			
BD	3	1'6			
No responde	8	4'1			
Total	193	100			

ÍTEM 5. Ordene los valores según expresen mayor correlación entre dos variables 0'5, -0'8, 0'2, -0'4, 0

_____ mayor correlación

_____

_____

_____

_____ no existe correlación

Tabla VI-5. Frecuencia y porcentaje de las respuestas al ítem 5

Ordenación	Frecuencia	Porcentaje
-0'8,0'5,-0'4,0'2,0	89	46'1
0'5,0'2,0,-0'4,-0'8	33	17'1
0'5,0'2,-0'4,-0'8,0	26	13'5
0'5,0'2,-0'8,-0'4,0	9	4'7
0,0'2,-0'4,0'5,-0'8	3	1'6
-0'8,-0'4,0'5,0'2,0	3	1'6
0,0'2,0'5,-0'4,-0'8	2	1'0
0'5,0'2,0,-0'8,-0'4	2	1'0
0'5,-0'8,-0'4,0'2,0	2	1'0
-0'8,-0'4,0'2,0'5,0	2	1'0
Otras*	15	7'8
No responde	7	3'6
Total	193	100

*Otras ordenaciones con frecuencia absoluta 1

ÍTEM 6. Si el coeficiente de correlación entre dos variables es nulo

a) Ambas rectas de regresión de Y sobre X y de X sobre Y son paralelas

b) La covarianza también es nula

c) Ambas rectas de regresión de Y sobre X y de X sobre Y coinciden

d) Las variables pueden tener una relación no lineal

e) Ambas rectas de regresión de Y sobre X y de X sobre Y son perpendiculares

Tabla VI-6. Frecuencia y porcentaje de las respuestas al ítem 6

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
bde	9	4'7	e)	96	49'7
bd	14	7'3	b)	84	43'5
bc	3	1'6	d)	64	33'2
be	32	16'6	A)	28	14'5
de	11	5'7	C)	19	9'8
b	15	7'8			
e	42	21'8			
d	18	9'3			
Abd	2	1'0			
Ab	9	4'7			
Ae	1	0'5			
Ad	5	2'6			
Cd	5	2'6			
Ce	1	0'5			
A	11	5'7			
C	10	5'2			
No responde	5	2'6			
Total	193	100			

ÍTEM 7. Si r es el coeficiente de correlación de dos variables, indique qué afirmaciones son correctas

a) $r = 0$ indica que las variables son independientes

b) Si $r = 0.6$ la correlación entre las variables X e Y es doble que cuando $r = 0.3$

c) Una relación funcional entre variables se corresponde con un valor de r de +1 ó -1

d) El coeficiente de correlación puede interpretarse como un porcentaje de la varianza

Tabla VI-7. Frecuencia y porcentaje de las respuestas al ítem 7

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
ac	66	34'2	a)	138	71'5
a	37	19'2	c)	112	58'0
c	29	15'0	B)	44	22'8
acD	4	2'1	D)	20	10'4
aBc	3	1'6			
aBcD	1	0'5			
aB	19	9'8			
aD	7	3'6			
Bc	8	4'1			
cD	1	0'5			
aBD	1	0'5			
B	8	4'1			
D	2	1'0			
BD	4	2'1			
No	3	1'6			
Total	193	100			

ÍTEM 8. Al estudiar las superficies sembradas de trigo en miles de hectáreas y las cosechas obtenidas en millones de quintales métricos, en cinco años consecutivos, el coeficiente de correlación obtenido fue de 0'91. Luego

a) Podría haber otros factores que hagan variar los resultados

b) Deberíamos tomar una muestra más grande para poder expresar la relación entre la superficie plantada y la cosecha obtenida

c) La cosecha obtenida presenta una alta correlación con la superficie plantada

d) Si plantamos doble superficie, obtendremos con seguridad doble cosecha

Tabla VI-8. Frecuencia y porcentaje de las respuestas al ítem 8

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
ac	39	20'2	c)	172	89'1
a	7	3'6	a)	56	29'0
c	93	48'2	D)	43	22'3
aBc	2	1'0	B)	6	3'1
acD	4	2'1			
aB	1	0'5			
aD	3	1'6			
Bc	2	1'0			
cD	32	16'6			
B	1	0'5			
D	4	2'1			
No responde	5	2'6			
Total	193	100			

ÍTEM 9. ¿En qué predicción tendría más confianza?

a) La predicción de la estatura de un hombre a partir de su peso

b) La predicción del peso de un hombre a partir de su estatura

c) Las dos me dan la misma confianza

Tabla VI-9. Frecuencia y porcentaje de las respuestas al ítem 9

Respuestas	Frecuencia	Porcentaje
abc	1	0'5
a	12	6'2
b	66	34'2
c	113	58'5
No responde	1	0'5
Total	193	100

ÍTEM 10. Las rentas se usan para predecir los ahorros, ambos medidos en miles de pesetas. Para la ecuación de regresión $y = 1000 + 0.1x$, ¿cuál de las siguientes afirmaciones es verdadera?

a) Y es la renta, X es el ahorro, la renta es la variable independiente

b) Y es la renta, X es el ahorro, el ahorro es la variable independiente

c) Y es el ahorro, X es la renta, el ahorro es la variable independiente

d) Y es el ahorro, X es la renta, la renta es la variable independiente

Tabla VI-10. Frecuencia y porcentaje de las respuestas al ítem 10

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
d	50	25'9	d)	70	36'3
Bd	20	10'4	B)	66	34'2
A	15	7'8	C)	56	29'0
AC	11	5'7	A)	26	13'5
B	42	21'8			
C	41	21'2			
BC	4	2'1			
No responde	1	0'5			
Total	193	100			

ÍTEM 11. ¿Qué valor ha de tener *r* si las dos rectas de regresión tienen una pendiente idéntica?

a) 0

b) 1

c) -1

d) 0.5

Tabla VI-11. Frecuencia y porcentaje de las respuestas al ítem 11

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
bc	75	38'9	b)	149	77'2
b	72	37'3	c)	80	41'4
c	3	1'6	A)	29	15'0
Abc	2	1'0	D)	7	3'6
A	27	14'0			
D	7	3'6			
No responde	6	3'1			
Total	193	100			

ÍTEM 12. Si la correlación entre X e Y es perfecta, el ángulo que forman las rectas de regresión es de

a) 120º
b) 90º
c) 45º
d) 0º

Tabla VI-12. Frecuencia y porcentaje de las respuestas al ítem 12

Respuestas	Frecuencia	Porcentaje	Opción	Frecuencia	Porcentaje
d	125	64'8	d)	129	66'8
Ad	3	1'6	B)	32	16'6
Cd	1	0'5	C)	24	12'4
A	2	1'0	A)	5	2'6
B	32	16'6			
C	23	11'9			
No responde	7	3'6			
Total	193	100			

Anexo VII

Tablas de respuestas de los alumnos a los problemas

1. DESCRIPCIÓN DE LOS PROCEDIMIENTOS EMPLEADOS POR LOS ALUMNOS EN EL PROBLEMA 1

10 = Se ha usado esta codificación cuando el alumno no ha contestado a la pregunta.

11 = Dado que el enunciado del problema no especifica qué variable es la dependiente y cuál la independiente, la respuesta recoge ambas rectas de regresión. Ejemplo:

Sujeto 67: *Suponemos la recta Y/X:* $x = 0, y = 4, y = ax + b, 4 = 0 + b, b = 4$

$$b = \bar{y} - a\bar{x}, 4 = \bar{y} - 16 \cdot 8, 4 = \bar{y} - 128, \bar{y} = 132$$

Suponemos la recta X/Y: $x = a'y + b, 0 = 16 \cdot 4 + b, b = -64$

$$b = \bar{x} - a'\bar{y}\ , -64 = \bar{x} - 16 \cdot 8, \bar{x} = 64$$

12 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, y únicamente toma en consideración la ecuación explícita de la recta de regresión de Y sobre X: $y = mx + n$. Ejemplo:

Sujeto 110: $\bar{y} = a + b\bar{x}, \bar{y} = 4 + 16 \cdot 8 = 132$

13 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, y únicamente toma en consideración la ecuación punto-pendiente de la recta de regresión de Y sobre X: $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$. Ejemplo:

Sujeto 14: $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}), y - \bar{y} = 16(x - \bar{x}), 4 - \bar{y} = 16(0 - 8),$

$$-\bar{y} = -128 - 4, \bar{y} = 132$$

14 = En este caso el alumno utiliza ambas expresiones de la ecuación de la recta de regresión de Y sobre X para resolver el problema planteado. Ejemplo:

Sujeto 56: $x = 1, y = 20,\ 20 - \bar{y} = 16\ (1 - 8),\ \bar{y} = \frac{-112}{-20} = 5'6$

15 = En este caso el alumno estima que Y es la variable independiente y que X es la variable dependiente, y únicamente toma en consideración la ecuación explícita de la recta de regresión de X sobre Y: $x = m'y + n'$. Ejemplo:

Sujeto 117: $x = a' + b'y, (0,4), a' = x - b'\cdot 4, a' = -16\cdot 4, a' = \bar{x} - b'\bar{y}$,

$$\bar{x} = -16\cdot 4 + 16\cdot 8, \bar{x} = 64$$

16 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero utiliza una expresión incorrecta de la ecuación de la recta de regresión de Y sobre X: $y - \bar{y} = \frac{S_{xy}}{S_x S_y}(x - \bar{x})$. Ejemplo:

Sujeto 1: $pendiente = b = \frac{S_{xy}}{S_x S_y} = 16, y = 4, \bar{x} = 8, \bar{y} = ?, y/x$

$$y - \bar{y} = \frac{S_{xy}}{S_x S_y}(x - \bar{x}), 4 - \bar{y} = 16\,(x - 8), 4 - \bar{y} = 16x - 128,$$

$$\bar{y} = -16x + 124, \textit{Dependiendo del punto } x$$

17 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero utiliza una expresión incorrecta de la ecuación de la recta de regresión de Y sobre X: $y - \bar{y} = \frac{S_{xy}}{S_x}(x - \bar{x})$. Ejemplo:

Sujeto 2: $x = 0, y = 4, \bar{x} = 8, \bar{y} = ?, \frac{S_{xy}}{S_x} = 16, y - \bar{y} = \frac{S_{xy}}{S_x}(x - \bar{x}),$

$$4 - \bar{y} = 16 - 8, \bar{y} = -16 + 8 + 4 = -4$$

18 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero utiliza una expresión incorrecta de la ecuación de la recta de regresión de Y sobre X: $y - \bar{y} = \frac{S_{xy}}{S_x^2} - (x - \bar{x})$. Ejemplo:

Sujeto 4: $\bar{x} = 8, y = 4, y - \bar{y} = \frac{S_{xy}}{S_x^2} - (x - \bar{x}), y - \bar{y} = 16 - 8, \bar{y} = 8 - 4, \bar{y} = 4$

19 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, aunque toma en consideración la recta de regresión de X sobre Y: $x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$. Ejemplo:

Sujeto 122: $x = 0,\ y = 4, \bar{x} = 8, x - \bar{x} = \frac{\mathrm{Cov(X,Y)}}{S_y^2}(y - \bar{y}), x - 8 = 16(4 - \bar{y})$

$0 - 8 = 16(4 - \bar{y}), 16\bar{y} = 16 \cdot 4 + 8 = 72, \bar{y} = 4'5$

La ecuación de la recta viene definida como $x - x_1 = m(y - y_1)$ *siendo m la pendiente de la recta. Como nos dan la pendiente, el valor de y, y sabemos que corta al eje Y en el punto* 4, *en ese punto* $x = 0$. *Al sustituir, la media de la variable dependiente sale 4'5.*

110 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero utiliza una expresión incorrecta de la ecuación de la recta de regresión de X sobre Y: $x - \bar{x} = \frac{S_{xy}}{S_{y,x}}(y - \bar{y})$. Ejemplo:

Sujeto 13: *pendiente* = 16, *punto* $y = 4,\ \bar{y} = 8, x - \bar{x} = \frac{S_{xy}}{S_{y,x}}(y - \bar{y})$,

$4 = 8 - 16y, 4 - 8 = -16x, -4 = -16x, \bar{y} = \frac{16}{4} = 4$

111 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero utiliza una expresión incorrecta de la ecuación de la recta de regresión de Y sobre X: $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}$. Ejemplo:

Sujeto 16: $b = 16, y = 4, \bar{x} = 8,\ \bar{y} = ?, y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}, 4 - \bar{y} = 16 \cdot 8,$

$\bar{y} = -128 + 4 = 124, \bar{y} = 124$

112 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, usando la ecuación explícita de la recta de regresión de Y sobre X, $y = mx + n$, pero no interpreta de forma adecuada los coeficientes: m=8, n=16. Ejemplo:

Sujeto 146: $y = a + bx, y = 4, a = 16, x$?, $4 = 16 - 8x, 12 = 8x, x = 12 / 8 = 1'5$

113 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, contemplando el uso de la recta de regresión de Y sobre X, $y - \bar{y} = m\ (\ x - \bar{x}\ )$, pero no interpreta de forma adecuada cuál es el eje de ordenadas. Ejemplo:

Sujeto 10: $y - \bar{y} = 16(4 - 8),\ y - \bar{y} = -64,\ \bar{y} = 64 + y$

114 = En este caso el alumno estima que Y es la variable independiente y que X es la variable dependiente, contemplando el uso de la recta de regresión de X sobre Y, $x - \bar{x} = m(y - \bar{y})$, pero interpreta inadecuadamente la abscisa del punto de corte con el eje de ordenadas: $x = 4$. Ejemplo:

Sujeto 7: $x = 4, x - \bar{x} = 16(4 - \bar{y}), -4 = 64 - 16\bar{y}, 16\bar{y} = 68,\ \bar{y} = 4'25$

115 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, contemplando el uso de la recta de regresión de Y sobre X, $y - \bar{y} = m(x - \bar{x})$, pero interpreta inadecuadamente la variable independiente: $\bar{y} = 8$. Ejemplo:

Sujeto 57: $\frac{S_{xy}}{S_x^2} = 16,\ (0,4),\ \bar{y} = 8,\ y - 8 = 16(0 - \bar{x}), -4 = -16\bar{x}, \bar{x} = 1/4$

116 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, contemplando el uso de la recta de regresión de Y sobre X, $y = mx + n$, pero interpreta inadecuadamente la pendiente de la recta de regresión $y = ax + b$: $m = -\frac{b}{2a}$. Ejemplo:

Sujeto 19: $y = ax + b,\ 4 = 0 + b,\ y = ax + 4,\ pendiente = 16,\ 16 = -\frac{b}{2a} = -\frac{4}{2a}$

$$32a = -4,\ a = 1/8,\ y = \frac{1}{8}x + 4$$

117 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero interpreta de forma inadecuada la pendiente en la ecuación explícita de la recta de regresión Y sobre X ($m = 1/16$) y, también, la media de la variable dependiente ($a = \bar{y}$). Ejemplo:

Sujeto 34: $y = ax + b, y = 4, \bar{x} = 8, 4 = a + (1/16)\cdot 8, a = 4 - (1/16)\cdot 8 = 4'5, \bar{y} = 4'5$

118 = En este caso el alumno no utiliza la recta de regresión, basándose en la aplicación de la noción de media, pero de forma no pertinente. Además, no interpreta adecuadamente la pendiente de la recta. Ejemplo:

Sujeto 35: $x = 16, y = 4, 8 = 16 / z, z = 16 / 8 = 2, \bar{y} = 4 / 2 = 2$

119 = En este caso el alumno utiliza la recta de regresión de Y sobre X, pero, en la ecuación de la recta de regresión $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$, interpreta inadecuadamente quién es la variable independiente: $\bar{y} = 8$. Ejemplo:

Sujeto 38: $y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}), 4 - 8 = 16x - \bar{x}, -4 = 16x - \bar{x}, \bar{x} = 16x + 4$

120 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero, en la ecuación explícita de la recta de regresión de Y sobre X, $y = mx + n$, intercambia el significado de los coeficientes m y n de la misma: $m = 4, n = 16$. Ejemplo:

Sujeto 114: $a = 16, b = 4, \bar{x} = 8, \bar{y} = ?, y = a + bx = 16 + 4x,$
$y = 16 + 4\cdot 8 = 16 + 32 = 48,$ *la media de la variable dependiente es*
$\bar{y} = 48$

121 = En este caso el alumno utiliza la recta de regresión de Y sobre X, pero, en la ecuación explícita de la recta de regresión $y = mx + n$, interpreta inadecuadamente quién es la variable independiente: $\bar{y} = 8$. Ejemplo:

Sujeto 46: $y = a + bx, b =$ *pendiente* $= 16, a = 4, \bar{y} = 8, a = \bar{y} - b\bar{x}, 4 = 8 - 16\bar{x},$
$16\bar{x} = 8 - 4, 16\bar{x} = 4, \bar{x} = 4/16 = 1/4 = 0'25$ *media variable dependiente*

122 = En este caso el alumno utiliza la recta de regresión de Y sobre X, pero opera de forma algebraica inadecuada con la expresión de la ecuación de la recta. Ejemplo:

Sujeto 54: $y = a + bx,\ y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}), y = \bar{y}\ \frac{S_{xy}}{S_x^2} - \bar{x}$

$y = 16\,\bar{y}$ - $8x, x = 0, y = 4, 4 = 16\,\bar{y}$

123 = En este caso el alumno utiliza ambas rectas de regresión para determinar el punto de intersección de ellas, y, a partir del mismo, obtener la media de la variable dependiente. Ejemplo:

Sujeto 75: $x = 0, y = 4,$ $\left.\begin{array}{l} Recta\ Y/X\ \ y = 16x + 4 \\ Recta\ X/Y\ \ x = 16y - 64 \end{array}\right\} \left.\begin{array}{l} y = 16x + 4 \\ y = (x/16) + 4 \end{array}\right\}$

$y = 3'96$ *por tanto* $\bar{x} = 3'96$

124 = En este caso el alumno utiliza una recta vectorial, $y = kx$, como recta de regresión de Y sobre X para determinar la media de la variable dependiente. Ejemplo:

Sujeto 80: $\frac{S_{xy}}{S_x^2} = 16, y = 4, y = 16x,\ 16x = 4, x = 4$

125 = En este caso el alumno no utiliza una recta de regresión, aplicando de forma inadecuada el concepto de media estadística y el de pendiente. Ejemplo:

Sujeto 85: $\bar{x} = 16/8 = 2, x = 2, 2/2 = 1, \bar{x} = 1$

126 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, pero utiliza una expresión inadecuada de la ecuación de la recta de regresión de Y sobre X: $n = \bar{y} + m\,\bar{x}$. Ejemplo:

Sujeto 88: $a = 16, b = \bar{y} + a\,\bar{x}, y = ax + b, b = y - ax = 4 - 16 \cdot 0 = 4$

$4 = \bar{y} + 16 \cdot 8,\ \bar{y} = 4 - 128 = -124$

127 = En este caso el alumno estima que Y es la variable independiente y que X es la variable dependiente, contemplando el uso de la ecuación explícita de la recta de regresión de X sobre Y, $x = m'y + n'$, pero interpreta de forma inadecuada los coeficientes m' y n'. Ejemplo:

Sujeto 89: $a = \bar{x} - b\,\bar{y}$, $16 = 8 - 4\,\bar{y}$, $\bar{y} = \frac{-16+8}{4} = 2$

128 = En este caso el alumno utiliza una expresión inadecuada de la ecuación de la recta de regresión $y = \bar{y} - m\,\bar{x}$, y confunde cuál es la variable independiente. Ejemplo:

Sujeto 98: $y = ax + b$, $a = 16$, *punto* $(0,4)$, $\bar{y} = 8$, $4 = 16 \cdot 0 + b$, $b = 4$
$y = \bar{y} - a\,\bar{x}$, $4 = 8 - 16\,\bar{x}$, $-4 = -16\,\bar{x}$, $\bar{x} = -4/-16 = 1/4$

129 = En este caso el alumno estima que X es la variable independiente y que Y es la variable dependiente, contemplando el uso de la ecuación explícita de la recta de regresión de Y sobre X, pero muestra un concepto inadecuado de la pendiente de la recta: $m = sen\ 16°$. Ejemplo:

Sujeto 98: $y = 4 + \text{sen } 16° \cdot 8 = 4 + 2'20 = 6'20$

130 = En este caso el alumno interpreta de forma inadecuada el centro de gravedad. Ejemplo:

Sujeto 103: *Las rectas se cortan siempre en el punto* $(\bar{x}, \bar{y})$, *por lo que la variable dependiente* $= 4$ $(\bar{y} = 4)$.

131 = En este caso el alumno utiliza la recta de regresión de X sobre Y, contemplando el uso de la ecuación explícita de la recta de regresión de X sobre Y, $x = m'y + n'$, pero no interpreta de forma conveniente quién es la variable independiente. Ejemplo:

Sujeto 105: $x = 8$, $y = 4$, $b' = 16$, $x = a' + b'y$, $8 = a' + 16 \cdot 4$, $8 = a' + 64$, $a' = 64 - 8 = 56$

132 = En este caso el alumno estima que es X la variable independiente y que es Y la variable dependiente, contemplando el uso de la ecuación explícita de la recta de regresión de Y sobre X, pero interpreta inadecuadamente la abscisa del punto de corte con el eje de ordenadas: $x = 8$. Ejemplo:

Sujeto 111: $a = \bar{y} - b\bar{x}, a = \bar{y} - 16 \cdot 8, a = \bar{y} - 128, y = a + bx, 4 = a + 16 \cdot 8, a = 124$
$124 = \bar{y} - 128, \bar{y} = 124 + 128 = 252$

133 = En este caso el alumno estima que es X la variable independiente y que es Y la variable dependiente, contemplando el uso de la ecuación explícita de la recta de regresión de Y sobre X, pero interpreta de forma inadecuada la abscisa del punto de corte con el eje de ordenadas y el concepto de media estadística. Ejemplo:

Sujeto 118: $y = a + bx, y = 4 + 16, y = 20$, *la media de la variable dependiente es* 10.

134 = En este caso el alumno utiliza la recta de regresión de Y sobre X, pero interpreta de forma inadecuada la ordenada del punto de corte con el eje OY y quién es la variable independiente: $\bar{y} = 8$. Ejemplo:

Sujeto 126: $b = 16, a = 0, \bar{y} = 8, \bar{x} = ?, y = a + bx, 8 = 0 + 16\,x, x = 0'5$

135 = En este caso el alumno utiliza una expresión inadecuada de la recta de regresión de Y sobre X, $y - \bar{y} = a + b\,(x - \bar{x})$, y no interpreta de forma pertinente la pendiente de la recta. Ejemplo:

Sujeto 131: $\bar{x} = 8, m = 16 = -b/a, b = -16\,a, y - \bar{y} = a + b\,(x - \bar{x})$,
$4 - \bar{y} = a + -16a\,(x - 8), 4 - \bar{y} = a - 16ax + 64, -\bar{y} = a\,(1 - 16x) + 60$
$\bar{y} = -a\,(1 - 16x) - 60, a = \bar{y} + 16a \cdot 8 = \bar{y} + 128\,a, \bar{y} = -127\,a$

136 = En este caso el alumno estima que es X la variable independiente y que es Y la variable dependiente, contempla el uso de la ecuación explícita de la recta de regresión de Y sobre X, pero interpreta de forma inadecuada la ordenada del punto de corte con el eje OY: $n = 8$. Ejemplo:

Sujeto 133: $y = a + bx$, $4 = 8 + 16x$, $x = \frac{4-8}{16} = -0'25$, *la media de la variable pendiente es* -0'25

137 = En este caso el alumno utiliza la ecuación explícita de la recta de regresión de X sobre Y, pero interpreta de forma inadecuada quién es la variable independiente: $\bar{x} = 8$. Ejemplo:

Sujeto 137: $a = 4$, $b = 16$, $\bar{x} = 8$, $a = \bar{x} - b\,\bar{y}$, $4 = 8 - 16\,\bar{y}$, $\bar{y} = -4/8 = -0'5$
la media de la variable independiente $\bar{y} = -0'5$

138 = En este caso el alumno no utiliza una recta de regresión, aplicando de forma no pertinente la noción de covarianza y de pendiente de la recta de regresión. Ejemplo:

Sujeto 142: $S_{xy} = \frac{\sum xy}{n} - \bar{x}\,\bar{y}$, $(16/2) - \bar{x}\,\bar{y}$, $(16/2) / 8 = \bar{x}$, $\bar{x} = 1$

139 = En este caso el alumno no utiliza una recta de regresión, sino que aplica la media estadística de forma no adecuada. Ejemplo:

Sujeto 150: *pendiente* = 16, $\bar{y} = y = 4$, $y = 8$,
la media de la variable dependiente es 8 pues la media es aquella medida que deja el 50 % *de los valores arriba y el* 50 % *de los valores abajo.*

140 = En este caso el alumno utiliza la ecuación explícita de la recta de regresión de Y sobre X, pero interpreta de forma no pertinente quién es la variable independiente y la pendiente de la recta. Ejemplo:

Sujeto 153: $m = 16, y = 4, x = 0, \bar{y} = 8, y = ax + b, 4 = 0 + b, b = 4,$
$a = \bar{y} - b\bar{x} = 8 - 4\bar{x}, 4 = 8 - 4\bar{x} + 4, 4\bar{x} = 8, \bar{x} = 2$

141 = En este caso el alumno utiliza la recta de regresión de Y sobre X, contemplando el uso de la ecuación explícita de la recta de regresión de Y sobre X, pero no interpreta de forma adecuada quién es la variable independiente. Ejemplo:

Sujeto 158: $y = a + bx, b = 16, p(0,4), \bar{y} = 8, a = y - bx, a = 0 - 16 \cdot 4 = 64,$
$a = \bar{y} - b\bar{x}, a - \bar{y} = b\bar{x}, \bar{x} = \frac{a - \bar{y}}{b} = \frac{64 - 8}{16} = 3'5$

142 = En este caso el alumno estima que es X la variable independiente y que es Y la variable dependiente, contemplando el uso de la ecuación explícita de la recta de regresión de Y sobre X, pero interpreta de forma inadecuada los coeficientes m y n: $m = 4, n = 0$. Ejemplo:

Sujeto 165: *La recta de regresión es* $y = a + bx, \bar{x} = 8, a = \bar{y} - b\bar{x}, 0 = \bar{y} - 4 \cdot 8, \bar{y} = 32$
Como corta al eje de ordenadas en el punto 4 *podemos decir que el punto es* (0,4)

143 = En este caso el alumno estima que es X la variable independiente y que es Y la variable dependiente, pero interpreta de forma inadecuada la independencia aleatoria de las características de una variable estadística bidimensional. Ejemplo:

Sujeto 68: $y = a\,x + b, b = \bar{y} - a\,\bar{x}, a = 16, \bar{x} = 8,$ *como son independientes* $\bar{x} = \bar{y}$ *entonces* $\bar{y} = 8,$ *ya que* $\bar{x} = 8$

Tabla VII-1. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra en el Problema 1

Procedimientos	Frecuencia	Porcentaje
10	64	33'2
11	2	1'0
12	31	16'1
13	8	4'1
14	2	1'0
15	2	1'0
16	3	1'6
17	1	0'5
18	1	0'5
19	2	1'0
110	1	0'5
111	1	0'5
112	3	1'6
113	1	0'5
114	1	0'5
115	5	2'6
116	1	0'5
117	1	0'5
118	2	1'0
119	2	1'0
120	5	2'6
121	24	12'4
122	1	0'5
123	1	0'5
124	1	0'5
125	1	0'5
126	2	1'0
127	4	2'1
128	1	0'5
129	1	0'5
130	1	0'5
131	1	0'5
132	1	0'5
133	1	0'5
134	1	0'5
135	1	0'5
136	2	1'0
137	2	1'0
138	1	0'5
139	2	1'0
140	1	0'5
141	2	1'0
142	1	0'5
143	1	0'5
Total	193	100'0

Tabla VII-2. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra al Problema 1

Respuestas	Frecuencia	Porcentaje
Respuesta correcta	45	23'3
Respuesta incorrecta	84	43'5
No responde	64	33'2
Total	193	100'0

Tabla VII-3. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra en el Problema 2 apartado a)

Procedimientos	Frecuencia	Porcentaje
No contesta	25	13'0
$r = \sigma_{xy} / (\sigma_x \sigma_y)$	161	83'4
$r = S_{xy} / (S_x^2 S_y^2)$	4	2'1
Usa otras expresiones inadecuadas de r	1	0'5
Usa coeficiente de determinación	2	1'0
Total	193	100'0

Tabla VII-4. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra al Problema 2 apartado a)

Respuestas	Frecuencia	Porcentaje
Respuesta correcta	106	54'9
Respuesta incorrecta	62	32'1
No responde	25	13'0
Total	193	100'0

2. DESCRIPCIÓN DE LAS ARGUMENTACIONES EMPLEADAS POR LOS ALUMNOS EN EL PROBLEMA 2 APARTADO b)

20 = Se ha usado esta codificación cuando el alumno no ha contestado a la pregunta.

21 = En este caso el alumno deduce el tipo de relación existente entre las variables basándose en el signo del coeficiente de correlación.

Ejemplo:

Sujeto 33: *La relación es directa ya que el coeficiente de correlación es positivo.*

22 = En este caso el alumno deduce el tipo de relación existente entre las variables a partir de su variación conjunta: *Cuando aumenta (disminuye) una variable aumenta (disminuye) la otra.*

Ejemplo:

Sujeto 8: *Relación directa puesto que a medida que va pasando el tiempo, el número de bacterias aumenta.*

23 = En este caso el alumno deduce el tipo de relación existente entre las variables basándose en el signo de la covarianza.

Ejemplo:

Sujeto 112: *A través del coeficiente de correlación se puede afirmar que existe una dependencia lineal alta. Para determinar si es directa o inversa se utiliza la covarianza*

$$S_{xy} = 23'8$$

Al ser positiva será directa.

24 = En este caso el alumno deduce que la relación existente entre las variables es directa a partir del coeficiente de determinación.

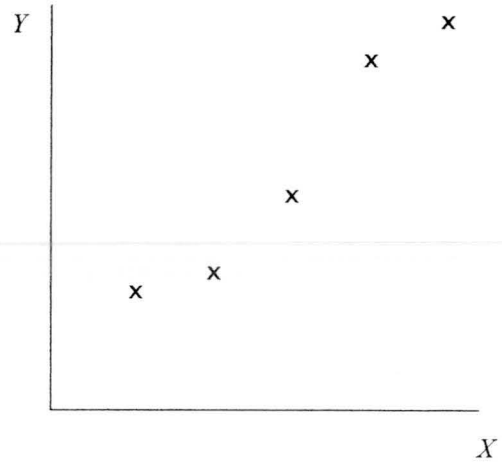
Ejemplo:

Sujeto 14: *La relación es directa e intensa ya que r^2 se aproxima a* 1.

25 = En este caso el alumno deduce el tipo de relación existente entre las variables basándose en el diagrama de dispersión que elabora a partir de la tabla de datos del problema.

Ejemplo:

Sujeto 100: *Según el gráfico que obtenemos al representar los datos la regresión es directa.*



26 = En este caso el alumno deduce que no existe relación entre las variables basándose en la proporcionalidad.

Ejemplo:

Sujeto 120: *Entre ambas variables existe una relación de independencia (no son proporcionales).*

27 = En este caso el alumno indica el tipo de relación existente entre las variables, pero no aporta argumentación alguna.

Ejemplo:

Sujeto 17: *Relación directa y casi perfecta.*

Tabla VII-5. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra en el Problema 2 apartado b)

Procedimientos	Frecuencia	Porcentaje
20	29	15'0
21	51	26'4
22	18	9'3
23	18	9'3
24	6	3'1
25	4	2'1
26	3	1'6
27	64	33'2
Total	193	100'0

Tabla VII-6. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra al Problema 2 apartado b)

Respuestas	Frecuencia	Porcentaje
Dependencia directa*	139	72'1
Dependencia inversa	4	2'1
Independencia	12	6'2
La relación es muy buena	2	1'0
La relación es dependiente	2	1'0
La relación es inversa con independencia débil	1	0'5
No hay relación	1	0'5
La relación no es buena	1	0'5
No responde	31	16'1
Total	193	100'0

*Respuesta correcta

Tabla VII-7. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra en el Problema 2 apartado c)

Procedimientos	Frecuencia	Porcentaje
No responde	31	16'1
Usa recta de regresión Y sobre X	129	66'8
Usa recta de regresión X sobre Y	4	2'1
Usa ambas rectas de regresión	29	15'0
Total	193	100'0

Tabla VII-8. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra al Problema 2 apartado c)

Respuestas	Frecuencia	Porcentaje
Respuesta correcta	81	42'0
Respuesta incorrecta	52	26'9
Da como respuesta ambas rectas de regresión	29	15'0
No responde	31	16'1
Total	193	100'0

Tabla VII-9. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra a la primera pregunta del Problema 2 apartado d)

Procedimientos	Frecuencia	Porcentaje
No responde	31	16'1
Usa recta de regresión de Y sobre X	142	73'7
Usa recta de regresión de X sobre Y	8	4'1
Usa proporcionalidad	9	4'6
Usa interpolación	1	0'5
No argumenta la respuesta	2	1'0
Total	193	100'0

Tabla VII-10. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra a la primera pregunta del Problema 2 apartado d)

Respuestas	Frecuencia	Porcentaje
Respuesta correcta	5	2'6
Respuesta incorrecta	69	35'7
Expresa la respuesta mediante un número decimal (31'625 bacterias)	88	45'6
No responde	31	16'1
Total	193	100'0

Tabla VII-11. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra en la segunda pregunta del Problema 2 apartado d)

Procedimientos	Frecuencia	Porcentaje
No responde	38	19'7
Usa recta de regresión Y sobre X	139	72'1
Usa recta de regresión X sobre Y	7	3'6
Usa proporcionalidad	7	3'6
No aporta argumentos	2	1'0
Total	193	100'0

Tabla VII-12. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra a la segunda pregunta del problema 2 apartado d)

Respuestas	Frecuencia	Porcentaje
Respuesta incorrecta	154	79'8
Expresa la respuesta mediante un número decimal	1	0'5
No responde	38	19'7
Total	193	100'0

Tabla VII-13. Frecuencia y porcentaje de los procedimientos utilizados por los sujetos de la muestra en el Problema 2 apartado e)

Procedimientos	Frecuencia	Porcentaje
No explicita estrategia	40	20'7
Usa recta de regresión de X sobre Y	97	50'3
Usa recta de regresión Y sobre X	3	1'6
Usa expresión inadecuada de la recta de regresión Y sobre X	42	21'8
Usa proporcionalidad	9	4'6
Usa ecuación $x_2 - x_1 = (a + by_2) - (a + by_1)$	1	0'5
Usa ecuación $x - x_m = S_{xy} (y - x_m) / S_y^2$	1	0'5
Total	193	100'0

Tabla VII-14. Frecuencia y porcentaje de las respuestas de los sujetos de la muestra al Problema 2 apartado e)

Respuestas	Frecuencia	Porcentaje
Respuesta correcta	49	25'4
Respuesta incorrecta	104	53'9
No responde	40	20'7
Total	193	100'0