UNIVERSIDAD DE GRANADA

DEPARTAMENTO DE DIDÁCTICA DE LA MATEMÁTICA


# DIFICULTADES DE ESTUDIANTES UNIVERSITARIOS EN ALGUNOS CONCEPTOS DE DISEÑO EXPERIMENTAL

*Osmar Darío Vera*

Directora: Dra. Carmen Batanero Bernabeau

TRABAJO FIN DE MASTER

Granada, 2008

Quiero agradecer,

A la Dra. Carmen Batanero, por permitirme estar a su lado, aprendiendo, guiándome, sólo impartiéndome frases alentadoras. Su sabiduría junto con sus ideas son infinitas, pero todo lo destina al servicio de los demás. En el peor momento con su sonrisa y practicidad me ha ayudado infinitamente. Al Dr. Juan Godino, quien con su presencia y sabios aportes contribuye a engrandecer al Programa.

A la Dra. Carmen Díaz, quien colaboró con nosotros facilitándonos el conjunto de ítems del cual se seleccionaron los del estudio, así como en la recogida de datos de los estudiantes.

A todos los profesores del Programa Master- Doctorado, que han influido en mí para desarrollar algunas capacidades en la Didáctica de la Matemática.

A las autoridades de la Universidad Nacional de Quilmes, encabezadas por su Vicerrector Prof. Jorge Flores, al Dr. Mario Lozano director del Departamento de Ciencia y Tecnología y a su Vicedirectora Dra. Maria Cristina Taira, finalmente al Dr. Martín Becerra; quienes me han dado la oportunidad de realizar con este pequeño trabajo un aporte a la didáctica de la enseñanza de la Estadística, y han depositado en mi toda su confianza. Espero no defraudarles.
Gracias por el esfuerzo realizado para otorgarme la licencia con goce de haberes que me permitió estar aquí, y poder finalizar esta primera etapa.

El apoyo brindado por la Prof. Cristina Wainmaier ha sido inestimable, a la Prof. Florencia Rembado quien con sus palabras de aliento me ayudo a seguir, cuando decaía.

A la Prof. Cristina Garbarini de Klein, por su entrega personal y profesional.

A Marlen y Pablo, mis vecinos de piso en Granada, quienes han sido muy generosos conmigo.

A mis amigos entrañables de Argentina, quienes a la distancia me han contenido con palabras de aliento y frases sabias.

"Trabajo realizado en el marco del Proyecto SEJ2007-60110, MEC-FEDER"

*A mi hijo Julián, por sus triunfos*

*A mi esposa Miriam, la persona más buena y entrañable*

# ÍNDICE

# INTRODUCCIÓN

En este trabajo llevamos a cabo un estudio exploratorio de evaluación de las posibles dificultades encontradas en una muestra de estudiantes de Psicología sobre las ideas elementales en el Diseño de Experimentos. Esta es una asignatura importante en la formación de especialistas en esta materia, dado el marcado carácter experimental de la disciplina.

Más en concreto analizamos las respuestas de una muestra de 93 estudiantes de segundo curso en la Licenciatura de Psicología a un cuestionario de opciones múltiples completado al finalizar una de las asignaturas dedicadas parcialmente al Diseño Experimental en el plan de estudios. Nos hemos restringido a algunos de los contenidos estadísticos elementales sobre los que se construye el resto de las ideas de diseño experimental y cuya comprensión debe apoyar la aplicación posterior en la vida profesional.

La Memoria se organiza en los siguientes apartados:

- En el primer capítulo se aborda el tema del trabajo, resaltando su importancia y analizando brevemente los objetos matemáticos que serán estudiados en el trabajo. Estos contenidos son una parte de los que pensamos abordar en el futuro. Trataremos además de usarlos en la definición preliminar de la variable objeto de un futuro estudio de evaluación. También hemos explicitado un breve resumen del marco teórico que seguimos. Además se presentan los objetivos del estudio de evaluación.
- Los antecedentes del estudio se muestran en el capitulo 2, donde revisamos algunas investigaciones sobre comprensión del contraste de hipótesis, intervalos de confianza, análisis de varianza, diseño experimental y otras.
- En el capítulo 3 se hace un análisis a priori del cuestionario, con el fin de obtener una primera aproximación a la definición semántica de la variable “comprensión de conceptos básicos del diseño experimental”; la misma se revisará en el futuro para usarlo en la construcción de un instrumento de evaluación.
- El capítulo 4 presenta los resultados de la prueba empírica de 20 ítems de opciones múltiples, analiza los errores encontrados, a fin de establecer hipótesis en una futura investigación. Se determinan las propiedades psicométricas del instrumento: índices

de dificultad y discriminación, dando además una aproximación a la fiabilidad del mismo.

- Finalizamos con las conclusiones finales, referencias y anexos.

El estudio es exploratorio, pues la muestra es de tamaño reducido y el conjunto de ítems utilizado es también limitado. Somos, por tanto, conscientes de las limitaciones de las posibles conclusiones, aunque el trabajo realizado nos ha permitido iniciarnos en la actividad investigadora, ensayar algunos ítems de evaluación que podrían sernos útiles en la construcción de un instrumento más válido y fiable, y familiarizarnos con la didáctica de la estadística a nivel universitario, donde pensamos continuar nuestra investigación en el futuro.

# 1. PROBLEMA DE ESTUDIO Y FUNDAMENTOS

## 1.1. INTRODUCCIÓN

El primer Capítulo de la Memoria se dedica a contextualizar el problema de estudio y proporcionar los fundamentos necesarios para abordarlo.

En lo que sigue delimitamos, en primer lugar, el significado específico que daremos al término Diseño de Experimentos en esta Memoria, justificando su importancia en la preparación de estudiantes de psicología y otras especialidades universitarias. Seguidamente hacemos un resumen muy breve del marco teórico empleado y de parte de la numerosa literatura sobre errores en la práctica e interpretación de la inferencia estadística. Finalmente presentamos los objetivos concretos de nuestro trabajo.

## 1.2. DISEÑO EXPERIMENTAL. CONTENIDOS INCLUIDOS EN ESTE TRABAJO

### 1.2.1 Diseño experimental.

Se conoce como diseño experimental a la metodología estadística orientada para la planificación y análisis de un "experimento", donde la palabra "experimento" se refiere a la concepción y realización de ensayos que verifiquen la validez de las hipótesis establecidas sobre las causas de un determinado problema o que afectan a una cierta variable, objeto de estudio.

Está basado en tres principios básicos: a) *asignación aleatoria* de los individuos a las diferentes condiciones en el experimento, de forma que se puedan determinar el efecto que producen rechazando explicaciones alternativas; b) *replicación*: o repetición del experimento en varios individuos para poder estimar el error experimental y c) *control de variables* que permite reducir la variabilidad del error.

En definitiva, se trata de una metodología basada en técnicas estadísticas cuyo objetivo es ayudar al investigador a: seleccionar la mejor estrategia para obtener la información buscada con el mínimo coste y evaluar los resultados experimentales obtenidos con la mayor fiabilidad posible (Ferré y Rius, 2008).

### 1.2.2 Variables y sus tipos

En el Diseño experimental, se consideran diferentes tipos de variables (Box, Hunter y Hunter, 1989):

- Las variables *dependientes* o *respuesta* que son las que se desea evaluar en el experimento y usualmente son cuantitativas (aunque podrían no serlo). También se contempla que la variable dependiente sea un vector formado por un conjunto de variables. En nuestro trabajo nos interesamos por los diseños de experimentos con una sóla variable dependiente cuantitativa.
- Las variables *independientes* o *factores,* que usualmente son cualitativas o bien tienen un conjunto finito de valores que se denominan niveles del factor. En el diseño de experimentos estamos interesados en determinar en qué medida los *factores*, podrían producir cambios (que se precisan mediante lo que denomina *efecto*) sobre las variable dependiente o respuesta debidos a la realización del experimento.

En cada experimento concreto se puede tener uno o varios factores, y además estos se pueden dividir en varios tipos:

- *Factor de efectos fijos:* Son aquello que incluyen en el estudio todos los niveles posibles del factor. Por ejemplo, al estudiar la diferencia de rendimiento escolar en niños y niñas, la variable repuesta es el rendimiento, el factor es el género y los dos niveles posibles “niño” y “niña” se incluyen en el estudio. Siempre que los experimentos de varios factores de efectos fijos se realicen combinando todos los niveles posibles, se hablará de *diseños completos.*
- *Factor de efectos aleatorios:* Se presentan cuando el factor tiene un conjunto muy amplio de niveles, que no se pueden incluir en su totalidad, de modo que se toma una muestra aleatoria en el estudio. En el ejemplo anterior podríamos incluir un segundo factor “colegio” y tomar una muestra aleatoria de colegios en una ciudad.

Otra división posible es factores intersujetos y factores intrasujeto:

- *Factor intrasujeto* es aquél en que los varios niveles se toman en la misma unidad estadística. Por ejemplo, podríamos tomar la tensión arterial (variable respuesta) a un conjunto de personas antes y después de un tratamiento médico. El factor intrasujeto sería el tiempo en que se toma la tensión con dos niveles (antes y después del tratamiento)
- *Factor intersujeto*, es cuando los diversos niveles se contemplan en diferentes unidades estadísticas; en el ejemplo anterior seria el caso de diferenciar el factor género (hombre/mujer).

Además, pueden aparecer variables que, no siendo de interés al investigador, afecten a la variable dependiente; estas variables se denominan *extrañas* y el investigador ha de tratar de controlarlas para que no afecten al experimento. La finalidad del diseño experimental es poder comprobar las hipótesis objeto de estudio, controlando el error aleatorio y las variables extrañas.

Las variables extrañas pueden controlarse manteniéndolas constantes o por medio de la asignación aleatoria de los sujetos experimentales a cada nivel de cada factor, pero esto suele ser difícil. Otra posibilidad es medir las variables extrañas e incluirlas como parte del análisis estadístico; en este caso se suelen denominar *covariables*. Por ejemplo, si se quieren comparar dos medicamentos y sospechamos que la edad puede afectar a su efecto sobre los pacientes, podríamos incluir la edad como covariable en el estudio cuando no se pueda asignar aleatoriamente los medicamentos a los pacientes teniendo en cuenta la edad.

### 1.2.3 Técnicas estadísticas del diseño de experimentos

En el diseño de experimentos se utilizan técnicas estadísticas para determinar en que medida los *factores*, podrían producir cambios (que se precisan mediante lo que denomina *efecto*) sobre las variables dependiente ó respuesta como consecuencia de la realización del experimento. En especial el conjunto de técnicas y objetos estadísticos que se conoce como Análisis de la Varianza (también encontrado a menudo en la literatura mediante la sigla ANOVA), es el más utilizado en el Diseño de Experimentos.

El objetivo general del ANOVA es comprobar si $k$ muestras provienen o no de

poblaciones con la misma media, cuando se dan ciertas condiciones establecidas. Los procedimientos que se aplican permiten descomponer la variabilidad total en variabilidades aportadas por los distintos factores puestos en juego, además de sus interacciones mutuas.

El modelo más simple es el conocido como *análisis de varianza de un factor, con efectos fijos:* Supongamos que disponemos de $k$ grupos, de modo que en el grupo $i$-ésimo hay $n_i$ observaciones. Sea $y_{ij}$ la $j$-ésima observación en el grupo $i$ (donde el subíndice $j$ varía entre 1 y $n_i$ , mientras que $i$ lo hace entre 1 y $k$). Todos los elementos del mismo grupo $i$ se dice que están "sujetos al tratamiento $i$", terminología que proviene de las primitivas aplicaciones del análisis de varianza en el trabajo experimental, sobre todo, en agricultura. Haremos las siguientes hipótesis (Dunn y Clark, 1997):

- *Independencia***:** Cada una de las observaciones es independiente de las demás.
- *Linealidad:* Cada uno de los valores observados $y_{ij}$ puede descomponerse en la forma:

(1) $$y_{ij}=\mu+\alpha_i+\varepsilon_{ij} \qquad \begin{cases} i = 1,2,....,k \\ j = 1,2,....,n_i \end{cases}$$

con la condición de que $\Sigma\alpha_i=0$. Es decir la suma de las diferencias entre la media de cada grupo y la media global es igual a cero. En la expresión (1) $\mu$ es una constante, que representa la media global del conjunto de las $k$ muestras; $\alpha_i$ es una constante dentro del grupo $i$ y representa la diferencia de la media de este grupo, anotada $\mu_i$, con la media del grupo total. El valor $\alpha_i$ se suele conocer como "efecto debido al tratamiento $i$". Por último, $\varepsilon_{ij}$ es una variable aleatoria, a través de esta $y_{ij}$ hereda la aleatoriedad presente en (1), pues $\varepsilon_{ij}$ tiene una distribución normal $N(0, \sigma^2)$.

- *Homocedasticidad*. Se supone una varianza común $\sigma^2$ para todos los grupos. Por tanto, podemos decir que la variable aleatoria $y_{ij}$ sigue una distribución normal $N(\mu+\alpha_i, \sigma^2)$.

El análisis de la varianza (tomando como ejemplo el modelo más simple), consiste en la realización de un contraste estadístico para decidir entre las dos hipótesis siguientes:

$H_0 \equiv \alpha_1=\alpha_2=.....=\alpha_k=0$ $H_1 \equiv$ al menos un $\alpha_i$ es diferente de cero.

Dicho de otro modo, bajo la hipótesis nula, todos los grupos provienen de poblaciones con igual media, y, en consecuencia, todos los tratamientos producen un efecto nulo. Bajo la hipótesis alternativa, al menos un tratamiento produce un efecto no nulo, por lo que algunas de las poblaciones tendrán diferentes medias.

Una vez rechazada la hipótesis se calculan intervalos de confianza para los efectos, usando fórmulas especiales que tienen en cuenta el nivel de significación global, permitiendo reconocer los efectos no nulos de los que lo son. En la literatura del diseño experimental se suele encontrar este estudio bajo la denominación: análisis a posteriori.

Otros modelos de análisis de varianza consideran nuevos factores o bien alguno de ellos aleatorios. Se aplican pruebas de hipótesis básicas sobre estos modelos, éstos se centran en contrastar si los factores realmente logran alterar los resultados de los experimentos al fijarlos en sus distintos niveles. Por ejemplo, si considerásemos dos factores, se obtiene el modelo siguiente:

$$y_{ijk}=\mu+\alpha_i+\beta_j+ (\alpha\beta)_{ij}+\varepsilon_{ijk}$$

donde $\mu$ es una constante que representa la media global, $\alpha_i$ representa el efecto del nivel $i$ del factor $A$, $\beta_j$ representa el efecto de nivel $j$ del factor $B$, $(\alpha\beta)_{ij}$ representa la interaccion del nivel $i$ del factor $A$ con el $j$ del factor $B$, $\varepsilon_{ijk}$ es el efecto del azar, término que dota al modelo de aleatoriedad, y se considera $N(0, \sigma^2)$.Suponemos, además, $\Sigma\alpha_i=\Sigma\beta_j=\Sigma(\alpha\beta)_{ij}=0$.

Llamamos *interacción* entre dos variables al efecto de una de ellas cuando éste depende del nivel de la otra. Así en el ejemplo anterior, existiría interacción si la mejora de puntuación de un curso al superior sólo se verificara en uno de los dos sexos.

### 1.2.4 Delimitación de los contenidos del estudio

En este trabajo nos centraremos en las herramientas estadísticas básicas del diseño experimental, más concretamente en el análisis de varianza, sus supuestos, modelos e interpretación, así como de los prerrequisitos para llevarlo a cabo. Por

tanto, tratamos de llevar a cabo una prueba inicial de evaluación que aporte alguna información de la comprensión de un grupo de estudiantes sobre los intervalos de confianza, contraste de hipótesis; ANOVA de efectos fijos: modelos unifactorial y bifactorial y medidas repetidas. Estos contenidos se delimitan con más detalle en el análisis a priori del cuestionario inicial.

## 1.3. DISEÑO EXPERIMENTAL EN LOS CURSOS UNIVERSITARIOS

El diseño experimental constituye una herramienta fundamental en el trabajo de científicos y profesionales, ayudándoles en la planificación de experimentos eficaces y a la determinación de factores que incidan en las variables dependientes de interés con objeto de reducir la variabilidad no explicada en los procesos de producción ó en fenómenos naturales, e incidir en la predicción ó control de los mismos. Es por ello que una asignatura con estos contenidos se incluye en muchas especialidades universitarias, como la psicología, la ingeniería o las ciencias relacionadas con la química y con la biología**,** así como en los cursos de postgrado o de formación profesional.

## 1.4. MARCO TEÓRICO

Entendemos, de acuerdo con Godino, Batanero, y Font (2007) que, para especificar los problemas de investigación en didáctica de las matemáticas, así como para afrontarlos, es necesaria una estrategia metodológica y pensar teorías como herramientas. Por ello, creemos que las nociones que componen el modelo teórico propuesto por Godino y colaboradores (Godino y Batanero, 1994; Godino, 2003) es el mas adecuado pora nuestra investigación y nuestras investigaciones futuras.

Dicho modelo, denominado de “enfoque ontosemiótico” (EOS) de la cognición matemática, nos da un punto de vista pragmático-antropológico, partiendo del papel clave que tiene la actividad de resolución de problemas (sistemas de practicas operativas y discursivas). Esta herramienta, que viene creciendo desde hace más de 15 años, comprende varias etapas, durante las cuales se ha ido perfeccionado.

Las etapas de las que hablamos son tres: en la primera han ido precisando y

desarrollando las nociones de "significado institucional y personal de un objeto matemático" (entendidos ambos en términos de sistemas de prácticas en las que el objeto es determinante para su realización) y su relación con la noción de comprensión. (Godino, Batanero y Font, 2007). Creemos que la distinción de estas dos nociones es fundamental para nuestra investigación, ya que al observar las respuestas de los alumnos en el Instrumento de Evaluación que hemos elaborado, estaremos trabajando sobre los significados personales (sistemas de prácticas adquiridos por un sujeto). Luego de un período de instrucción, el resultado será exitoso, si los alumnos han logrado acoplarse a los significados instituciones (prácticas referidas al objeto, emanadas desde el seno de una institución, las cuales podrían venir representadas ya sea por material bibliográfico, apuntes del profesor, etc.). Siguiendo esta noción, y como el fin último es lograr el aprendizaje del alumno, diremos que lo ha logrado si se han acoplado los significados personales con los institucionales. En una segunda etapa han pretendido elaborar una ontología suficientemente rica para describir la actividad matemática y los procesos de comunicación de sus "producciones". En esta etapa han tratado de progresar en el desarrollo de una ontología y una semiótica específica que estudie los procesos de interpretación de los sistemas de signos matemáticos puestos en juego en la interacción didáctica, y han ampliado las investigaciones realizadas sobre los significados institucionales y personales completando también la idea de función semiótica y de la ontología matemática asociada.

Finalmente, y como última etapa, en la cual se encuentra la teoría en este momento, proponen distinguir en un proceso de instrucción matemática seis dimensiones, siendo posible cada una ser modelizada como un proceso estocástico con sus respectivos espacios de estado y trayectorias: epistémica (relativa al conocimiento institucional), docente (funciones del profesor), discente (funciones del estudiante), mediacional (relativa al uso de recursos instruccionales), cognitiva (génesis de significados personales) y emocional (que da cuenta de las actitudes, emociones, etc. de los estudiantes ante el estudio de las matemáticas). Los constructos teóricos elaborados durante estos tres periodos constituyen el modelo ontológico semiótico de la cognición matemática.

Este marco teórico podría complementar y enriquecer el análisis que hacemos de los significados personales logrados, analizando las respuestas de los alumnos dadas a los ítems en nuestra evaluación, pues el análisis de la noción de significado,

desde un punto de vista didáctico puede ayudar a comprender las relaciones entre las distintas formulaciones teóricas en esta disciplina y permitir estudiar bajo una nueva perspectiva las cuestiones de investigación, particularmente las referidas a la evaluación de los conocimientos.

Nos resulta práctico a nuestra investigación la clasificación que este marco realiza para los significados institucionales y personales (Figura 1.1), a saber:

### 1.4.1 Tipos de significados Institucionales

*Implementado:* en un proceso de estudio específico es el sistema de prácticas efectivamente implementadas por el docente, en nuestra investigación se trata de las prácticas relativas a las nociones básicas del diseño experimental propuestas por el docente en el seno de la institución. Este significado no será evaluado en nuestra investigación.

*Evaluado:* el subsistema de prácticas que utiliza el docente para evaluar los aprendizajes. En nuestra investigación, se trata de los objetos elementales del diseño experimental que ha evaluado el docente a través del cuestionario elaborado.

Figura 1.1. Sistemas de prácticas operativas y discursivas ligadas a campos o tipos de problemas TSS (Teoría de los Significados Sistémicos)

*Pretendido:* sistema de prácticas incluidas en la planificación del proceso de estudio. En nuestra investigación, hay otras prácticas incluidas en la planificación que no se han incluido, por escaparse a nuestros objetivos.

*Referencial:* sistema de prácticas que se usa como referencia para elaborar el significado pretendido. En una institución de enseñanza concreta este significado de referencia será una parte del significado holístico del objeto matemático. La determinación de dicho significado global requiere realizar un estudio histórico – epistemológico sobre el origen y evolución del objeto en cuestión, así como tener en cuenta la diversidad de contextos de uso donde se pone en juego dicho objeto. En nuestra investigación, por limitaciones de tiempo no se ha realizado este estudio, se incluirá en investigaciones futuras. Nuestro Instrumento de evaluación se basó en la experiencia previa y en errores de comprensión encontrados de algunos objetos relacionados directamente con el diseño experimental.

### 1.4.2 Tipos de Significados personales

*Global:* corresponde a la totalidad del sistema de prácticas personales que son capaces de manifestar potencial los sujetos, relativos a un objeto matemático.

*Declarado*: da cuenta de las prácticas efectivamente expresadas a propósito de las pruebas de evaluación propuestas, incluyendo tanto las correctas como las incorrectas desde el punto de vista institucional. Este es el significado que analizamos en nuestra investigación, a través del cuestionario de evaluación implementado.

*Logrado*: corresponde a las prácticas manifestadas que son conformes con la pauta institucional establecida. En el análisis del cambio de los significados personales que tiene lugar en un proceso de estudio interesará tener en cuenta los significados iniciales o previos de los estudiantes y los que finalmente alcancen. Este significado en nuestra investigación lo analizamos de acuerdo con el éxito manifestado en la elección de las respuestas a los ítems. Somos concientes que con esta forma de análisis solamente no seremos capaces de dar una respuesta de los significados personales logrados por nuestros alumnos, y de observar la comprensión de los objetos matemáticos puestos en juego, pero si es un primer paso para obtenerlo.

Esta clasificación la podemos apreciar gráficamente a través de la Figura 1.1, donde en la parte central los autores, indican las relaciones dialécticas entre enseñanza y aprendizaje, lo cual da como consecuencia el acoplamiento (que ya indicamos) progresivo entre los significados personales e institucionales.

### 1.4.3 Objetos intervinientes y emergentes de los sistemas de prácticas

En los sistemas de prácticas matemáticas, y como consecuencia de éstas, siempre intervienen objetos matemáticos que necesitamos evocar, los mismos emergen, ya sea en forma de texto, oral, gráfica, y también a través de gestos. Además de los sistemas de prácticas matemáticas, aparecen otros objetos que dan cuenta de su organización y estructura. Podrán emerger, objetos personales ó institucionales dependiendo en que sitio se llevan a cabo tales prácticas (si es de tipo institucional, será por que aparecen desde el seno de una institución). El EOS propone la siguiente tipología de objetos matemáticos primarios:

- *Lenguaje* (términos, expresiones, notaciones, gráficos, ...) en sus diversos registros (escrito, oral, gestual, ...). Para nuestra investigación tales objetos los tenemos presentes atravesándola, los mismos en cuanto a término se refiere pueden ser: intervalo de confianza, contraste de hipótesis, ANOVA, modelos, supuestos, etc; en cuanto a expresiones nos podemos referir a: aleatoriedad de la toma de datos para el diseño, homocedasticidad de la muestra, interpretación de una tabla de análisis de la varianza; en cuanto a notaciones, podemos hallar muchos ejemplos en el apartado 1.2.. Merece la pena nos detengamos especialmente en el uso del lenguaje gráfico, ya que hemos incluido ítems donde se pide analizar tablas (Ítems 13, Ítem 16) y responder de acuerdo a la observación de gráficos (Ítem 17).
- *Situaciones-problemas* (aplicaciones extra-matemáticas, ejercicios, ...). Son múltiples e indispensables estos objetos en nuestra investigación, ya que las aplicaciones se refieren en su mayoría a ejercicios inherentes a situaciones provenientes de la Psicología. Debemos notar aquí que si variamos el contexto de aplicación, emergerán otras situaciones y fenómenos equivalentes.
- *Conceptos- definición* (introducidos mediante definiciones o descripciones), ejemplos para nuestra investigación tendremos: parámetro muestral, parámetro poblacional, definiciones de error Tipo I, error Tipo II en un contraste de hipótesis,

descripción de la composición de una tabla ANOVA, etc.

- *Proposiciones* (teoremas y propiedades), por ejemplo el teorema central del límite.
- *Procedimientos* (algoritmos, operaciones, técnicas de cálculo,...), en nuestra investigación encontramos algoritmos con sus operaciones para la determinación de los intervalos de confianza, algoritmos para la aplicación de un contraste de hipótesis, las técnicas para la determinación de los elementos que conforman una tabla ANOVA, etc.
- *Argumentos* (enunciados usados para validar o explicar las proposiciones y procedimientos, deductivos o de otro tipo,...). En nuestra investigación, el alumno debe argumentar para diferenciar la respuesta correcta de los distractores. De lo correcto ó incorrecto de los argumentos que utilice para validar o explicar las proposiciones propuestas dependerá el éxito de su ítem.

Creemos muy importante esta clasificación en los seis tipos de entidades primarias postuladas, pues amplían la tradicional distinción entre entidades conceptuales y procedimentales, al considerarlas insuficientes para describir los objetos intervinientes y emergentes de la actividad matemática.

## 1.5. OBJETIVOS DEL ESTUDIO DE EVALUACIÓN

Como hemos indicado, el trabajo presentado en esta Memoria es un primer paso en un estudio más completo que pensamos abordar en nuestra tesis doctoral y que consistirá en la construcción de un instrumento de evaluación. En dicho trabajo futuro se pretende seguir el proceso de construcción seguido por Díaz (2004, 2007) y Olivo (En prensa) quienes elaboraron un cuestionario de comprensión de la probabilidad condicional y de los intervalos de confianza respectivamente. Ambos autores forman parte de nuestro grupo de investigación y han seguido las normas psicométricas de APA, AERA y NCME (1999) en la construcción de sus instrumentos.

Las fases principales recomendadas en estas normas, así como en Martínez Arias (1995), en el proceso de elaboración de un cuestionario serían: a) definir la variable a partir del análisis de contenido, incluyendo la construcción de una tabla de especificaciones del contenido que se desea evaluar mediante el instrumento; b) recopilar, poner a prueba y depurar un conjunto de ítems que puedan formar parte

del mismo; c) seleccionar los ítems que conformarán la versión definitiva del cuestionario, para lo que se puede usar las pruebas empíricas de los ítems así como la valoración de los mismos mediante juicio de expertos; d) estimar las características finales del cuestionario construido y los ítems que lo forman, esto es, acumular evidencias de fiabilidad y validez del cuestionario, y calcular los índices de dificultad y discriminación de los ítems.

Estas fases se suelen solapar en diversos momentos, pues el proceso de elaboración de un cuestionario no es secuencial sino cíclico, ya que en cada una de las etapas es necesario revisar sus resultados y a veces se requiere volver a las etapas anteriores. Por ello, en esta Memoria presentamos los resultados de los primeros pasos en nuestro estudio global, abordando los dos primeros puntos descritos, esto es, tratamos de iniciar la definición de la variable objeto de medición y al mismo tiempo iniciar la recopilación y pruebas de ítems que nos puedan ser útiles en el futuro.

A continuación describimos los dos objetivos de nuestro trabajo.

Objetivo 1. *Definición preliminar de la variable "Comprensión de objetos estadísticos básicos del diseño experimental".*

Dada la amplitud de contenidos que pueden englobarse dentro del título "Diseño experimental" se hace necesario limitar el tema de investigación para poder abarcar la investigación en un periodo razonable de tiempo. Se trataría de definir el significado institucional evaluado en la investigación. Una primera limitación ya señalada en nuestro estudio es centrarnos en contenidos específicamente estadísticos, y, entre ellos, los dedicados al análisis de los datos obtenidos en los diseños experimentales.

Un punto ya señalado por Díaz (2004, 2007) y Olivo (En prensa), así como por los textos de psicometría, es la diferencia entre constructo y variable. El constructo es el rasgo psicológico en que estamos interesados en la evaluación (en nuestro caso la comprensión de conceptos elementales de diseño experimental); como constructo psicológico no lo podemos observar ni el sujeto conoce la cantidad que tiene de este constructo; tampoco se puede medir directamente sino indirectamente por medio de una variable relacionada que es observable (León y Montero, 2002).

En nuestro caso, la variable que queremos observar es una puntuación

obtenida a partir de las respuestas de los alumnos a los que se pasará el cuestionario en los ítems que los componen; variable que también se puede descomponer como un vector formado por el conjunto de respuestas del estudiante a cada uno de los ítems, lo que nos permitirá obtener un indicador cuantitativo (puntuación en el cuestionario) y una serie de indicadores cualitativos (conjunto de respuestas específicas) de la comprensión de cada estudiante.

Es claro que la definición de la variable relacionada con el constructo no es única, y de ahí la dificultad de la construcción de un cuestionario. El proceso tampoco es sencillo y requiere un gran nivel de reflexión sobre cuáles serían los componentes del constructo de interés (cómo descomponer la comprensión de ideas elementales del diseño experimental) y cómo medir cada uno de estos componentes. Habría que comenzar en primer lugar por fijar cuáles serán los objetos estadísticos elementales que contemplaremos como componentes de nuestra variable y para cada uno de ellos, a su vez, analizar cuáles serían sus constituyentes esenciales.

Con este proceso se iniciaría la definición semántica de la variable objeto de medición; ésta consiste en la descomposición de la variable en unidades de contenido semántico, que podrían reflejarse en la elaboración de una tabla de especificaciones del contenido del instrumento (Martínez Arias, 1995).

Este objetivo se ha abordado inicialmente en la Memoria fijando la lista de contenidos a incluir en la variable objeto de medición a partir de nuestra propia experiencia docente y de la consulta de algunos libros elementales de diseño experimental y se ha descrito en el apartado 1.2..

Contrastaremos este contenido pretendido con el contenido cubierto por los ítems que ponemos a prueba para revisar la definición preliminar de la variable en la sección 3.3.. Pensamos depurar esta definición en el futuro, mediante análisis de otros libros de texto y consultas a expertos en diseño experimental.

Objetivo 2. *Puesta a prueba de algunos ítems que podrían constituir parte de un posible cuestionario y que evalúen los contenidos incluidos en la variable.*

Además de iniciar la definición de la variable, nos hemos interesado por comenzar a recopilar posibles ítems de utilidad en nuestro estudio y recoger algunos indicadores de su utilidad para nuestros fines y de la comprensión que los estudiantes manifiestan en sus respuestas a los mismos. Estos ítems han sido seleccionados de un conjunto más amplio que se utilizan en las evaluaciones de una

asignatura, que se describirá en la sección 3.2., y que cubren los contenidos estadísticos en los que estamos interesados.

En la selección de ítems hemos tratado de apegarnos a los criterios de Osterlind (1989), como son la adecuación del ítem particular al contenido que se pretende evaluar, la independencia de las respuestas entre diferentes ítems y la adecuación del tipo de formato. También se tuvo en cuenta que el número de ítems cubra el contenido pero que al mismo tiempo los alumnos tengan tiempo suficiente para responder al cuestionario completo.

Ello nos llevó a usar ítems en un formato de opción múltiple, donde se proporciona una pregunta y se pide al examinado elegir una entre tres alternativas, de las cuáles sólo una es correcta. Las alternativas tratan de ver errores frecuentes en los estudiantes, algunos de ellos descritos en investigaciones previas. Tiene la ventaje de requerir menor esfuerzo; además los estudiantes a los que se pasó la prueba están familiarizados con este formato, pues en sus exámenes se suele utilizar en todas las asignaturas.

En el Capítulo 3 se presentan los ítems puestos a prueba y su análisis a priori, obteniendo al finalizar una primera revisión de la definición de la variable objeto de medición.

También presentaremos los resultados de las pruebas de estos ítems en una muestra de 93 estudiantes de psicología que habían seguido un curso que incluye todos los contenidos evaluados en los ítems y cuyas características y proceso de estudio seguido se describen igualmente.

Finalmente, aunque no es uno de nuestros objetivos principales, el estudio de los ítems permite informar sobre algunos errores y dificultades de los estudiantes, que permitirán posteriormente construir hipótesis sobre la comprensión de los objetos estadísticos incluidos en las variables que se incluyan en investigaciones futuras. Con ello informaremos del significado personal logrado por los estudiantes.

# 2. ANTECEDENTES

## 2.1. INTRODUCCIÓN

Las investigaciones que presentamos en este apartado están centradas sobre la enseñanza y el aprendizaje de los objetos estadísticos que se incluyen en el cuestionario de evaluación y que son básicos del diseño de experimentos. No todos estos objetos han sido analizados en la investigación didáctica con la misma intensidad. Específicamente, ya que una parte de nuestra investigación consiste en la selección de ítems para explicar la comprensión de algunos objetos matemáticos explicitados en el capítulo 1, serán más desarrolladas aquellas que han utilizado ítems que podrían sernos útiles.

En lo que sigue comenzamos analizando las investigaciones previas sobre el objeto contraste de hipótesis que es el que ha recibido más atención, por ser la herramienta básica e ineludible para la comprensión y análisis de los diseños de experimentos. Seguidamente tratamos el objeto intervalo de confianza, para finalmente explicitar otras investigaciones que se ocupan de objetos estadísticos que son prerrequisitos para los objetos sobre los que se centra la investigación. Nos ocupamos a continuación de las pocas investigaciones que han tratado el diseño experimental y finalizamos con otras, relacionadas con las distribuciones muestrales que son prerrequisito para el diseño experimental.

## 2.2. INVESTIGACIONES SOBRE COMPRENSIÓN DEL CONTRASTE DE HIPÓTESIS

Los contrastes de hipótesis son herramientas básicas en el diseño de experimentos, tanto en la confirmación de los modelos empleados, como en la demostración de la existencia de efectos. Sin embargo, su uso en las ciencias empíricas en general y en particular en la psicología ha sido muy objetado casi desde su creación, pues Yates (1951) avisó que la mayoría de los investigadores y científicos se preocupaban básicamente en los resultados obtenidos de sus pruebas de hipótesis de sus investigaciones, sin prestar atención a la estimación de los efectos sobre la variable respuesta aportados por la presencia de los factores en el modelo.

Una investigación que estudia a fondo el contraste de hipótesis es la de Vallecillos (1994), para la que construye un cuestionario de evaluación y describe numerosos errores en una amplia muestra de estudiantes. La autora clasifica estos errores en varios apartados:

- *Contraste de hipótesis como problema de decisión:* Los estudiantes intercambian la hipótesis nula y alternativa. Este error, según la autora, está indicando la falta de comprensión en el alumno de la diferencia entre la demostración matemática de una hipótesis y el contraste estadístico de hipótesis. También se confunden los errores tipo I y tipo II o a veces se consideran complementarios o incompatibles
- *Interpretación de las probabilidades de error y sus relaciones:* Los estudiantes no aprecian la posibilidad de ocurrencia de los errores de tipo I y II, también confunden las dos probabilidades condicionales que intervienen en la definición del nivel de significación, es decir se confunden $\alpha = P(\text{rechazar } H_0 / H_0 \text{ cierta})$ con $P(H_0 \text{ cierta/ se ha rechazado } H_0)$ lo cual sería una interpretación bayesiana, correspondiente a una probabilidad a posteriori. Los alumnos realizan otras explicaciones erróneas sobre el significado de $\alpha$ , llegando a suprimir la condición en la probabilidad ó confundiéndola con su probabilidad complementaria. Vallecillos (1994) indica que este es el error más citado en las investigaciones didácticas sobre el tema.
- *Nivel de significación y potencia como riesgos del decisor*: Se confunde el nivel de significación con la probabilidad de obtener un resultado correcto, lo que indica en general la confusión de resultado significativo con resultado correcto. También suelen pensar que ambos riesgos están predeterminados por el tamaño de la muestra
- *Parámetro y contraste de hipótesis:* Los estudiantes no dan cuenta de la variabilidad de la distribución muestral del estadístico, así como tampoco se reconoce la dependencia de dicha distribución muestral respecto al parámetro poblacional.
- *Nivel de significación y el criterio de decisión*: Se observa una confusión entre región de rechazo y de aceptación para la hipótesis nula/ alternativa, así como de cómo hay que construirlas al tener un test unilateral ó bilateral.
- *Nivel de significación y la distribución del estadístico*: Falta de apreciación de que la hipótesis alternativa determina, junto con el nivel de significación, la región crítica, y que un mayor nivel de significación da una menor área determinada por la

función de densidad de la distribución del estadístico en el muestreo y la región crítica bajo la hipótesis nula cierta.

- *Interpretación de resultados:* Confunden el nivel de significación con la probabilidad de ocurrencia de un resultado significativo, así como un resultado significativo lo ven como el resultado que corrobora la hipótesis nula. También afirman que un resultado significativo ocurre con un error cometido en el proceso de contraste.
- *Lógica global del proceso:* Una confusión frecuente es tomar como hipótesis nula aquella que se desea probar. Confunden el tipo de prueba de la hipótesis que proporciona un contraste y consideran que el test a ocupar para una prueba depende de que los resultados sean conocidos antes de la realización del experimento.

Las controversias sobre el papel de los contrastes estadísticos de hipótesis en la investigación experimental, han sido largamente estudiadas por muchos autores y se resumen entre otras muchas publicaciones en Harlow, Mulaik y Steiger (1997) y Batanero (2000). Batanero en su trabajo describe la lógica de los contrastes de hipótesis desde la filosofía de Fisher y Neyman-Pearson y analiza las interpretaciones erróneas de los conceptos que intervienen en la lógica del contraste de hipótesis. Son analizadas también importantes cambios en las políticas editoriales referidas al uso del contraste de hipótesis, en diversas revistas científicas. Estas políticas son impulsadas desde la Americam Psychological Association y otras asociaciones científicas, quienes recomiendan a la hora de publicar sus investigaciones, no solo seleccionen los artículos a través de los resultados obtenidos en los contrastes de hipótesis (usando el p-valor), sino se tomen en cuenta las estimaciones tanto de los efectos como de los parámetros que están en juego usando intervalos de confianza.

Batanero (2000) resume la investigación de Vallecillos (1994) y otras previas como las de Birnbaum (1982) , Falk (1986), Pollard y Richardson (1987), en las cuales el error más frecuente es intercambiar los dos términos de probabilidad condicional en cuanto a nivel de significación se refiere, ó sea interpretarlo como la probabilidad de que la hipótesis nula sea cierta si hemos tomado la decisión de rechazarla. Además se listan una serie de interpretaciones erróneas del nivel de significación y el valor *p*.

Además, de los ya señalados en relación a la investigación de Vallecillos (1994), se confunde significación estadística y significación desde el punto de vista práctico. Skipper, Guenter y Nass (1970) advierten sobre el uso abusivo en la literatura de investigación de valores de significación .05, .01 y .001 en todo tipo de problemas, ya que esto puede traer como consecuencia que los resultados de investigación que se publiquen solo se seleccionen en base al valor p obtenido, y no así a la diferencia o efecto encontrado cuando tiene una magnitud importante. Por otro lado, en algunas ocasiones interesa mejor controlar el Error Tipo II. Avisa que lo optimo es buscar un equilibrio entre la potencia del contraste, el error Tipo II, para acomodar el valor del error Tipo I más adecuado a ese estudio.

Otro error frecuente es pensar que el valor p es la probabilidad de que el resultado se deba al azar, siendo el valor p es la probabilidad de obtener el resultado particular u otro más extremo cuando la hipótesis nula es cierta y no hay otros factores posibles que influencien el resultado.

Además existe la creencia en la conservación del valor del nivel de significación cuando se realizan contrastes consecutivos en el mismo conjunto de datos, lo que produce el problema de las comparaciones múltiples. Si por ejemplo hacemos 100 contrastes sobre el mismo conjunto de datos con nivel de significación .05, habrá que esperar que 5 de las 100 pruebas sean significativas por azar (Moses, 1992).

Otro error citado en Batanero (2000) consiste en la confusión entre las diferentes hipótesis que aparecen en un contraste: Hipótesis sustantiva o teórica, hipótesis de investigación, hipótesis alternativa y nula. Ello hace que, rechazada la hipótesis nula, los investigadores piensen que han demostrado su hipótesis de investigación, aún cuando ella sea mucho más amplia que la hipótesis alternativa. La explicación dada por Batanero es que, por un lado, las personas no diferencian el razonamiento inductivo del deductivo y por otro tienen una fe excesiva en el resultado de los cálculos matemáticos.

## 2.3. INVESTIGACIONES SOBRE COMPRENSIÓN DEL INTERVALO DE CONFIANZA

En relación al intervalo de confianza las investigaciones didácticas no son tan numerosas, aunque están aumentando recientemente. Cumming, William y Fidler (2004) estudian sistemáticamente los errores en la interpretación de intervalos de

confianza de los investigadores que han publicado artículos en los que usan dichos intervalos en revistas internacionales. Uno de los puntos en los que se interesaron fue la predicción de cómo sería un intervalo de confianza si se replica el experimento, observando que muchos investigadores piensan (erróneamente) que hay una alta probabilidad de que el parámetro caiga de nuevo en el intervalo de confianza original, cuando esta probabilidad es menor de lo supuesto por estos investigadores. Belia, Fidler y Cumming (2005) pidieron a algunos investigadores mediante una encuesta hecha por correo electrónico que juzgaran cuando eran significativamente diferentes las medias de dos grupos independientes a partir de la interpretación de los intervalos de confianza para dichas medias, observando que sólo unos cuantos sujetos se acercaron a la respuesta correcta.

En un estudio con alumnos universitarios Fidler y Cumming (2005) tratan de ver si relacionan los intervalos con los contrastes de hipótesis y si saben interpretar su significado práctico en artículos publicados en revistas científicas. Para ello les preguntan si los intervalos y el valor p en los contrastes proporcionan o no evidencia empírica a favor de la hipótesis nula o de la alternativa. Observa que el número de interpretaciones incorrectas es mayor en los contrastes de hipótesis aunque alrededor de un 20 % de alumnos en su estudio también interpretan incorrectamente un intervalo de confianza. Por ejemplo, algunos alumnos suponen que el intervalo se refiere a los valores de la media de la muestra o a los valores individuales de la variable.

Behar (2001) prepara un cuestionario que aplica un grupo de expertos en estadística, y un grupo de estudiantes universitarios para valorar la comprensión del intervalo de confianza. Entre otras dificultades señala falta de comprensión de la manera como se relacionan los distintos objetos que emergen asociados con un intervalo de confianza, en especial el ancho del intervalo y el nivel de confianza. Una buena proporción de los participantes de ambos grupos, no interpretan correctamente el coeficiente de confianza, pues suponen que da la probabilidad de que el parámetro se encuentre en el intervalo, mientras que la verdadera definición es el porcentaje de intervalos calculados a partir de muestras de igual tamaño en la población que contiene al parámetro. Tampoco se comprende la utilidad de los intervalos en la toma de decisiones.

Terán (2006), utilizando ideas del marco teórico de Godino (2003) investiga el significado asignado a los intervalos de confianza por dos estudiantes cuando trabajan resolviendo problemas con el ordenador en un estudio de tipo cualitativo. Su análisis de

las entrevistas y el trabajo de los alumnos muestran como ellos usan propiedades y conceptos, argumentos y procedimiento, esto es, los diversos elementos de significado definidos por Godino.

Un trabajo importante es el de delMas, Garfield y Chance (2004), quienes tuvieron un proyecto para desarrollar materiales didácticos relacionados con la inferencia, entre otros, en relación al intervalo de confianza. Indican que los estudiantes han de tener los siguientes conocimientos previos para comprender el intervalo de confianza: Población, muestra, parámetro, estadístico, variabilidad del muestreo/ error del muestreo, error estándar, distribución muestral. Han de entender que las estimaciones provienen de muestras, y el verdadero valor puede únicamente ser obtenido conociendo toda la población. Esas estimaciones son cercanas al parámetro poblacional, pero no exactas.

La contribución más importante que hemos encontrado es la investigación de Olivo (En prensa) que abarca diversos aspectos, utilizando también el marco teórico de Godino (2003) y otros trabajos posteriores de este autor. Comienza con un estudio histórico, donde muestra que la primera aproximación a la idea de intervalo surge de Bayes y su discípulo Laplace, quien estudia la estimación de la proporción y la media a partir del teorema de Bayes, llegando a una idea (no explícita) de lo que sería el intervalo de credibilidad y de este modo aparecen los fundamentos de la actual inferencia bayesiana. La alternativa de los intervalos fiduciarios introducidos por Fisher en su teoría fiducial, no generó éxito pues no brinda una distribución sobre los parámetros desconocidos, aunque, a través de ellos es posible determinar los valores de verosimilitud de los parámetros. También indica que la verosimilitud y la máxima verosimilitud, representan hoy las ideas principales sobre las que circula la teoría de estimación. Destaca que *los intervalos de confianza tal como se los conoce en la actualidad, surgen luego de la resistencia de muchos investigadores a aceptar las probabilidades subjetivas que son básicas para el procedimiento bayesiano* (Olivo, En prensa, pp. 268).

Este estudio servirá para justificar uno de los conflictos encontrados en los estudiantes en su investigación, que consiste en dar una interpretación bayesiana al intervalo y al coeficiente de confianza; interpretación que se presentó en un 36,5% de la muestra estudiada.

Un segundo punto en la investigación de Olivo (En prensa) es analizar y clasificar elementos de significado de los intervalos de confianza, utilizando una muestra de libros de texto adecuada al caso. Describe de este modo una gran

diversidad de campos de problemas, acompañados con una gran variedad de lenguaje, argumentos y propiedades. El autor se ha basado en estudios de textos universitarios de estadística (siguiendo el método de Tauber, 2001; Alvarado, 2007), y ha confirmado sobre todo conflictos de significado en cuanto a la definición del objeto de estudio se refiere, además en un muy alto porcentaje de libros se les da prioridad a las definiciones y propiedades y no se tratan las aplicaciones, dando lugar primordial a la teoría.

El tercer punto abordado en el trabajo de Olivo (En prensa) es la construcción de un instrumento válido para evaluar la comprensión del intervalo de confianza, luego de recibir instrucción sobre el objeto estadístico. En la construcción del mismo utiliza el método sugerido Batanero y Díaz (2005). Nos señala que la información recogida en la prueba del cuestionario ha permitido confirmar los resultados de investigaciones anteriores sobre el tema, así como describir nuevas dificultades, que no solo afectan a los intervalos de confianza, sino a los elementos relacionados que lo conforman.

Analizando las respuestas a los ítems abiertos, destaca y pormenoriza un elenco de conflictos semióticos, observados para los alumnos de las carreras de ingeniería, y con respecto a distintos significados del intervalo de confianza. Resumimos estos conflictos a continuación, clasificándolos dentro de cada tipo de objeto matemático primario emergente de los sistemas de prácticas, indicando si se trata de conflictos nuevos encontrados o se confirman resultados de otros autores:

- *Definición, propiedades y otros conceptos.* No se comprende que el coeficiente de confianza da un porcentaje de intervalos tomados en las mismas condiciones que contienen al parámetro. Se trata de un resultado confirmatorio (Behar, 2001). Interpretación Bayesiana del intervalo, este es un resultado aportado por la esa investigación. También detecta que el alumno confunde estadístico con parámetro; del mismo modo que confunde varianza poblacional y muestral. Este es un resultado confirmado por otras investigaciones (Vallecillos y Batanero, 1997; Behar, 2001 y delMas; Garfield, Ooms y Chance, 2007). Tampoco asocian correctamente el ancho del intervalo con el nivel de confianza, conflicto también denunciado en las investigaciones de Fidler y Cumming (2005) y Behar (2001). Finalmente encuentra un conflicto, consistente en que los alumnos visualizan los intervalos de confianza

como estadísticos descriptivos; este resultado también es confirmado por las investigaciones de Cumming y Fidler (2004).

- *Campos de problemas y procedimientos.* Encuentra un conflicto cuando el alumno confunde las condiciones de diferentes campos de problemas, por ello no individualiza la distribución de muestreo apropiada, y también confunden los parámetros en las distribuciones muestrales, por ejemplo el número de grados de libertad. Se trata en ambos casos de resultados aportados por la investigación de Olivo (En prensa). Observa además que no determina correctamente un valor crítico a partir de la tabla de la distribución. Este resultado lo confirma desde la investigación de Schuyten (1991)
- *Lenguaje y argumentación.* También la investigación de Olivo (En prensa) aporta otros resultados, denunciando que los alumnos hacen una interpretación incorrecta del intervalo de confianza a partir de una salida de ordenador, confunden los símbolos de varianza poblacional con desviación típica y observa una gran cantidad de conflictos con la notación. Finalmente un conflicto semiótico de este tipo lo encuentra, pues realizan una interpretación incorrecta de intervalos de confianza a partir de gráficos.

Como conclusión de este análisis sugiere que la enseñanza trate de organizarse para que el estudiante relacione los campos de problemas con cada una de las distribuciones muestrales, y que adquiera la lógica de los procedimientos para la construcción e interpretación de todos los tipos de intervalos que surgen, que se defina, se enuncien y demuestren propiedades, pero además se intensifiquen las aplicaciones en los campos de problemas.

## 2.4. INVESTIGACIONES RELACIONADAS CON EL ANÁLISIS DE VARIANZA Y DISEÑO EXPERIMENTAL

Son casi inexistentes las investigaciones relacionadas con el aprendizaje o enseñanza de estos objetos matemáticos, por lo cual pensamos que nuestro trabajo aportará resultados novedosos. A continuación exponemos los escasos trabajos que hemos encontrado en una búsqueda exhaustiva en publicaciones de congresos y revistas de educación estadística.

### 2.4.1 Comprensión de tipos de variables e interacción

Rubin y Rosebery (1990) planificaron y observaron un experimento de enseñanza dirigido a estudiar las dificultades de los profesores con las ideas estocásticas. Informaron que tanto los alumnos como su profesor interpretaron incorrectamente algunas de las ideas básicas del diseño experimental.

Una de las lecciones del mencionado experimento usó una actividad de lanzamiento a una canasta de baloncesto, en la que se varió la distancia de lanzamiento (de 1 a nueve metros) y el ángulo posicional del lanzador (para ángulos de 0, 45 y 90 grados). El objetivo de la lección era explorar los efectos separados de la distancia y el ángulo y la interacción entre las variables.

La observación de la discusión entre el profesor y los alumnos sobre la idea de variables independientes, dependientes y extrañas en el experimento de lanzamiento mostró la confusión entre estos conceptos. Algunos estudiantes sugirieron como posibles variables independientes características individuales del lanzador, como su altura o su habilidad para encestar. Incluso la altura de la canasta, que se conservó inalterable durante el experimento fue considerada como variable independiente por algunos estudiantes.

Otros estudiantes sugirieron que la iluminación del gimnasio podría ser diferente para las distintas combinaciones de ángulo y distancia, de modo que tanto el profesor como los alumnos quedaron con la creencia de que la presencia de tales influencias podría hacer imposible la obtención de conclusiones sobre el efecto de las variables ángulo y distancia. Finalmente, Rubin y Rosebery resaltaron la dificultad en distinguir entre las características de los sujetos que no tenían influencia sobre el resultado del experimento de otras variables que si podrían tenerla. El papel de la asignación aleatoria como medio de compensar estas diferencias individuales tampoco fue comprendido.

### 2.4.2 Control experimental y condiciones de aplicación de los métodos

Una dificultad importante en las ciencias humanas es el control experimental. Puesto que no es posible controlar todas las variables relevantes, la estadística sugiere la necesidad de aleatorización, que garantiza la existencia de técnicas para medir la probabilidad de que un efecto haya sido producido aleatoriamente. En la práctica esta situación no se da. Por un lado, el número de observaciones dificulta que se pueda controlar simultáneamente todas las variables de interés. Por otro lado,

quizás no todas las variables relevantes hayan sido medidas y algunas variables "confundidas" podrían no ser controladas (Selvin, 1970).

La dificultad de tomar muestras aleatorias implica que no se alcanzan las condiciones exigidas para aplicar los métodos estadísticos en forma correcta. En otras ocasiones se violan otros supuestos, por ejemplo, la frecuencia es demasiado pequeña para aplicar un test de ji-cuadrado (Díaz, Batanero y Wilhelmi, en prensa).

## 2.5. OTRAS INVESTIGACIONES RELACIONADAS

También se producen errores que inciden en los resultados e interpretación de la estadística por una incorrecta identificación de la población en estudio, tomar una muestra de tamaño insuficiente o interpretar resultados de muestras no aleatorias como si el muestreo hubiese sido aleatorio, o bien no se especifica con claridad cual es la población a la que se quiere extender los resultados (Díaz, Batanero y Wilhelmi, en prensa).

Un prerrequisito para comprender los intervalos de confianza y el contraste de hipótesis son las distribuciones muestrales, tema sobre el que resumimos a continuación los estudios previos.

Well, Pollatsek, A. y Boyce, S. J.(1990) realizan cuatro experimentos donde se investiga como los alumnos comprenden las distribuciones muestrales para la media. Se aplica a grupos de estudiantes de psicología, a los que no se les da una instrucción a priori (sólo se les da instrucción en el cuarto y último experimento). La finalidad es hallar las causas que los llevan a aplicar heurísticas de representatividad. Concluyen que cada vez que se les preguntaba sobre le exactitud de las medias muestrales ó sobre la zona central de la distribución muestral, los sujetos utilizan la información del tamaño muestral en forma mas apropiada; no así cuando se les pedía información acerca de las colas de las distribuciones muestrales. De aquí se desprende que entre las variables puestas en juego, la más importante es el parecido que existe entre la media muestral y la poblacional para el éxito en la tarea.

También hemos encontrado investigaciones para estudiar el aprendizaje de las distribuciones muestrales, usando el ordenador, realizando simulaciones. Hodgson (1996) señala que este recurso podría sumar un esfuerzo al alumno, ya que introduce más información en el proceso de aprendizaje. La simulación agrega también la posibilidad de obtener nuevas concepciones erróneas, a pesar que mejora la comprensión en los alumnos y es un recurso didáctico inestimable a la hora de enseñar los conceptos relacionados a las distribuciones muestrales.

DelMas Garfield y Chance (1999) diseñaron actividades educativas describiendo el software Sampling Distribution guiando a los alumnos en la exploración de las distribuciones muestrales. En este experimento, observaron que los estudiantes cambiaban la forma de la distribución teórica de la población sin inconvenientes, y podían simular la distribución muestral de diversos estadísticos, incluso con varios tamaños de muestras. A pesar de ello, los autores dan aviso que el uso de tecnología no siempre genera una adecuada comprensión de las distribuciones muestrales en los estudiantes. Coincidiendo con Well, Pollatasek y Boyce (1990), sugieren que el agregado de actividades usando software puede generar una exigencia adicional en los alumnos. Esta nueva información basada en software puede interferir para el aprendizaje de las distribuciones muestrales

## 2.6. CONCLUSIONES

El estudio de los antecedentes muestra una gran variedad de dificultades en relación a objetos estadísticos tales como muestreo, distribución muestral, intervalo de confianza y contraste de hipótesis y también alguna investigación aislada sobre otros objetos relacionados con el diseño experimental.

En el cuestionario que pensamos construir incluiremos algunos ítems relacionados con la comprensión del intervalo de confianza y contraste de hipótesis pues son objetos básicos en el tema. Haremos énfasis en aspectos como la potencia en las pruebas de hipótesis que no han sido tan tratados en la investigación previa.

Por otro lado no hemos encontrado investigaciones didácticas relacionadas con la comprensión del análisis de varianza, sus diversos modelos, elección del modelo, supuestos y comprobación de los mismos o la interpretación y comprensión

de cómo se obtiene una tabla de análisis de varianza. Por ello todos estos puntos serán incluidos en el cuestionario, y los mismos tendrán un gran peso.

# 3. ANÁLISIS DEL CUESTIONARIO

## 3.1. INTRODUCCIÓN

En este capítulo tratamos de responder al primer objetivo planteado en nuestro estudio, es decir, se trataría de definir el significado institucional evaluado en la investigación, a través de una primera definición semántica de la variable objeto de evaluación. Para ello analizaremos algunos ítems que podrían ser útiles en un futuro cuestionario de evaluación de la comprensión de objetos estadísticos elementales en el diseño experimental. Además trataremos de ver el grado en que el conjunto de ítems cubren el contenido pretendido de nuestra variable, que se describió someramente en la sección 1.2.

En lo que sigue realizamos un análisis a priori de los ítems sometidos a prueba, finalizando con una tabla de especificaciones del contenido cubierto por el conjunto de ítems, lo que constituye una primera aproximación a la definición futura del contenido de la variable que se pretende medir. En el enunciado de cada ítem en negrita se indica la respuesta correcta.

## 3.2. ANÁLISIS A PRIORI DEL CUESTIONARIO DE EVALUACIÓN

A continuación analizaremos el contenido a priori de cada uno de los ítems pasados en la prueba. Como se ha indicado, se decidió utilizar el formato de los ítems de opción múltiple, para poder evaluar un mayor número de contenidos en el tiempo disponible, teniendo en cuenta también que los estudiantes de la muestra estaban habituados a este tipo de formato, ya que lo usan habitualmente en sus exámenes.

Ítem 1. Queremos conocer si los sujetos extrovertidos e introvertidos difieren en la puntuación media en autoestima y no disponemos de ninguna información previa. El tipo de hipótesis nula razonable que debo plantear es:

a) $\mu_I \leq \mu_E$
b) $\mathbf{\mu_I = \mu_E}$
c) $\mu_I \geq \mu_E$

Se pretende evaluar a través del ítem la comprensión sobre el test de hipótesis en general, y en particular la asignación de hipótesis, y dentro de ésta a su vez, la asignación a la hipótesis nula. El enunciado plantea una hipótesis de investigación (en términos verbales), y el alumno debe traducirlo a una expresión simbólica,

diferenciando entre hipótesis nula y alternativa y entre un test unilateral y bilateral. Vallecillos (1994) plantea un ítem similar, en el contexto de estimar una proporción y encuentra un 31,4% de errores en su muestra, indicando que los estudiantes confunden hipótesis estadística nula con hipótesis de investigación en este ítem.

La respuesta correcta es la b), ya que para este caso no se dispone de información previa. Además, como se quiere conocer si los sujetos extrovertidos e introvertidos "difieren" en la puntuación, esa opción deberá caer en la alternativa. En términos de este problema, que $H_o$ (hipótesis nula) sea verdadera significa que: no difieren en la puntuación media en autoestima para los dos tipos de sujetos involucrados en el estudio. Si designamos por $\mu_I$ : la puntuación media poblacional de los sujetos introvertidos y con $\mu_E$ : la puntuación media poblacional de sujetos introvertidos, resulta $H_0 : \mu_I = \mu_E$ . Los errores que se evalúan en este ítem son los siguientes:

- El distractor c) lo elegiría un alumno que usa sus ideas previas, pensando que las personas introvertidas tienen menos autoestima que las extrovertidas. No se usa el enunciado del problema
- El distractor a) además del error anterior el alumno confunde las hipótesis nula y alternativa, confusión que se ha encontrado en investigaciones previas, como la de Vallecillos (1994).

Ítem 2. Un psicólogo escolar desea estimar la puntuación media de la población en un test de rendimiento de lectura. Para ello administra el test a una muestra de 36 estudiantes, obteniendo una media de 48 y desviación típica 10. Calcular, al nivel de significación $a$=0,05, los límites del intervalo de confianza para la puntuación en el test. (Error típico de la media: $S_d/\sqrt{n}$).

a) [38 , 58]
b) [15,33 , 80,66]
c) **[44,73 , 51,27]**

Se pretende evaluar a través de esta pregunta la comprensión sobre el procedimiento de construcción del intervalo de confianza, particularizado para la media de una población. En este caso se desea estimar: "el rendimiento medio de lectura en una población de 36 estudiantes". Llamemos $X$ a la variable aleatoria que mide la puntuación en el test de rendimiento de lectura; se da como dato la media $(\bar{x})$ y desviación típica de la muestra ($s_d$), tamaño de muestra ($n$) y la fórmula del error típico de la media ($s_d/\sqrt{n}$). El procedimiento que se deberá emplear aquí para la

determinación del intervalo pedido es componer la fórmula $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{S_d}{\sqrt{n}}$, con los datos dados, es decir: $\bar{x} = 48;\ \ \alpha = 0{,}05$ de donde resulta $z_{0{,}025} = 1{,}96$ ; $S_d = 10$ y *n=36.*

- El distractor c) nos proporciona el resultado correcto, ya que $48 \pm 1{,}96 \cdot \frac{10}{\sqrt{36}} = 48 \pm 1{,}96 \cdot \frac{10}{6} = 48 \pm 3{,}27.$ LI = 44,73 y LS = 51,27. [1] El intervalo de Confianza, resulta : **[44,73 , 51,27]**

Los errores que se evalúan en este ítem son los siguientes:

- En el a) se hace un error en la fórmula sumando y restando simplemente la desviación típica $48 \pm 10;$ LI = 38 y LS = 58.
- En c) se trata de ver si el alumno lee correctamente las tablas de la distribución, pues estaría considerando como valor del percentil 19,62; en lugar del 1,962. A través de este error, estaría considerando como coeficiente de confianza prácticamente $\alpha \approx 0$ ; ya que $P(X \geq 19{,}62) \approx 0$ , si $X \sim N(0,1)$ . De este modo su intervalo sería $48 \pm 19{,}62 \cdot \frac{10}{\sqrt{36}} = 48 \pm 32{,}67$ LI = 15,33 y LS = 80,67.

Ítem 3. En relación con la pregunta anterior, también podríamos afirmar que:
a) **Si extrajéramos 100 muestras y calculáramos en cada una el intervalo de confianza para la media, en 95 de ellos, se encontraría el verdadero valor del parámetro.**
b) La puntuación del test está en el intervalo estimado con una probabilidad del 95%
c) Si extrajéramos 100 muestras y calculáramos en cada una el intervalo de confianza, en 95 de los intervalos, se encontraría la puntuación de los sujetos de la muestra.

Se pretende evaluar la interpretación que se hace sobre el intervalo de confianza a través de esta unidad, pues tanto Behar (2001) como Olivo (En prensa) encuentran errores en esta interpretación. En esta última investigación se reporta que los estudiantes no alcanzan a interpretar que se representa mediante un intervalo de confianza, dando en general una interpretación bayesiana y sólo el 36,5 % de los estudiantes dieron una respuesta correcta.

La respuesta correcta es la a), ya que si hacemos la interpretación frecuentista de

[1] LI : Límite Inferior del Intervalo de confianza; LS: Límite Superior del Intervalo de Confianza.

la probabilidad se tendría que usando $\alpha = 0{,}05$ dado para el ítem anterior, el valor 0,95 (95% si se da en porcentaje), indica el nivel de confianza del intervalo. Por tanto, la $P\left(LI\left(\overline{X}\right) < \mu < LS\left(\overline{X}\right)\right) = 0{,}95$ nos está diciendo que sobre 100 muestras distintas extraídas de la misma población, 95 de los intervalos de confianza obtenidos, incluirían el verdadero valor de la media poblacional $\mu$ . En definitiva, el coeficiente de confianza sólo nos da la proporción de intervalos calculados de la misma población con tamaño de muestra dado que cubrirían el valor del parámetro, pero no si el intervalo calculado lo cubre o no (Cumming, Williams y Fidler, 2004). Nótese que los extremos del intervalo varían aleatoriamente de muestra a muestra, mientras que el valor del parámetro es fijo (constante) aunque desconocido.

Los errores que se evalúan en este ítem son los siguientes:

- El distractor b) da la interpretación bayesiana del intervalo de confianza, la cual consiste en considerar los extremos como fijos y el parámetro como aleatorio, es decir, interpretar el coeficiente de confianza como coeficiente de credibilidad que da una probabilidad a posteriori para el parámetro, dado el conocimiento de los datos de la muestra.
- En el caso del distractor c) el alumno que lo eligiera estaría haciendo una confusión entre el valor de la variable y el parámetro a estimar, que se ha descrito también en investigaciones previas.

Ítem 4. Supongamos un contraste bilateral sobre la media, siendo la variable estudiada la inteligencia. Para $H_0$: μ = 100, $H_1$: μ = 110, α = 0,05 y β = 0,4406. ¿Cuál es la probabilidad de rechazar la $H_0$ cuando no es 'cierta'?

a) 0,05
b) 0,4406
c) **0,5594**

Se pretende evaluar la comprensión sobre el test de hipótesis en general, los errores que pueden cometerse (Errores de Tipo I y Tipo II), y la potencia determinada para un test en particular. Tanto en la respuesta correcta como en los distractores, los estudiantes tienen que interpretar una probabilidad condicional, pues la potencia del test y los dos tipos de errores, vienen definidos por probabilidades condicionales. Ellas son las que producen, dentro del contraste de hipótesis mayor cantidad de errores y concepciones erróneas, tanto en los estudiantes universitarios como los científicos que usan a través de su trabajo la inferencia estadística (Vallecillos, 1994, Batanero, 2000).

En las pruebas de Neyman-Pearson, contrastes como reglas de decisión entre dos hipótesis, se consideran estos dos tipos de errores (Batanero, 2000). El error Tipo I se considera la posibilidad de: *rechazar Ho, siendo en realidad Ho verdadera* y su probabilidad es constante (nivel de significación α).

Por el contrario, la probabilidad de Error Tipo II o probabilidad de *aceptar una hipótesis nula Ho, cuando Ho es falsa* resulta variable, pues en caso de ser verdadera la hipótesis alternativa $H_1$ hay un conjunto infinito de valores que hacen que la hipótesis nula no sea verdadera (condición para determinar la probabilidad que se pide para este ítem). Al complemento de $\beta$ ó probabilidad de rechazar la hipótesis nula cuando no sea cierta se lo denomina potencia del test, la cual también es variable pues depende del verdadero valor que se le de al parámetro. A la potencia del test se la suele simbolizar por $\pi$, de esta forma resultará que: $1-\beta=\pi$.

En este item se está calculando la potencia para un valor particular dado de $\mu$ =110. La respuesta correcta es la c), ya que suponemos que se calcula la potencia para los valores de $\alpha$ y $\beta$ dados, los cuales representan respectivamente las probabilidades de cometer Errores de Tipo I y de Tipo II . Al ser $\beta$ = 0,4406; la probabilidad pedida es $\pi$ =1-0,4406=0,5594.

Los errores que se evalúan en este ítem son los siguientes:

- Al elegir como respuesta correcta el ítem a) estaría confundiendo la potencia del test con el nivel de significación, ya que aquí se da como dato $\alpha = 0{,}05$, siendo esta la probabilidad de cometer Error de Tipo I, $P(\text{rechazar Ho, siendo Ho cierta})$, mientras se pide determine la $P(\text{rechazar Ho, siendo Ho falsa})$.
- Al elegir como respuesta correcta el ítem b) estaría confundiendo la potencia del test con la probabilidad de cometer Error de Tipo II ó *P(no rechazar Ho, siendo Ho falsa)*, ya que aquí se da como dato $\beta = 0{,}4406$.

Ítem 5. Una maestra cree que unas nuevas actividades de lectura ayudarán a mejorar la capacidad lectora de los niños de primaria. La maestra tiene una clase con 21 niños a los que le pasa la prueba de lectura DRP (Degree of Reading Power), para conocer el nivel del que parten. Después realiza estas actividades en clase durante 8 semanas. Al final del período vuelve a pasarles la prueba. De las siguientes técnicas, ¿cuál debería aplicar los investigadores para comprobar si las actividades modifican las capacidades lectoras?

a) Contraste de hipótesis sobre dos medias independientes
b) **Contraste de hipótesis sobre dos medias relacionadas**

c) ANOVA de un factor completamente aleatorizado

A través de este ítem se desea evaluar si al conocer las características de un problema, el alumno es capaz de escoger el modelo que mejor se adecua al mismo.

La respuesta correcta es la b), ya que la maestra pasa a los mismos individuos la prueba antes y después de ser sometidos a un tratamiento. Se trata de dos poblaciones $X_1, X_2, ....., X_{21}$ e $Y_1, Y_2, ....., Y_{21}$, para las cuales sus medidas están relacionadas, ya que los individuos son los mismos. Si fuese posible suponer normalidad de los datos generados con sus diferencias, es decir: $D_1 = Y_1 - X_1, D_2 = Y_2 - Y_1, ....., D_{21} = Y_{21} - X_{21}$, el contraste que se aplica es la prueba *t* para una muestra, utilizando los valores de $D_1, D_2, ....., D_{21}$ obtenidos . Los errores que se evalúan en este ítem son los siguientes:

- Al elegir la respuesta a) estaría confundiendo el concepto de independencia de muestras, ya que aquí se toma la misma población para pasar la prueba al cabo de un período, claramente las poblaciones son dependientes desde la naturaleza del problema.
- Al elegir como correcta la respuesta c) estaría desconociendo los supuestos mínimos de ANOVA de un factor completamente aleatorizado, el cual no es otro que el de independencia entre los individuos a los cuáles se les aplica los distintos tratamientos. Además para elegir este modelo, se requiere al menos comparar tres muestras independientes, no dos como en este caso.

Ítem 6. Supongamos que conocemos la 'verdad absoluta' sobre la eficacia de dos tratamientos (A y B), y sabemos que existen diferencias en la efectividad de ambos para curar la depresión. Un investigador que realice un estudio y parta de la hipótesis '*no existen diferencias en la efectividad de los tratamientos A y B para curar la depresión*' cometerá un *error tipo II* cuando:

a) Concluya que A y B no son efectivos para curar la depresión
b) **Concluya que A y B no difieren en su efectividad para curar la depresión**
c) Concluya que A y B difieren en su efectividad para curar la depresión

Con este ítem se quiere evaluar la comprensión de la diferencia entre hipótesis estadísticas e hipótesis de investigación; y si el alumno diferencia el Error Tipo I del Error Tipo II.

La respuesta correcta es la b). Consideremos $\mu_A$ y $\mu_B$ el verdadero valor medio de la puntuación de la efectividad al aplicar el tratamiento A y B respectivamente a la población para curar la depresión. Con esto, las hipótesis nulas

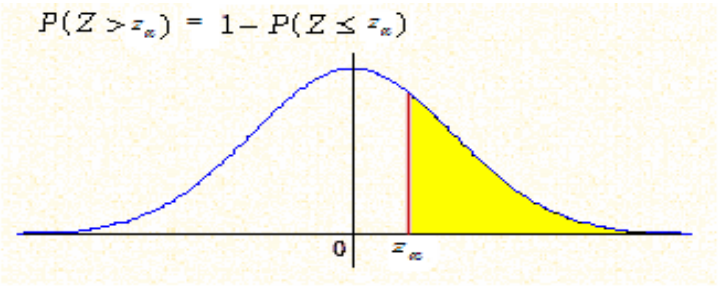
y alternativas elegidas para este problema son: $H_0 : \mu_A = \mu_B$ vs $H_1 : \mu_A \neq \mu_B$. Con lo cual cometerá un Error de Tipo II cuando decida que no existe diferencia entre los dos tratamientos para curar la depresión, cuando en realidad si existen. Además le interesará controlarlo, ya que en estos términos ha enunciado su hipótesis de investigación. Los errores que se evalúan en este ítem son los siguientes:

- En el caso que escogiera el distractor c) como respuesta a este ítem, estaría confundiendo el Error de Tipo I con el de Tipo II, ya que el error de Tipo I en este problema es afirmar que existe diferencia entre los dos tratamientos aplicados para curar la depresión, cuando en realidad no hay diferencia.
- En el caso que eligiese como correcto el distractor a) estaría haciendo caso omiso a una hipótesis explícita en el problema de investigación, ya que se afirma que debemos suponer que conocemos la 'verdad absoluta' sobre la eficacia de los dos tratamientos A y B. Es decir tanto el tratamiento A, cuanto el B han sido probados separadamente y resultaron eficaces para la cura de la enfermedad; pero la investigación que tiene como consecuencia este problema, es que hay "diferencias" entre ellos.

Ítem 7. La puntuación típica correspondiente a una α= 0,01 en un contraste unilateral derecho es:
a) **2,33**
b) -2,33
c) 3,10

Mediante este ítem tratamos de evaluar la comprensión de la relación entre el nivel de significación y el valor crítico, la diferenciación entre test unilateral y bilateral, y el concepto de puntuación típica. Implícitamente el enunciado asume que se puede aplicar la distribución normal. Además evaluamos la comprensión alcanzada para el manejo de tablas estadísticas. Usaremos la Figura 3.2 (gráfico de la función de densidad asociada con una variable aleatoria *Z* normal con media 0 y varianza 1), para analizar los criterios, en los cuales podría estar pensando un alumno que escogiera cada distractor:

Figura 3.2. Cálculo de áreas en la normal




La respuesta correcta es la a), ya que la puntuación $z_\alpha$ , es aquél valor que deja un área de $\alpha$ unidades cuadradas a la derecha bajo la curva de densidad de una variable aleatoria *N(0,1)* (parte sombreada del gráfico 3.2), teniendo en cuenta que se trata de un test unilateral derecho, ó sea bajo las hipótesis estadísticas: $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ . Como el área bajo la curva debe ser de una unidad cuadrada, se tiene que si el valor dado de $\alpha = 0,01$ , $z_\alpha$ debe ser positivo. Los errores que se evalúan en este ítem son los siguientes:

- Si la respuesta elegida fuese el distractor b), lo que hace es tomar el test unilateral izquierdo o lo que es lo mismo confunde hipótesis nula y alternativa o bien la región crítica y de aceptación en un contraste unilateral. Además, si la puntuación típica fuese negativa, el valor de $\alpha$ debería ser superior a 0,5 unidades.
- Si la respuesta elegida fuese el distractor c), hay un error en la lectura de la tabla, ya que $P\left(Z_{N(0,1)} > 3,10\right) \approx 0,00097$ .

Ítem 8. La prueba que se utiliza para comprobar el supuesto de esfericidad en un estudio con una V.I. con medidas repetidas es:

a) La prueba de Levene
b) La prueba de Kolmogorov
c) **La prueba de Mauchly**

En este ítem se quiere evaluar la comprensión de las diversas pruebas estadísticas que se utilizan para validar los diseños de experimentos, y si sabe elegir las que se aplican a los diseños con medidas repetidas.

Un supuesto para la correcta aplicación de ANOVA en medidas repetidas es que no se verifiquen interacciones. La respuesta correcta es la c), ya que la prueba de Mauchly, también llamada prueba de esfericidad es la que se realiza para comprobar este supuesto. Esta prueba contrasta la hipótesis nula de que la matriz de covarianza del error de variables dependientes transformadas es proporcional a una matriz identidad. Se

supone, si se calculan las diferencias entre todas las parejas de valores de varianzas del factor, que estas diferencias son estadísticamente iguales. Esto implica que la matriz de covarianza define una esfera. Cuando se rechaza la hipótesis de esfericidad deben corregirse los grados de libertad, multiplicándolos por un factor, épsilon, que es proporcional en el resultado del test y mide el alejamiento de la esfericidad; si esta fuese perfecta, épsilon sería igual a 1. Los errores que se evalúan son los siguientes:

- De escoger como correcto el distractor a) estaría confundiendo con la prueba de Levene utilizada para testear la igualdad de varianzas. En esta prueba se contrasta la hipótesis nula de que las varianzas de los grupos poblacionales son iguales, versus la alternativa que existe al menos un par de ellas que son distintas.
- De escoger como correcto el distractor b) estaría confundiendo con la prueba de Kolmogorov (prueba de bondad de ajuste) para chequear el grado de acercamiento de los datos a alguna distribución prefijada. En esta prueba de Kolmogorov, se contrasta la hipótesis nula que la distribución observada se ajusta a una distribución teórica, versus la alternativa, que la distribución observada no se ajusta a la distribución teórica puesta en la hipótesis nula. En general esta prueba es muy utilizada cuando se desea comprobar que es posible afirmar que la muestra proviene de una población cuyos elementos distribuyen en forma normal, como una media y una desviación estándar especificadas de antemano.

Ítem 9 ¿Cuál de las siguientes hipótesis está bien formulada?:
a) Ho: μ= 3; Hi: μ ≠ 4
b) Ho: μ= 3; Hi: μ ≥ 3
c) **Ho: μ= 3; Hi: μ ≠ 3**

En este ítem se pretende evaluar la comprensión del grupo de alumnos acerca de la formulación de hipótesis estadísticas, un tema en el cuál también Vallecillos (1994) encontró dificultades. La respuesta correcta es la c). Recordemos que las hipótesis estadísticas son dos y complementarias entre ellas; recorriendo para este caso específico el conjunto de los números reales pues es sólo uno el parámetro el de interés. Así, la media poblacional $\mu$, se tendrá que en el caso c) $\mu \in \{3\}$ bajo $H_0$ y $\mu \in R - \{3\}$ bajo $H_1$. De este modo se observa que los conjuntos de valores bajo ambas hipótesis, $\{3\}$ y $R - \{3\}$, cumplen la propiedad de ser complementarios para los reales. Los errores que

se evalúan en este ítem son los siguientes:

- En el caso de elegir el distractor b) estaría desconociendo que los conjuntos a los cuales pertenece el parámetro bajo ambas hipótesis deben ser complementarios (en este caso la constante 3 está como valor posible del parámetro, tanto para la hipótesis nula como para la alternativa).
- En el caso de elegir el distractor a) además de cometer un error semejante al del distractor a) se está tomando un test unilateral, pero nunca debe aparecer la igualdad bajo la hipótesis alternativa al desarrollar una prueba de hipótesis.

Ítem 10. Cuando realizamos un contraste, la regla de decisión nos lleva a rechazar la hipótesis nula siempre que:
a) El estadístico de contraste caiga en la región de rechazo
b) La probabilidad asociada al estadístico de contraste (el valor de significación) sea menor que el valor de alfa
c) **a) y b) son correctas**

Evaluamos a través de este ítem el conocimiento de las condiciones en las que se tomará la decisión de rechazar la hipótesis nula, pues Vallecillos (1996) sugirió que algunos estudiantes no comprenden la lógica del contraste de hipótesis. También se evalúa la comprensión de la región de aceptación y rechazo. Una regla de decisión para rechazar la hipótesis nula es, (luego de construir ambas zonas de acuerdo con las hipótesis que se están contrastando) observar si el estadístico de prueba cae en la zona de rechazo. En el caso de confirmarse se decide que existe suficiente evidencia como para rechazar la hipótesis nula. Es decir bajo esta condición es correcta la respuesta a).

Para un valor observado del estadístico de prueba, digamos $\frac{\sqrt{n}\left(\bar{x}-\mu_0\right)}{\sigma}$ [2], se tiene que la probabilidad asociada al estadístico de contraste es: $P\left(Z_{N(0,1)} > \frac{\sqrt{n}\left(\bar{x}-\mu_0\right)}{\sigma}\right)$.

Que el valor de esta probabilidad sea menor que el valor de $\alpha$ dado (nivel de significación del test) es equivalente a decir que el valor observado del estadístico de prueba cae en la zona de rechazo para este test, de donde la respuesta b) también es correcta. Finalmente la respuesta correcta a esta unidad de medida es la c), ya que tanto a) como b) son correctas.

[2] Test de la media con desviación estándar conocida y distribución de la población Normal con media bajo la hipótesis nula

Si eligiera sólo la respuesta a) ó la b) estaría desconociendo la relación de equivalencia entre las dos respuestas.

Ítem 11. El modelo lineal utilizado para representar las fuentes de variabilidad presentes en un ANOVA de un factor, efectos fijos completamente aleatorizado es:

a) **$Y_{ij}=\mu + \alpha_j + E_{ij}$**
b) $Y_{ij}=\mu + \alpha_j + \beta_i + E_{ij}$
c) $Y_{ijk}=\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + E_{ijk}$

En este ítem tratamos el análisis de varianza y queremos evaluar la comprensión del grupo de alumnos para asociar un modelo estadístico, de acuerdo con el estudio que se desea realizar. Se pregunta por el modelo lineal que representa las fuentes de variabilidad en un ANOVA de un factor fijo completamente aleatorizado.

Al tratase de un Análisis de la Varianza de un factor fijo, las fuentes de variación serán: variación total, variación entre los tratamientos (del único factor que se estudia) y variación residual (García, 2004). El modelo lineal que se debe elegir como el que mejor se adecua al problema es: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ , siendo $\mu$ la media de la población sometida al i-ésimo tratamiento, $\alpha_i$ el efecto del i-ésimo tratamiento y $\varepsilon_{ij}$ una cantidad aleatoria, supuesta para este modelo con distribución Normal. De este modo, la respuesta correcta es la a). Los errores que se evalúan en este ítem son los siguientes:

- En el caso de elegir el ítem b) estaría considerando un modelo aditivo, con dos factores cuyos efectos serían $\alpha_i$ y $\beta_j$ . El alumno confunde los modelos para uno y dos factores. Además, este modelo es el que se aplicaría para un ANOVA de un factor con medidas repetidas, pues no se esta considerando interacción.
- En el caso de elegir el ítem c) estaría considerando un modelo completo aditivo-multiplicativo para dos factores, que tiene en cuenta la presencia de dos factores, pero además al ser completo, se analiza la presencia de un efecto interacción entre los mismos, a través del término $(\alpha\beta)_{ij}$

Ítem 12. Si en un ANOVA de un factor, y medidas repetidas encuentro que la F empírica u observada toma un valor de 8,16 esto quiere decir que:

a) CM intergrupos / CM error = 8,16
b) **CM intrasujetos / CM error = 8.16**
c) CM intergrupos / CM intrasujetos = 8,16

A través de este ítem evaluamos la comprensión del grupo sobre algunos de los

cálculos realizados para completar una tabla ANOVA para el diseño de un factor con medidas repetidas, esto es, cuando a un mismo sujeto se le aplican varios tratamientos y se quiere ver la diferencia del tratamiento en diferentes momentos. También se particulariza la construcción del valor F observado para el único factor presente, el factor intra-sujetos. La respuesta correcta es la b), pues el diseño es de medidas repetidas y sólo hay un factor, por tanto, se analiza en este tipo de diseños la variabilidad intra sujetos, ya que la aleatorización sólo se aplica a los sujetos dentro de cada grupo. El análisis de la variabilidad ínter grupos no es un objetivo en este tipo de experimento, pues no tiene mayor relevancia. Los errores que se evalúan en este ítem son los siguientes:

- De elegir el distractor a) como respuesta correcta, el sujeto estaría analizando la significación de la variabilidad introducida por la formación de los grupos, que no es relevante para este tipo de experimentos, sino para el diseño ordinario de un factor completamente aleatorizado. El alumno confunde medidas repetidas con factor de efectos fijos
- De elegir como correcto el distractor c) consideraría el análisis de la interacción entre dos factores, que no están presentes para este modelo, donde sólo se considera un factor.

Ítem 13 y 14. Se quiere estudiar el efecto de ciertas variables motivacionales sobre el rendimiento en tareas de logro. Se manipularon dos variables: "tipo de entrenamiento motivación" (A1: instrumental; A2: atribucional y A3: control) y "clima de clase" (B1: cooperativo; B2: competitivo y B3: individual). Se seleccionaron a 45 sujetos y se dividieron en grupos para cada condición experimental.
A continuación presentamos la tabla del ANOVA incompleta. A partir de la información de la tabla contesta a las preguntas del Ítem 13 y del Ítem 14.

| Fuente de variación | SC | Gl | MC | F |
|---|---|---|---|---|
| Factor A | 70 | | | |
| Factor B | | | 20 | |
| Interacción AB | | | | 3,91 |
| Error | 46 | | 1,278 | |
| Total | 176 | 44 | | |

Ítem 13. (CUESTIONARIO B)
El valor del sumatorio cuadrado para el factor B es:
a) 15,65
b) 35
c) **40**

En esta unidad se quiere evaluar la comprensión de los pasos que se siguen en el análisis estadístico del diseño factorial de dos factores con efectos fijos. En particular el

formato para la descomposición de la variación total (a través de la suma de cuadrados total, $SS_T$), la distribución de los grados de libertad (*GL*) y el cálculo de la medida de cuadrados (*MC*) para todas las fuentes de variación; elementos éstos prioritarios para el análisis de la significación tanto de los factores principales (en este caso son A y B), y de la interacción AB presentes en el problema. Recordemos que la descomposición de la variación total, a través de la $SS_T$ es:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E \quad (1)$$

Cada una de las sumas de cuadrados presentes en la igualdad última anterior, se le puede asociar un número entero positivo que represente su *grado de libertad.* Este número se determina contando la cantidad de elementos independientes que hay en dicha suma de cuadrados. Así, $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{...}\right)^2$ para un modelo con *a* niveles del factor *A*, *b* niveles del factor *B* y con *n* datos recogidos en cada celda (por esto se lo denomina modelo balanceado) tiene *abn* -1 elementos independientes, por ello diremos que *GL* ($SS_T$) = *abn* -1.

Análogamente $SS_A = bn\sum_{i=1}^{a}\left(\overline{Y}_{i..} - \overline{Y}_{...}\right)^2$, suma de cuadrados correspondiente al factor *A*, tiene *a-1* elementos independientes, de donde *GL* ($SS_A$)=*a-1*, y en forma análoga *GL* ($SS_B$)=*b-1* serán los grados de libertad en la suma de cuadrados que se corresponde con *B*. Así completamos la descripción de los *GL* para los factores principales.

La respuesta correcta es la c), como vemos en la Tabla 3.1, donde damos la tabla ANOVA completa. En la columna PERCENTIL se han colocado los valores críticos de la distribución *F* obtenidos, usando los correspondientes grados de libertad y para el nivel de significación del 5%. Estos valores son los que han de compararse con el valor observado del estadístico para ver si caemos en la zona de rechazo ó de aceptación del test. Por ejemplo, el 3,25 se corresponde con el percentil obtenido de la tabla *F* con nivel de significación 0,05, con 2 y 36 grados de libertad en el numerador y denominador respectivamente.

Tabla 3.1. Tabla ANOVA, solución del problema. Ítem 13

| | Fuente de variación | SC | GL | MC | F | PERCENTIL |
|---|---|---|---|---|---|---|
| Fila 1 | FACTOR A | 70 | 2 | 35 | 27,37 | 3,25 |
| Fila 2 | FACTOR B | 40 | 2 | 20 | 15,65 | 3,25 |
| Fila 3 | INTERACCION | 20 | 4 | 5 | 3,91 | 2,63 |
| | ERROR | 46 | 36 | 1,278 | | |
| | TOTAL | 176 | 44 | | | |

Los errores que se evalúan en este ítem son los siguientes:

- De escoger como correcto el distractor a) confunde la suma de cuadrados pedida para el factor B con el $F_{obs}$ (valor observado del estadístico de prueba).
- De escoger como correcto el distractor b) confunde con media de cuadrados del factor A.

Se presentó un ítem similar para el Cuestionario A en esta evaluación, donde se pide analice la validez para otra puntuación presente en la tabla ANOVA, el mismo es:

Ítem 13. (Cuestionario A)
El valor de la media cuadrática para el factor A es:
a) 5
b) **35**
c) 15,65

Ambos ítems serán analizados por separado a la hora de observar las respuestas de los alumnos, aunque el objetivo perseguido es el mismo: interpretación de las puntuaciones presentes en una Tabla ANOVA de dos factores fijos.

Ítem 14. Una de las conclusiones del estudio sería (alfa = 0,05)
a) **Hay efecto del factor A ("entrenamiento") sobre el rendimiento en tareas de logro**
b) No hay efecto del factor B ("clima de clase") sobre el rendimiento en tareas de logro
c) No hay interacción de los factores

En este ítem se quiere evaluar la interpretación de los valores obtenidos en la tabla completa, para dar una respuesta estadísticamente válida al problema. El modelo de los efectos ajustado específico para este problema es:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,3 \\ j = 1,2,3 \\ k = 1,2,...,5 \end{cases} \tag{2}$$

Siendo $\mu$ el efecto promedio global, $\alpha_i$ es el efecto del i-ésimo nivel del factor *A*, $\beta_j$ es el efecto del j-ésimo nivel del factor *B*, $(\alpha\beta)_{ij}$ el efecto de la interacción entre

$\alpha_i$ y $\beta_j$. Finalmente, $\varepsilon_{ijk}$ es un componente del error aleatorio presente en el problema. Recordemos que los contrastes que aquí se analizan ponen a prueba los siguientes pares de hipótesis en las Filas 1, 2 y 3 de la tabla ANOVA respectivamente (ver Tabla 3.1):

- $H_0^A : \alpha_1 = \alpha_2 = \alpha_3 = 0$ vs $H_1^A : (\exists i)(i = 1,2,3) / \alpha_i \neq 0$ en la Fila 1. Se rechaza la hipótesis nula cuando el estadístico *(F)* observado supera al percentil correspondiente (en esta tabla se debe rechazar $H_o^A$).
- $H_0^B : \beta_1 = \beta_2 = \beta_3 = 0$ vs $H_1^A : (\exists i)(i = 1,2,3) / \beta_i \neq 0$ en la Fila 2, Se rechaza la hipótesis nula cuando el estadístico *(F)* observado supera al percentil correspondiente (en este caso se debe rechazar $H_o^B$).
- $H_0^{AB} : (\alpha\beta)_{ij} = 0$ vs $H_1^{AB} : (\alpha\beta)_{ij} \neq 0$ para algún $i \neq j$ en la Fila 3. Se rechaza la hipótesis nula cuando el estadístico *(F)* observado supera al percentil correspondiente, (en este caso se debe rechazar $H_o^{AB}$).

La respuesta correcta es a), ya que para asegurar una influencia tanto de factores principales (*A* y/o *B*) como de la interacción *AB* sobre el rendimiento en las tareas de logro, deberá comprobarse que el estadístico de prueba cae en la zona de rechazo del test. Esto último es equivalente a decir que el valor del estadístico *F* observado es mayor al el percentil correspondiente. Los errores que se evalúan en este ítem son los siguientes:

- De elegir como correcto el distractor b) estaría confundiendo el criterio para rechazar la hipótesis nula con el criterio de aceptación, para el factor B,
- De elegir como correcto el distractor c) estaría confundiendo criterio para rechazar la hipótesis nula de igualdad de efectos para las interacciones, pues aquí, por todo lo explicado existe efecto de la interacción de ambos factores sobre la respuesta.

Ítem 15. Los supuestos del ANOVA de dos factores, efectos fijos y completamente aleatorizado son:
a) Independencia de las observaciones, normalidad de las distribuciones, y aditividad
b) Independencia de las observaciones, igualdad de varianza y aditividad
c) **Independencia de las observaciones, normalidad de las distribuciones e igualdad de varianzas**

En esta pregunta se quiere evaluar la comprensión que el grupo posee sobre los supuestos que deben cumplir las observaciones para aplicar el modelo ANOVA. Estos supuestos son: independencia en la recogida de los datos, aleatoriedad (que las observaciones constituyen una muestra aleatoria de la población), que las varianzas $\sigma_{ij}^2$ sea la misma en todas las poblaciones (homocedasticidad), y que la variable se distribuya Normalmente en cada población con media $\mu_{ij}$ y desvío $\sigma$. De donde se desprende que la respuesta correcta es la c).

Sabemos que los modelos nunca se ajustarán exactamente a la realidad presente en los datos observados y la descomposición de la variabilidad presente en las observaciones, expresada en la ecuación (1), no es más que una relación algebraica. Pero la aplicación de los contrastes requiere del cumplimiento de ciertos supuestos para verificar si el modelo ajustado (2) describe de forma adecuada las observaciones. No se aconseja confiar en los resultados obtenidos con ANOVA, sin antes verificar el cumplimiento de los supuestos, algunos de los cuáles se pueden comprobar examinando los residuos (diferencias entre el valor observado y el ajustado con el modelo).. Si anotamos con $Y$ a la variable aleatoria $y_{ijk}$ los datos el problema, el residual que se corresponde a la j-ésima observación lo definimos como:

$$e_{ijk} = y_{ijk} - \hat{y}_{ijk}$$

Siendo $\hat{y}_{ijk}$ el valor ajustado mediante el modelo (2), en otras palabras es una "estimación" de la observación $y_{ijk}$. Fácilmente es posible probar que para cualquier subíndice ($jk$) la estimación viene dada por el promedio de los valores de la casilla correspondiente. Los errores que se evalúan en este ítem son los siguientes:

- De escoger como correcto el distractor a) no estaría considerando el supuesto de homocedasticidad que debe estar presente en el modelo. Este defecto en general sale a relucir con la gráfica de los residuos contra los valores ajustados. Por ejemplo, en algunas ocasiones la varianza de las observaciones se incrementa con la magnitud de las observaciones; en este caso, los residuos se harían mayores conforme las observaciones se hicieran más grandes (Montgomery, 2002). Si se violara el supuesto de la homogeneidad de varianzas sólo se afectará la prueba F ligeramente para un modelo balanceado (igual cantidad de observaciones por casilla) con efectos

fijos. Pero el problema podría volverse grave, si alguna de la varianzas se disparara en medida con respecto a las demás.

Para solucionar este problema comúnmente se busca aplicar transformaciones a las observaciones para estabilizar la varianza, y así volver a correr el mismo análisis, pero ahora con los datos transformados. La literatura indica distintos tipos de transformaciones de acuerdo con la distribución de las observaciones. Se deberá buscar empíricamente una transformación adecuada que estabilice u homogeinice la varianza. Además existen pruebas estadísticas para la igualdad de varianza, un procedimiento es el llamado: procedimiento de Bartlett (Montgomery, 2002).

- De escoger como correcto el distractor b) no estaría considerando el supuesto de normalidad que debe estar presente en el modelo pues se parte de ese argumento para aplicar las pruebas *F* y tomar la decisión estadística adecuada. Es posible realizar la verificación del supuesto de normalidad, haciendo un histograma de los residuos, y pruebas estadísticas de bondad de ajuste (test de Kolmogorov, por ejemplo). Si no se cumple, se debe transformar las variables para que se ajusten a la normal o bien usar métodos no paramétricos adecuados. Es importante destacar aquí que una desviación moderada de la normalidad no es motivo de gran preocupación en el análisis de la varianza con efectos fijos, pues la prueba F que se aplica solo se ve afectada ligeramente. Dicho de otro modo la prueba ANOVA es *robusta* con respecto al supuesto de normalidad (Montgomery, 2002).

El estudio HBSC sobre Conductas de los Escolares relacionadas con la Salud en su edición de 2006 realizó encuestas a niños de 38 países en edad escolar. Hemos extraído los datos de un grupo de niños a los que se les pidió que evaluaran la "calidad en la relación con su mejor amigo" en una escala de 0 a 10. También se clasificó a los niños según su "sexo" y la "autoestima" medida con la escala de Rosenberg considerándose dos valores (baja / alta autoestima). Los resultados obtenidos tras analizar los datos fueron los siguientes:

Tabla 1. Estadísticos descriptivos

Variable dependiente: CALIDAD EN LA RELACION CON EL AMIGO ESPECIAL

| SEXO | AUTOESTIMA | Media | Desv. típ. | N |
|---|---|---|---|---|
| CHICO | BAJA | 3,877 | ,9808 | 13 |
| | ALTA | 4,096 | ,6636 | 26 |
| CHICA | BAJA | 4,386 | ,6637 | 29 |
| | ALTA | 4,590 | ,4932 | 10 |

Tabla 2. Pruebas de efectos inter-sujetos.

| Fuente | Suma de cuadrados tipo III | gl | Media cuadrática | F | Significación |
|---|---|---|---|---|---|
| Modelo corregido | 4,090(a) | 3 | 1,363 | 2,721 | ,050 |
| Intersección | 1149,723 | 1 | 1149,723 | 2294,721 | ,000 |

| | | | | | |
|---|---|---|---|---|---|
| Sexo | 4,027 | 1 | 4,027 | 8,038 | ,006 |
| autoestima | ,716 | 1 | ,716 | 1,429 | ,236 |
| sexo * autoestima | ,001 | 1 | ,001 | ,002 | ,965 |
| Error | 37,076 | 74 | ,501 | | |
| Total | 1437,320 | 78 | | | |
| Total corregida | 41,166 | 77 | | | |

R cuadrado = ,099 (R cuadrado corregida = ,063)


Ítem 16. El tipo de análisis aplicado en este estudio es:

a) ANOVA de dos factores, efectos fijos y con medidas repetidas
b) **ANOVA de dos factores, efectos fijos y completamente aleatorizado**
c) ANOVA de un factor, efectos aleatorios y completamente aleatorizado

A través de este ítem se quiere evaluar si el grupo ha comprendido en general, el análisis de las salidas que ofrece un software estadístico, para este caso, se trata del SPSS; pero se particulariza la evaluación al tipo de modelo que se aplica en la resolución de este problema. La respuesta correcta es la b), pues se analiza la influencia a la variación de una variable dependiente, a saber: "calidad en la relación con el amigo especial", que tienen dos factores independientes: la autoestima y el sexo, en cada uno de estos factores se tienen en cuenta dos niveles: autoestima bajo/alto; sexo: hombre/mujer. De donde se trata de un ANOVA de dos factores (Autoestima y Sexo) con efectos fijos (alta/baja niveles para la autoestima; hombre/mujer niveles para el sexo), y completamente aleatorizado. Los errores que se evalúan en este ítem son:

- De escoger como correcto el distractor a) estaría confundiendo el modelo de medidas repetidas con el completamente aleatorizado. En el modelo de medidas repetidas entra en juego un factor de bloqueo, para evitar problemas que podrían surgir por la heterocedasticidad, ó para aquellos casos en los que no sea posible realizar todas las corridas en un mismo momento. Estos modelos, de medidas

repetidas, se utilizan mucho en psicología pues se suelen analizar reacciones como consecuencia de tratamientos en distintos tiempos. De donde es central la recogida de datos en distintos momentos, por esta razón se decide bloquear por momentos.

- De escoger como correcto el distractor c) estaría confundiendo el modelo unifactorial con el bifactorial. De elegirlo, solo tiene en cuenta un factor, claramente aquí los factores presentes en el modelo deben ser dos: Autoestima y Sexo.

Ítem 17. La condición experimental en la que el valor para la variable "calidad en la relación con el amigo especial" es menor es

a) **Chicos con baja autoestima**
b) Chicas con baja autoestima
c) Chicos con alta autoestima

Este ítem también evalúa si el grupo ha comprendido el análisis de las salidas de un software estadístico. La respuesta correcta es la a), se desprende del gráfico que otorga las *medias marginales estimadas de calidad en la relación con el amigo especial*. Se trata de un gráfico para evaluar la interacción entre los factores presentes en el modelo. Se dice que existe una interacción entre los factores, siempre que en el experimento la respuesta entre los niveles de un factor, no sea la misma para todos los niveles del otro factor. Dicho en otras palabras, siempre que el efecto de un factor A, dependa del nivel en que se elige el otro factor B, se observará que existe interacción.

Cada uno de los cuatro puntos en el gráfico nos indica el promedio de puntuación de la calidad en la relación con el amigo especial con respecto a una baja/alta autoestima, siendo chico ó chica. En el ejemplo no hay interacción, ya que los chicos tanto con baja como con alta autoestima, mantienen igual diferencia de la calidad en relación con el amigo especial como las chicas. Claramente la puntuación más baja la logran los chicos con baja autoestima. Los errores que se evalúan en este ítem son:

- De escoger como correcto el distractor b) estaría desconociendo la interpretación del gráfico de interacciones, ya que las chicas, tengan tanto alta como baja autoestima siempre mantienen los niveles más altos en relación con la calidad en la relación con el amigo especial.
- De escoger como correcto el distractor c) también desconocería la interpretación del gráfico de interacciones, pues los chicos con alta autoestima presentan un nivel más alto en cuanto a la calidad en relación con el amigo especial que los de baja autoestima.

Ítem 18. Entre las posibles conclusiones del estudio se encuentra, considerando un α=0,05:
a) Existe interacción entre el "sexo" y la "autoestima".
b) La "autoestima" influye sobre la "calidad en la relación con los amigos".
c) **El "sexo" influye sobre la "calidad en la relación con los amigos".**

Este ítem también evalúa si el grupo ha comprendido el análisis de las salidas de un software estadístico; pero se particulariza el análisis de los resultados obtenidos en la tabla ANOVA para dar una respuesta al problema, es decir se desea evaluar si interpreta correctamente los resultados del experimento.

Desde la tabla, claramente es posible responder a tres preguntas bien concretas: ¿influye el factor sexo sobre la calidad en la relación con el amigo especial? y/o ¿influye el factor autoestima sobre la calidad en la relación con el amigo especial? y/o ¿existe interacción entre ambos factores?. Con la finalidad de responder a estas preguntas planteamos el diseño del experimento. La respuesta correcta es la c), ya que en la última columna de la tabla arrojada por el ordenador, correspondiente a Significación, se observa un valor de significación o valor p igual a 0,006, el cual es mucho menor que el alfa considerado para el estudio, a saber: 0,05.

Aunque es posible reportar los resultados de una prueba de hipótesis es estableciendo que la hipótesis nula fue rechazada para un valor alfa especificado a priori, sin embargo, esta forma "*es con frecuencia inadecuada porque no le ofrece al responsable de la toma de decisiones idea alguna de si el valor calculado del estadístico de prueba apenas rebasó la región de rechazo o si se adentró bastante en la misma"* (Montgomery, 2002 p. 37). Por ello es que se ha adoptado muy extensamente en la práctica reportar el valor *p* ó probabilidad de que el estadístico que se prueba sea al menos igual al valor que se observa del estadístico, siempre que la hipótesis nula sea verdadera. Se decide rechazar la hipótesis nula siempre que el valor p observado es menor ó a lo sumo igual que el nivel de significación alfa establecido de antemano. Los errores que se evalúan en este ítem son:

- Si eligiera como correcto el distractor a) estaría respondiendo que la interacción sería significativa, es decir que el factor Sexo y el factor Autoestima inciden juntos sobre la calidad en la relación con el amigo especial. Esta afirmación es falsa ya que de la última columna se observa un *p*- valor superior al nivel 0,05 establecido.
- De elegir como correcto el distractor b) afirmaría que el factor Autoestima, ó sea el otro factor principal presente en el modelo incidiría sobre la calidad en la relación

con el amigo especial. Esta afirmación también es falsa, pues el *p*-valor que arroja el software resulta también superior a la puntuación 0,05.

Posteriormente se realizó una comparación en la misma variable ("calidad en la relación con el amigo especial") entre niños de familias monoparentales y biparentales. Los resultados son los siguientes:

Tabla 3. Prueba de muestras independientes

| | | Prueba de Levene | | Prueba T para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| | | | | | | | | | Inferior | Superior |
| CALIDAD EN LA RELACION CON EL AMIGO ESPECIAL | Se han asumido varianzas iguales | ,900 | ,346 | 4,437 | 87 | ,000 | ,7327 | ,1651 | ,4045 | 1,0609 |
| | No se han asumido varianzas iguales | | | 3,342 | 24,079 | ,003 | ,7327 | ,2192 | ,2803 | 1,1851 |

Ítem 19. De acuerdo con la prueba de Levene:

a) No se cumple el supuesto de homocedasticidad
b) **Se cumple el supuesto de homocedasticidad**
c) No se puede concluir sobre la homocedasticidad

A través de este ítem se quiere evaluar si el grupo ha comprendido el análisis de las salidas de un software estadístico; se particulariza sobre la interpretación para uno de los supuestos del modelo, la homocedasticidad. Este supuesto, como ya se ha afirmado anteriormente, se lo analiza a través de la llamada Prueba de Levene. De este modo al analizar las puntuaciones arrojadas en la Tabla 3 segunda columna, encabezada por Sig., el *p* – valor es 0,346. Estimamos que, luego de esta evidencia asumir desigualdad de varianzas sería un riesgo muy alto que se correría. Para este ítem la decisión recae sólo sobre la puntuación arrojada sobre el *p* – valor. De allí que la respuesta correcta sea la b), pues de no cumplirse la homocedasticidad, esa puntuación en la tabla quisiésemos fuese como máximo 0,05. Los errores que se evalúan para este ítem son:

- De elegir como correcto el distractor a) no haría una interpretación correcta de la Tabla 3, pues claramente de la misma se desprende que, cuando no se asuma igualdad de varianzas (ó sea si decidiera no se cumpliría el supuesto de homocedasticidad) para este problema por cada 100 muestras tomadas en iguales condiciones, en casi 35 de ellas, las varianzas no resultarían estadísticamente iguales. Si asumiera que este distractor fuese correcto el error que cometería

(estadísticamente hablando) sería riesgoso.

- Si eligiese el distractor c) cometería error pues, aunque generalmente cuando el valor *p* es menor que el nivel de significación no tomamos conclusiones, en este caso, como no rechazamos la hipótesis que las varianzas son diferentes, la única opción es asumir que son estadísticamente iguales

Ítem 20. La conclusión que podemos sacar de la prueba es:

a) **Hay diferencias entre los niños con familias monoparentales y biparentales en cuanto a la "calidad en la relación con el amigo especial"**
b) No hay diferencias entre los niños con familias monoparentales y biparentales en cuanto a la "calidad en la relación con el amigo especial"
c) No es posible concluir debido al incumplimiento del supuesto de aplicación.

A través de este ítem se quiere evaluar si el grupo ha comprendido el análisis de las salidas de un software estadístico, acerca de la interpretación de una prueba *t* para igualdad de medias.

Se trata de analizar una prueba *t* para igualdad de medias del tipo: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$, se trata de una prueba bilateral. En este problema interpretamos los valores medios poblacionales como: $\mu_1 =$ Puntuación media de la calidad en la relación con el amigo especial en niños con familias monoparentales; $\mu_2 =$ Puntuación media de la calidad en la relación con el amigo especial en niños con familias biparentales.

Bajo el supuesto de homocedasticidad (no se rechaza la hipótesis nula de igualdad de varianzas con una Sig. = 0,346), la puntuación arrojado por el *p*–valor en la columna Sig., bilateral es 0,000. De aquí se concluye como respuesta que hay suficiente evidencia estadística como para rechazar $H_0$ (que las puntuaciones medias de la calidad en relación con el amigo especial es la misma tanto para los niños que provengan de familias monoparentales como biparentales), de este modo, la respuesta correcta es la a). Los errores que se evalúan en este ítem son:

- De elegir como correcto el distractor b) el alumno no recuerda el criterio de decisión, pues está aceptando la hipótesis nula; es decir confunde el criterio de aceptación y rechazo.
- Si elige el distractor c) asume que las varianzas son diferentes. Por un lado, está interpretando incorrectamente el test de homogeneidad de varianzas, cuyo resultado

es que no puede rechazar la hipótesis de igualdad de varianzas. Por otro lado, incluso cuando las varianzas fuesen diferentes, el programa da como salida un contraste que es válido si las varianzas fuesen estadísticamente diferentes.

## 3.3. RESUMEN DE CONTENIDOS DEL CUESTIONARIO

Para sintetizar el análisis a priori, en la Tabla 3.2 mostramos un resumen de los contenidos estadísticos evaluados en el cuestionario, que hemos clasificado en relación a los objetos estadísticos principales. Esta tabla constituye una primera aproximación a la definición semántica de la variable "comprensión de objetos estadísticos en diseño experimental", es decir a la delimitación del significado institucional evaluado. De esta forma cumplimos el primer objetivo de nuestro trabajo.

En concreto, el contenido revisado del instrumento de evaluación, incluye los siguientes objetos estadístico:

- *Prerrequisitos:* Lecturas de tablas de una distribución (lenguaje) y cálculo de probabilidades (procedimiento), diferenciar muestras independientes y relacionadas (propiedades).
- *Intervalo de confianza:* construcción (procedimiento), interpretación (argumento)

Tabla 3.2. Contenidos evaluados en el cuestionario

| | Ítem | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Contenido evaluado | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| *Prerrequisitos* | | | | | | | | | | | | | | | | | | | | |
| -Lecturas de las tablas de la distribución y cálculo de probabilidades | | x | | | | | | | | | | | | | | | | | | |
| - Diferenciar muestras independientes y relacionadas | | | | | x | | | | | | | | | | | | | | | |
| *Intervalo de confianza* | | | | | | | | | | | | | | | | | | | | |
| - Construcción de un intervalo de confianza | | x | | | | | | | | | | | | | | | | | | |
| - Interpretación de un intervalo de confianza | | | x | | | | | | | | | | | | | | | | | |
| *Contraste de hipótesis* | | | | | | | | | | | | | | | | | | | | |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - Asignación de hipótesis | x | | | | | | x | | | | | | | | | | | | |
| - Diferenciar hipótesis nula y alternativa | x | | | | | | | | | | | | | | | | | | |
| - Valor p, nivel de significación | | | | | | | | x | | | | | | | | | | x | x |
| -Diferenciar potencia, nivel de significación, probabilidad de error II | | | | x | | x | | | | | | | | | | | | | |
| - Selección de un contraste adecuado | | | | | x | | | | | | | | | | | | | | |
| - Región crítica y de aceptación | | | | | | | x | | x | | | | | | | | | | |
| - Contraste unilateral y bilateral | | | | | | | x | | | | | | | | | | | | |
| - Regla de decisión | | | | | | | | x | | | x | | | | | | | x | x |
| *ANOVA* | | | | | | | | | | | | | | | | | | | |
| - Supuesto de independencia | | | | | x | | | | | | | | x | | | | | | |
| - Supuesto de normalidad e igualdad de varianzas | | | | | | | | | | | | x | x | | | | | x | |
| -Comprobación de supuestos | | | | | | | | | | | | | | x | | | | x | |
| - Modelo lineal correspondiente | | | | | | | | | x | | | | | | | | | | |
| - Cálculo del estadístico *F* | | | | | | | | | | x | x | | | | | | | | |
| - Interpretación de la tabla de ANOVA | | | | | | | | | | | x | | | | | x | x | | x |
| - Interacción | | | | | | | | | | x | | | | | | | x | | |
| - Modelo de efectos fijos | | | | | | | | | x | | x | | | | x | x | x | | |
| - Modelo de un factor | | | | | | | | | x | | x | | | | x | | | | |
| - Modelo de medidas repetidas | | | | | | | | | | x | x | | | | x | | | | |
| - Modelo de dos factores | | | | | | | | | | | x | | | | x | x | x | | x |

- *Contraste de hipótesis:* Asignación de hipótesis (procedimiento), diferenciar hipótesis nula y alternativa (argumento), determinación del valor p (propiedad), nivel de significación del test (propiedad), diferenciar potencia de nivel de significación y de probabilidad de error de tipo II (argumento), seleccionar un contraste adecuado (situación-problema), distinguir las regiones de rechazo y aceptación (procedimiento), regla de decisión (argumento) y finalmente, contrastes unilateral y bilateral (concepto).
- *ANOVA:* Supuestos de independencia (argumento), supuestos de homogeneidad y normalidad (argumento), comprobación de supuestos (procedimiento), modelo lineal adecuado (lenguaje), cálculo de estadístico *F* (procedimiento), interpretación de tablas (argumento), interacciones (definición), modelos de efectos fijos y

aleatorios (lenguaje), modelos de uno y dos factores (lenguaje) y modelos de medidas repetidas (lenguaje).

Creemos importante agregar que somos concientes que la definición semántica de la variable, objeto de medición, precisa refinarse si queremos conseguir en el futuro un cuestionario que cumpla los requisitos de validez y fiabilidad indicados en los estándares psicométricos y educativos (`por ejemplo, APA, AERA y NCME, 1999). Por ello, en una segunda fase futura de nuestro trabajo pensamos mejorar esta primera lista de contenidos, basándonos para ello en el análisis de contenido de una muestra de libros de texto usados en la enseñanza de diseño experimental a estudiantes universitarios.

No obstante, pensamos que la Tabla 3.2 proporciona un punto de partida, tanto para la revisión de dichos libros de texto, como para la selección de otros posibles ítems a utilizar en nuestro cuestionario.

# 4. RESULTADOS DEL ESTUDIO DE EVALUACIÓN

## 4.1. INTRODUCCIÓN

En este capítulo presentaremos los resultados de la prueba empírica del cuestionario para cumplir el segundo objetivo de nuestro trabajo, pues a partir de la prueba podremos revisar los ítems y descartar aquellos que no cumplan los requisitos de dificultad y discriminación. También revisaremos aquellos que tengan problemas en la redacción (que no sean bien comprendidos por los estudiantes) o los distractores (si las respuestas se reparte desigualmente en los distractores).

Finalmente, aunque no es uno de nuestros objetivos principales, el estudio de los ítems permitirá informar sobre algunos errores y dificultades de los estudiantes, y construir hipótesis sobre la comprensión de los objetos estadísticos incluidos en las variables que se incluyan en investigaciones futuras. Con ello informaremos sobre el significado personal logrado por los estudiantes.

Seguidamente, comenzamos describiendo la muestra y contexto. A continuación analizamos cada uno de los ítems y luego llevamos a cabo un estudio global, finalizando con algunas conclusiones al respecto, luego de analizar algunas de las propiedades psicométricas.

## 4.2. DESCRIPCIÓN DE LA MUESTRA Y DEL CONTEXTO EDUCATIVO

La muestra de participantes ha estado formada por un total de 93 alumnos de segundo año de la Licenciatura en Psicología de la Universidad de Huelva, que cursaban una asignatura de Análisis de Datos II, cuya profesora colaboró con nosotros facilitándonos el conjunto de ítems del cual se seleccionaron los del estudio, así como en la recogida de datos de los estudiantes. Los contenidos que abarca esta asignatura se organizan en ocho temas, que se recogen en la Figura 4.1, donde hemos incluido sólo el detalle de los cuatro primeros, que son los que se relacionan con los contenidos que pretendemos evaluar.

Esta asignatura se enmarca dentro del área de Métodos de Investigación en las Ciencias del Comportamiento que incluye varias asignaturas troncales y obligatorias de universidad y ofrece a los alumnos un cuerpo de conocimientos teóricos y prácticos cuya compresión y utilización les permitirá analizar y entender los

resultados obtenidos tanto en la investigación como en la práctica psicológica, así como generar sus propios análisis de datos en aquellos campos que resulten de su interés.

Figura 4.1. Contenidos estadísticos de la asignatura cursada por los estudiantes

TEMA 1. INTRODUCCIÓN: CONCEPTOS BÁSICOS DE LA INFERENCIA ESTADÍSTICA.
1. Marco general de la inferencia estadística en Psicología
2. Variables y su clasificación
3. Población, muestra, parámetro y estadístico
4. Distribución muestral
5. Estimación de parámetros: estimación puntual y estimación por intervalos
6. Contraste de hipótesis

TEMA 2. CONTRASTE DE HIPÓTESIS SOBRE MEDIAS Y PROPORCIONES
1. Contraste de hipótesis sobre una media
2. Contraste de hipótesis sobre dos medias
2.1 Dos medias independientes
2.2 Dos medias relacionadas
3. Contraste de hipótesis de una proporción
4. Contraste de hipótesis sobre dos proporciones

TEMA 3. ANÁLISIS DE LA VARIANZA DE UN FACTOR
3.1. El modelo lineal general
3.2. Modelos de ANOVA
3.3. La lógica del ANOVA
3.4. ANOVA de un factor

TEMA 4. ANÁLISIS DE LA VARIANZA DE DOS O MÁS FACTORES
4.1. La interacción entre factores
4.2. ANOVA de dos factores
4.3. Comparaciones múltiples
4.4. ANOVA de más de dos factores

TEMA 5. CORRELACIÓN

TEMA 6. REGRESIÓN LINEAL SIMPLE

TEMA 7. REGRESIÓN LINEAL MÚLTIPLE

TEMA 8. INFERENCIA BAYESIANA

Los estudiantes completaron los ítems como parte de una de las evaluaciones de la asignatura, que tiene un total de 6 créditos (60 horas lectivas), de los cuales la mitad son créditos prácticos, donde los alumnos realizan proyectos y estudios estadísticos en el laboratorio de informática, utilizando el programa SPSS. Mientras que en las sesiones de teoría se trabaja en gran grupo (de unos 100 alumnos), en las

sesiones prácticas los estudiantes trabajan en parejas con grupos de 40.

Los estudiantes habían cursado el año anterior Análisis de Datos en Psicología I, que tiene el mismo enfoque cuyos contenidos incluyen la estadística descriptiva univariante y bivariante, así como el cálculo de probabilidades simples y compuestas, teorema de Bayes y estudio de las distribuciones de probabilidad normal, *t* de *Student,* Chi cuadrado y *F*.

## 4.3. ANÁLISIS DE RESULTADOS POR ÍTEM

En primer lugar volvemos a presentar cada ítem, y a continuación de cada uno presentamos las tablas de frecuencia de respuestas a cada uno de ellos, comentando brevemente sobre los principales errores observados. Remitimos al capítulo 3 para la descripción del contenido de cada ítem. Se marca en negrita la respuesta correcta para cada caso, como ya se tiene en el Capítulo 3.

Ítem 1. Queremos conocer si los sujetos extrovertidos e introvertidos difieren en la puntuación media en autoestima y no disponemos de ninguna información previa. El tipo de hipótesis nula razonable que debo plantear es:

d) $\mu_I \leq \mu_E$
e) **$\mu_I = \mu_E$**
f) $\mu_I \geq \mu_E$

Tabla 4.1. Distribución de respuestas al Ítem 1

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 4 | 4,3% |
| **b** | **86** | **93,5%** |
| c | 2 | 2,2% |
| sin contestar | 1 | 1,1% |
| Total | 93 | 100,0% |

Este ítem da un 93,5% de respuestas correctas, por lo que ha resultado fácil para los estudiantes, quienes parecen comprender el planteamiento de las hipótesis en un contraste estadístico y la diferencia entre hipótesis estadística con la hipótesis de investigación. Vallecillos (1994) obtuvo un 68,6% de respuestas correctas en este ítem en su estudio, donde la pregunta también se hacía en un contexto de aplicación (78.6% en los alumnos de Psicología). Un 4,3% elige el distractor a) alumno confunde las hipótesis nula y alternativa, confusión que se ha encontrado en

investigaciones previas, como la de Vallecillos (1994); sólo un 2,2% elige el distractor c), de donde solo son dos los alumnos del total de la muestra que usarían sus ideas previas, pensando que las personas introvertidas tienen menos autoestima que las extrovertidas sin tener en cuenta el enunciado del problema. Finalmente, todos los alumnos parecen conocer la importancia que conlleva el contenido de este ítem ya que solo uno no lo contesta, es decir, todos declaran el tipo de hipótesis más adecuado. Los distractores funcionan de manera eficiente, ya que existe un cierto equilibrio entre ellos (4,3% y 2.2%).

Ítem 2. Un psicólogo escolar desea estimar la puntuación media de la población en un test de rendimiento de lectura. Para ello administra el test a una muestra de 36 estudiantes, obteniendo una media de 48 y desviación típica 10. Calcular, al nivel de significación α=0,05, los límites del intervalo de confianza para la puntuación en el test. (Error típico de la media: $S_d/\sqrt{n}$).

d) [38 , 58]
e) [15,33 , 80,66]
f) **[44,73 , 51,27]**

El porcentaje de respuestas correctas es acá el 78,3% (Tabla 4.2), lo que supone una buena capacidad de construcción de los intervalos, siendo los resultados próximos a los de Olivo(En prensa) quien encuentra 79% en un ítem similar, aunque en su caso no da la fórmula del error típico, mientras que nosotros la damos. En ambos casos parecen advertir bien el procedimiento de construcción de in intervalo de confianza, de acuerdo a los datos dados. Sin embargo el 19% no nos responde, hay de donde 18 alumnos de 93 que dudan al seguir un procedimiento de construcción del intervalo; mientras que en Olivo(En prensa), sólo un 1,6% deja el ítem en blanco, lo que podrían deberse a que los alumnos en esa investigación, siguen la carrera de Ingeniería y tienen mayores conocimientos de naturaleza algebraica. También en este ítem se observa un equilibrio entre los distractores (1,1% y 2,2%).

Tabla 4.2. Distribución de respuestas al Ítem 2

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 1 | 1,1% |
| b | 2 | 2,2% |
| **c** | **72** | **78,3%** |
| sin contestar | 18 | 19,6% |
| Total | 93 | 100,0% |

Ítem 3. En relación con la pregunta anterior, también podríamos afirmar que:

d) **Si extrajéramos 100 muestras y calculáramos en cada una el intervalo de confianza para la media, en 95 de ellos, se encontraría el verdadero valor del parámetro.**

e) La puntuación del test está en el intervalo estimado con una probabilidad del 95%

f) Si extrajéramos 100 muestras y calculáramos en cada una el intervalo de confianza, en 95 de los intervalos, se encontraría la puntuación de los sujetos de la muestra.

Tabla 4.3. Distribución de respuestas al Ítem 3

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| **a** | **10** | **10,9%** |
| b | 35 | 38,0% |
| c | 4 | 4,3% |
| sin contestar | 44 | 47,8% |
| Total | 93 | 100,0% |

De acuerdo con el porcentaje de respuestas correctas en el ítem último anterior (Tabla 4.2) pocos alumnos presentan dificultades para construir un intervalo de confianza, es decir comprenden el procedimiento, pero son muy pocos los que realmente comprenden su definición (10,9% Tabla 4.3). Olivo(En prensa) encuentra un 36,5% de interpretaciones correctas en un ítem similar, en estudiantes de ingeniería, aunque también encontró un alto porcentaje de fallos, a pesar de ser estudiantes con mayor base matemática que los nuestros. Casi un 50% para nuestra muestra parece no saber la respuesta. La interpretación bayesiana (cambio de condición y condicionado en la probabilidad condicional que define el coeficiente de confianza) la da un 38% mientras que en la de Olivo(En prensa) el porcentaje fue del 34%. La interpretación bayesiana del Intervalo de Confianza es un error que se puede recoger de varias investigaciones (Batanero, 2000b; Batanero y Díaz , 2005, Olivo(En prensa)). Hay un pronunciado desequilibrio entre los distractores (38% y 4,3%).

Ítem 4. Supongamos un contraste bilateral sobre la media, siendo la variable estudiada la inteligencia. Para $H_0$: $\mu = 100$, $H_1$: $\mu = 110$, $\alpha = 0,05$ y $\beta = 0,4406$. ¿Cuál es la probabilidad de rechazar la $H_0$ cuando no es 'cierta'?

d) 0,05

e) 0,4406

f) **0,5594**

En este ítem obtenemos 76,1% de aciertos (Tabla 4.4); mientras que Vallecillos (1994) obtiene 22,9% de respuestas correctas en un ítem similar (20% en

Psicología), aunque ella da los distractores mediante símbolos; por ello puede ser más difícil que en nuestro caso, en que están dados en valor numérico. Un 10,9% confunde los errores tipo I y tipo II (el porcentaje de alumnos con este error fue el 6,2% en la investigación de Vallecillos (1994) en la muestra general y el 7,1% en los estudiantes de Psicología). Un 6,5% de nuestros estudiantes confunde β y 1- β; (mientras que los porcentajes fueron el 16,1% en Vallecillos (1994) en la muestra general y el 30% en los estudiantes de Psicología). Vallecillos (1994), ya alerta que esta confusión entre β y 1- β , está bastante extendida especialmente entre estudiantes de Psicología. Hay equilibrio entre los distractores (10,9% y 6.5%) con los que no dan respuesta (7,6%).

Tabla 4.4. Distribución de respuestas al Ítem 4

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 10 | 10,9% |
| b | 6 | 6,5% |
| **c** | **70** | **76,1%** |
| sin contestar | 7 | 7,6% |
| Total | 93 | 100,0% |

Ítem 5. Una maestra cree que unas nuevas actividades de lectura ayudarán a mejorar la capacidad lectora de los niños de primaria. La maestra tiene una clase con 21 niños a los que le pasa la prueba de lectura DRP (Degree of Reading Power), para conocer el nivel del que parten. Después realiza estas actividades en clase durante 8 semanas. Al final del período vuelve a pasarles la prueba. De las siguientes técnicas, ¿cuál debería aplicar los investigadores para comprobar si las actividades modifican las capacidades lectoras?

d) Contraste de hipótesis sobre dos medias independientes
e) **Contraste de hipótesis sobre dos medias relacionadas**
f) ANOVA de un factor completamente aleatorizado

Tabla 4.5. Distribución de respuestas al Ítem 5

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 3 | 3,3% |
| **b** | **76** | **82,6%** |
| c | 3 | 3,3% |
| sin contestar | 11 | 12,0% |
| Total | 93 | 100,0% |

Con este ítem pretendemos dar aportaciones nuevas sobre la comprensión de los alumnos acerca de elementos estadísticos básicos del Diseño Experimental. Notamos (Tabla 4.5) que el 82,6% de los alumnos contestan correctamente al ítem, lo que nos dice que el alumno es capaz de escoger el modelo que mejor se adecua al

problema. Al elegir la respuesta a), un 3,3% estaría confundiendo el concepto de independencia de muestras. Al elegir como correcta la respuesta c), en el mismo porcentaje que el otro distractor, estaría desconociendo los supuestos mínimos de ANOVA de un factor completamente aleatorizado, el cual no es otro que el de independencia entre los individuos a los cuáles se les aplica los distintos tratamientos. Hay un 12% que no contesta la pregunta. Encontramos equilibrio entre los distractores (3,3% para cada uno).

Ítem 6. Supongamos que conocemos la 'verdad absoluta' sobre la eficacia de dos tratamientos (A y B), y sabemos que existen diferencias en la efectividad de ambos para curar la depresión. Un investigador que realice un estudio y parta de la hipótesis '*no existen diferencias en la efectividad de los tratamientos A y B para curar la depresión*' cometerá un *error tipo II* cuando:

d) Concluya que A y B no son efectivos para curar la depresión
e) **Concluya que A y B no difieren en su efectividad para curar la depresión**
f) Concluya que A y B difieren en su efectividad para curar la depresión

Tabla 4.6. Distribución de respuestas al Ítem 6

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 1 | 1,1% |
| **b** | **60** | **65,2%** |
| c | 17 | 18,5% |
| sin contestar | 15 | 16,3% |
| Total | 93 | 100,0% |

Se tiene un 65,2% (Tabla 4.6) de respuestas correctas. Vallecillos (1994) obtiene 22,9% de respuestas correctas en un ítem similar (20% en Psicología), aunque ella da los distractores mediante símbolos lo que puede ser más difícil que en nuestro caso, en que están dados en valor numérico. Un 18,5% piensa que hay error cuando no existe, confunde la definición de Error Tipo II. Sólo un 1,1% no interpretan enunciado, ya que el problema afirma que se conoce que ambos medicamentos han sido probados separadamente, resultando efectivos. Hay un alto porcentaje, 16,3%, de alumnos que no responden al ítem, lo que conlleva el desconocimiento sobre la diferencia entre hipótesis estadísticas e hipótesis de investigación; y entre Error Tipo I y Tipo II. Estos son los que dudan sobre el significado del Error Tipo II. Se observa un fuerte desequilibrio entre los distractores (1,1% y 18,5%).

Ítem 7. La puntuación típica correspondiente a una $\alpha = 0,01$ en un contraste unilateral derecho es:

d) **2,33**
e) -2,33
f) 3,10

El porcentaje de aciertos para este ítem es del 34,8% (Tabla 4.7). La mayoría se ha decidido por escoger el distractor b), lo que hacen es tomar el test unilateral izquierdo o lo que es lo mismo confunde hipótesis nula y alternativa o bien la región crítica y de aceptación en un contraste unilateral. Este bajo porcentaje de aciertos, lo significa como un ítem de difícil resolución. Tauber (2001) en su investigación sobre la construcción de significados acerca de la distribución Normal, nos muestra un ítem similar, pero con otro fin (p. 181) observar el conocimiento de la gráfica de la función de densidad. Un 46,7% elige el distractor b); aunque lo más claro sería elegir en un más alto porcentaje el distractor c) (5,4%), por estar del mismo lado del contraste que se da en el enunciado. Existe un gran desequilibrio a la hora de elegir alguno de los distractores propuestos. Un alto porcentaje responde (85%). Hay un fuerte desequilibrio entre los distractores.

Tabla 4.7. Distribución de respuestas al Ítem 7

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| **a** | **32** | **34,8%** |
| b | 43 | 46,7% |
| c | 5 | 5,4% |
| Sin contestar | 13 | 14,1% |
| Total | 93 | 100,0% |

Ítem 8. La prueba que se utiliza para comprobar el supuesto de esfericidad en un estudio con una V.I. con medidas repetidas es:

d) La prueba de Levene
e) La prueba de Kolmogorov
f) **La prueba de Mauchly**

Tabla 4.8. Distribución de respuestas al Ítem 8

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 5 | 5,4% |
| b | 5 | 5,4% |
| **c** | **72** | **78,3%** |
| Sin contestar | 11 | 12,0% |
| Total | 93 | 100,0% |

Visualizamos en este ítem, no tratado en la investigación previa, un 78,3% de respuestas correctas (Tabla 4.8). De acuerdo con el porcentaje de respuestas correctas elegidas, las no respuestas (12%), como por las incorrectas (10,8%),

estimamos que los alumnos logran comprender las diversas pruebas estadísticas que se utilizan para validar los diseños de experimentos, y saben elegir las que se aplican a los diseños con medidas repetidas. Debemos aceptar el fuerte equilibrio entre los distractores (5,4% para ambos).

Ítem 9. ¿Cuál de las siguientes hipótesis está bien formulada?:
d) Ho: μ= 3; Hi: μ ≠ 4
e) Ho: μ= 3; Hi: μ ≥ 3
f) **Ho: μ= 3; Hi: μ ≠ 3**

Tabla 4.9. Distribución de respuestas al Ítem 9

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 3 | 3,3% |
| b | 1 | 1,1% |
| **c** | **87** | **94,6%** |
| sin contestar | 2 | 2,2% |
| Total | 93 | 100,0% |

Este ítem resultó muy sencillo a los estudiantes (94,6% de respuestas correctas, ver Tabla 4.9) que conocen bien el planteamiento de hipótesis. Vallecillos (1994) encontró un 56% de respuestas correctas en un ítem similar (77,1% en Psicología), aunque en esa investigación se pregunta por la negativa (p. 319). Es de destacar que el mayor porcentaje de respuestas correctas, lo obtiene la especialidad de Matemáticas, y que el segundo mayor porcentaje lo tienen los alumnos de la carrera de Psicología; esto confirma los porcentajes altos de respuestas correctas obtenidos en esta investigación.

Un 3,3% no considera las hipótesis como complementaria (Vallecillos (1994) no incluye este distractor). Se trata de un total de 3 alumnos sobre un total de 93 en nuestra investigación. Es destacable el equilibrio y bajo porcentaje entre los distractores (3,3% y 1,1%). Finalmente es muy bajo el porcentaje que no responde.

Ítem 10. Cuando realizamos un contraste, la regla de decisión nos lleva a rechazar la hipótesis nula siempre que:
d) El estadístico de contraste caiga en la región de rechazo
e) La probabilidad asociada al estadístico de contraste (el valor de significación) sea menor que el valor de alfa
f) **a) y b) son correctas**

Obtuvimos 59,8% de respuestas correctas, Vallecillos (1994) encuentra un 33,7% en un ítem relacionado con éste (42,9% en estudiantes de Psicología), pero donde se da el valor obtenido del estadístico y se pide cuál es la decisión (qué hipótesis se debe aceptar o rechazar) por lo cual el enunciado es más complejo. También en otro muy similar donde pregunta qué se concluye si el resultado en un contraste de hipótesis es significativo, Psicología presenta un 30% de aciertos, para su par presenta un 11,4%. Esta diferencia podría darse también por la complejidad del enunciado. En nuestra investigación, a pesar de ser correctas las dos respuestas a) y b) a la vez, sólo un 2,2% decide por el distractor b) y un 26,1% por el distractor a).

Tabla 4.10. Distribución de respuestas al Ítem 10

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 24 | 26,1% |
| b | 2 | 2,2% |
| **c** | **55** | **59,8%** |
| sin contestar | 12 | 13,0% |
| Total | 93 | 100,0% |

Esta diferencia nos está mostrando la mayor complejidad que tiene para los alumnos relacionar el cálculo de probabilidades en el contraste de hipótesis (distractor b), que comparar si un valor numérico es ó no elemento de un subconjunto de la recta real (distractor a). Un muy bajo porcentaje (13%) que no da una respuesta sobre la regla de decisión.

Ítem 11. El modelo lineal utilizado para representar las fuentes de variabilidad presentes en un ANOVA de un factor, efectos fijos completamente aleatorizado es:

d) $\mathbf{Y_{ij}=\mu + \alpha_j + E_{ij}}$

e) $Y_{ij}=\mu + \alpha_j + \beta_i + E_{ij}$

f) $Y_{ij}=\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + E_{ijk}$

Tabla 4.11. Distribución de respuestas al Ítem 11

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| **a** | **49** | **53,3%** |
| b | 11 | 12,0% |
| c | 4 | 4,3% |
| sin contestar | 29 | 31,5% |
| Total | 93 | 100,0% |

En este ítem obtuvimos un 53,3% (Tabla 4.11) de respuestas correctas; un 31,5% no conteste el ítem, no establece una correspondencia entre la expresión verbal y simbólica del modelo de análisis de varianza. En este ítem tratamos el análisis de varianza y queremos evaluar la comprensión del grupo de alumnos para asociar un modelo estadístico, de acuerdo con el estudio que se desea realizar, si bien más de la mitad de la población evaluada parece comprender el objetivo propuesto para esta unidad; somos concientes que se debe continuar tratando el tema con mayor profundidad, sobre todo al tratarse de alumnos de la carrera Psicología, los cuales no tienen costumbre de manejar en su vida académica con modelos matemáticos. El bajo porcentaje que elige el distractor c) muestra que los alumnos distinguen que en este problema no se analizaría interacción entre factores. Se trata también de un ítem que nuestra investigación aporta.

Ítem 12. Si en un ANOVA de un factor, y medidas repetidas encuentro que la F empírica u observada toma un valor de 8,16 esto quiere decir que:

d) CM intergrupos / CM error = 8,16
e) **CM intrasujetos / CM error = 8.16**
f) CM intergrupos / CM intrasujetos = 8,16

Tabla 4.12. Distribución de respuestas al Ítem 12

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 58 | 63,0% |
| **b** | **7** | **7,6%** |
| c | 15 | 16,3% |
| sin contestar | 13 | 14,1% |
| Total | 93 | 100,0% |

El objetivo de este ítem es dar una medida empírica para comparar con una puntuación teórica, y dar una decisión del cambio de los sujetos ante un estímulo, ó una práctica a través del tiempo. Este ítem resultó muy difícil (7,6 % de respuestas correctas, Tabla 4.12) de modo que el alumno no parece comprender el argumento para interpretar esta puntuación en la tabla ANOVA. Al elegir en un 63% el distractor a) eso nos da una idea de la incomprensión del método de cálculo, por lo que recomendamos insistir sobre estos conceptos. Se observa desequilibrio entre los distractores (63% y 16,3%). No contamos con otra investigación de este estilo, de manera que podamos comparar respuesta de alumnos a una pregunta similar. Se trata también de un aporte de nuestra investigación.

Ítem 13 y 14. Se quiere estudiar el efecto de ciertas variables motivacionales sobre el rendimiento en tareas de logro. Se manipularon dos variables: "tipo de entrenamiento motivación" (A1: instrumental; A2: atribucional y A3: control) y "clima de clase" (B1: cooperativo; B2: competitivo y B3: individual). Se seleccionaron a 45 sujetos y se dividieron en grupos para cada condición experimental.
A continuación presentamos la tabla del ANOVA incompleta. A partir de la información de la tabla contesta a las preguntas del Ítem 13 y del Ítem 14.

| Fuente de variación | SC | Gl | MC | F |
|---|---|---|---|---|
| Factor A | 70 | | | |
| Factor B | | | 20 | |
| Interacción AB | | | | 3,91 |
| Error | 46 | | 1,278 | |
| Total | 176 | 44 | | |

Ítem 13. (TEMA A)
El valor de la media cuadrática para el factor A es:
d) 5
e) **35**
f) 15,65

Tabla 4.13(a). Distribución de respuestas al Ítem 13(A)

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| A | 0 | 0,0% |
| **B** | **34** | **73,9%** |
| C | 2 | 4,3% |
| sin contestar | 10 | 21,7% |
| Total | 46 | 100,0% |

En el Capítulo 3 de este trabajo, presentamos la Tabla ANOVA completa, que proporciona la respuesta, de manera que la misma podría ser nuevamente revisada aquí. Encontramos un 73,9% de respuestas correctas (ver Tabla 4.13(a)), lo cual indica un alto porcentaje de respuestas correctas. Pero comparativamente es alto el porcentaje de de alumnos que no responde (21,7%), lo que confirma otra vez la dificultad de comprensión del cálculo. Esta dificultad también podría ser atribuida a la diversidad presente en el lenguaje. Hay equilibrio a la hora de escoger un distractor (0% y 4,3%).

Ítem 13. (TEMA B)
El valor del sumatorio cuadrado para el factor B es:
a) 15,65
b) 35
c) **40**

Tal como se indicara oportunamente en el Capítulo 3, pretendemos en este espacio, interpretar las respuestas de los alumnos dadas a ambos ítems para cada tema de la evaluación. Hemos decidido tener resultados parciales sobre los distintos

procedimientos involucrados para completar una tabla ANOVA de una vía. Se observa un 63,8% de respuestas correctas (Tabla 4.13(b)), un equilibrio al elegir los distractores, pero un algo porcentaje que no responde (29,8%). El grupo pareciera comprender mejor cuando nos referimos a la frase "sumatorio cuadrado" que cuando hablamos de media cuadrática.

Tabla 4.13(b). Distribución de respuestas al Ítem 13(b)

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| a | 0 | 0,0% |
| b | 3 | 6,4% |
| **c** | **30** | **63,8%** |
| sin contestar | 14 | 29,8% |
| Total | 47 | 100,0% |

Ítem 14. Una de las conclusiones del estudio sería (alfa = 0,05)
d) **Hay efecto del factor A ("entrenamiento") sobre el rendimiento en tareas de logro**
e) No hay efecto del factor B ("clima de clase") sobre el rendimiento en tareas de logro
f) No hay interacción de los factores

Tabla 4.14. Distribución de respuestas al Ítem 14

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| **a** | **24** | **26,1%** |
| b | 6 | 6,5% |
| c | 5 | 5,4% |
| sin contestar | 58 | 63,0% |
| Total | 93 | 100,0% |

Encontramos un muy bajo porcentaje de respuestas correctas 26,1% (ver Tabla 4.14), y un alto porcentaje de no respuestas (63%). Esto nos indica la dificultad en este ítem encontrada por los alumnos. De acuerdo a los porcentajes recogidos en el ítem 13, para ambos temas, allí emerge que a las hora de realizar los cálculos para determinar las puntuaciones de la tabla, y de discriminar que representa cada puntuación parecería comprenderlo aunque parcialmente; pero cuando se trata de interpretar (dar argumentos de) esas puntuaciones se encuentra frente a una dificultad, por lo que recomendamos insistir en estos conceptos. También esta es una unidad que aporta nuestra investigación. A pesar de la dificultad, observamos un equilibrio entre los distractores (6,5% y 5,4%).

Ítem 15. Los supuestos del ANOVA de dos factores, efectos fijos y completamente aleatorizado son:

d) Independencia de las observaciones, normalidad de las distribuciones, y aditividad
e) Independencia de las observaciones, igualdad de varianza y aditividad
f) **Independencia de las observaciones, normalidad de las distribuciones e igualdad de varianzas**

Tal como lo indicáramos anteriormente, los modelos nunca se ajustarán exactamente a la realidad presente en los datos observados y no se aconseja confiar en los resultados obtenidos con ANOVA, sin antes verificar el cumplimiento de los supuestos. Aquí nos encontramos con un 51,1% de respuestas correctas (ver Tabla 4.15), lo que muestra una dificultad moderada encontrada por los alumnos en el ítem, también hay un alto porcentaje de no respuesta (31,5%) y una diferencia significativa a la hora de elegir un distractor adecuado, un 13% para el a) y un 5,4% para el b), lo que marca un fuerte desequilibrio. Tampoco existen investigaciones que nos aporten un estado de la cuestión sobre la comprensión de este tipo de conceptos relacionados específicamente con la prueba ANOVA.

Tabla 4.15. Distribución de respuestas al Ítem 15

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| A | 12 | 13,0% |
| B | 5 | 5,4% |
| **C** | **47** | **51,1%** |
| sin contestar | 29 | 31,5% |
| Total | 93 | 100,0% |

El estudio HBSC sobre Conductas de los Escolares relacionadas con la Salud en su edición de 2006 realizó encuestas a niños de 38 países en edad escolar. Hemos extraído los datos de un grupo de niños a los que se les pidió que evaluaran la "calidad en la relación con su mejor amigo" en una escala de 0 a 10. También se clasificó a los niños según su "sexo" y la "autoestima" medida con la escala de Rosenberg considerándose dos valores (baja / alta autoestima). Los resultados obtenidos tras analizar los datos fueron los siguientes:

Tabla 2. Estadísticos descriptivos

Variable dependiente: CALIDAD EN LA RELACION CON EL AMIGO ESPECIAL

| SEXO | AUTOESTIMA | Media | Desv. típ. | N |
|---|---|---|---|---|
| CHICO | BAJA | 3,877 | ,9808 | 13 |
| | ALTA | 4,096 | ,6636 | 26 |
| CHICA | BAJA | 4,386 | ,6637 | 29 |
| | ALTA | 4,590 | ,4932 | 10 |

Tabla 2. Pruebas de efectos inter-sujetos.

| Fuente | Suma de cuadrados tipo III | gl | Media cuadrática | F | Significación |
|---|---|---|---|---|---|
| Modelo corregido | 4,090(a) | 3 | 1,363 | 2,721 | ,050 |
| Intersección | 1149,723 | 1 | 1149,723 | 2294,721 | ,000 |
| sexo | 4,027 | 1 | 4,027 | 8,038 | ,006 |

| | | | | | |
|---|---|---|---|---|---|
| autoestima | ,716 | 1 | ,716 | 1,429 | ,236 |
| sexo * autoestima | ,001 | 1 | ,001 | ,002 | ,965 |
| Error | 37,076 | 74 | ,501 | | |
| Total | 1437,320 | 78 | | | |
| Total corregida | 41,166 | 77 | | | |

R cuadrado = ,099 (R cuadrado corregida = ,063)


Ítem 16. El tipo de análisis aplicado en este estudio es:

d) ANOVA de dos factores, efectos fijos y con medidas repetidas
e) **ANOVA de dos factores, efectos fijos y completamente aleatorizado**
f) ANOVA de un factor, efectos aleatorios y completamente aleatorizado

Tabla 4.16. Distribución de respuestas al Ítem 16

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| A | 7 | 7,5% |
| **B** | **64** | **68,8%** |
| C | 5 | 5,4% |
| sin contestar | 17 | 18,3% |
| Total | 93 | 100,0% |

En este ítem, el primero de la serie de ítems que le siguen, se quiere evaluar si el grupo ha comprendido en general, el análisis de las salidas que ofrece un software estadístico; es decir si comprende los resultados arrojados desde un ordenador, luego de aplicar un modelo adecuado a un conjunto de datos tomados después de realizar el diseño experimental. El porcentaje de respuestas correctas es alto, pareciera que el grupo comprende este objeto. Observamos que un 68,8% de los alumnos responden correctamente a este ítem (ver Tabla 4.16), pero también hay un 18,3% que no lo responde. Los porcentajes para la elección de distractores están equilibrados, un

7,5% y un 5,4%. Los alumnos han encontrado este ítem, de acuerdo con los porcentajes analizados, bastante fácil. No hay antecedentes para compararlo con otras investigaciones, se trata también de un aporte dado por nuestra investigación.

Ítem 17. La condición experimental en la que el valor para la variable "calidad en la relación con el amigo especial" es menor es

d) **Chicos con baja autoestima**
e) Chicas con baja autoestima
f) Chicos con alta autoestima

Tabla 4.17. Distribución de respuestas al Ítem 17

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| **a** | **79** | **85,9%** |
| b | 3 | 3,3% |
| c | 1 | 1,1% |
| sin contestar | 10 | 10,9% |
| Total | 93 | 100,0% |

Para este ítem el alumno debería ser capaz de interpretar desde un gráfico la respuesta correcta. Se trata de un gráfico para evaluar la interacción entre los factores presentes en el modelo; es decir deberá observar toda la salida del ordenador dada, y de eso deducir que le sirve para responder. Se han obtenido un 85,9% de respuestas correctas (ver Tabla 4.17), se trata de un alto porcentaje; pareciera les ha resultado fácil responder correctamente a este ítem. De los que contestan, la gran mayoría lo hace correctamente, pues tan sólo un 4,4% responde equivocademante. Hay equilibrio entre los distractores (3,3% y 1,1%). Debemos tener en cuenta que existe un 10,9% de alumnos que no responden a esta pregunta.

Ítem 18. Entre las posibles conclusiones del estudio se encuentra, considerando un $\alpha=0.05$:

d) Existe interacción entre el "sexo" y la "autoestima".
e) La "autoestima" influye sobre la "calidad en la relación con los amigos".
f) **El "sexo" influye sobre la "calidad en la relación con los amigos".**

Este ítem presenta una dificultad moderada para los estudiantes, ya que sólo hubo un 41,3% de respuestas correctas (ver Tabla 4.18). Hay un 34,8% que no responde correctamente al ítem, eligen alguno de los distractores a) ó b). Se evidencia desequilibrio a la hora de elegir un distractor. Este es otro ítem donde se les solicita alguna conclusión acerca de las puntuaciones, en este caso obtenido de la tabla ANOVA, pero que nos ha arrojado un ordenador (les resultad difícil

argumentar). Advertimos sobre este tipo de ítem, reforzar las prácticas en las sesiones de clase, para su mejor comprensión pues ésta es fundamental al momento de dar una respuesta al problema. Un 25% no responde.

Posteriormente se realizó una comparación en la misma variable ("calidad en la relación con el amigo especial") entre niños de familias monoparentales y biparentales. Los resultados son los siguientes:

Tabla 3. Prueba de muestras independientes

| | | Prueba de Levene | | Prueba T para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | T | Gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| | | | | | | | | | Inferior | Superior |
| CALIDAD EN LA RELACION CON EL AMIGO ESPECIAL | Se han asumido varianzas iguales | ,900 | ,346 | 4,437 | 87 | ,000 | ,7327 | ,1651 | ,4045 | 1,0609 |
| | varianzas iguales | | | 3,342 | 24,079 | ,003 | ,7327 | ,2192 | ,2803 | 1,1851 |

Tabla 4.18. Distribución de respuestas al Ítem 18

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| A | 22 | 23,9% |
| B | 10 | 10,9% |
| **C** | **38** | **41,3%** |
| sin contestar | 23 | 25,0% |
| Total | 93 | 100,0% |

Ítem 19. De acuerdo con la prueba de Levene:

d) No se cumple el supuesto de homocedasticidad
e) **Se cumple el supuesto de homocedasticidad**
f) No se puede concluir sobre la homocedasticidad

Tabla 4.19. Distribución de respuestas al Ítem 19

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| A | 27 | 29,3% |
| **B** | **41** | **44,6%** |
| C | 2 | 2,2% |
| sin contestar | 23 | 25,0% |
| Total | 93 | 100,0% |

Observamos también aquí un bajo porcentaje de respuestas correctas, 44,6% (ver Tabla 4.19), un alto porcentaje ha optado por algún distractor incorrecto, un

31,5%. También presenta un alto porcentaje sin ser respondido, un 25%. Hay que tener en cuenta además la presencia de un desequilibrio importante para la elección de los distractores, se trata de un 29,3% que no hace una interpretación correcta de la Tabla 3 (Prueba de muestras independientes) y un 2,2% que también yerra cuando debe leer la tabla arrojada. También se analiza la comprensión del grupo para el análisis de las salidas de un software estadístico.

Ítem 20. La conclusión que podemos sacar de la prueba es:

d) **Hay diferencias entre los niños con familias monoparentales y biparentales en cuanto a la "calidad en la relación con el amigo especial"**

e) No hay diferencias entre los niños con familias monoparentales y biparentales en cuanto a la "calidad en la relación con el amigo especial"

f) No es posible concluir debido al incumplimiento del supuesto de aplicación.

Éste ítem tampoco les ha resultado fácil al grupo, sólo lo responde correctamente el 27,2% (ver Tabla 4.20). También observamos un gran desequilibrio entre los que eligen alguno de los distractores b) ó c).

Tabla 4.20. Distribución de respuestas al Ítem 20

| Respuesta | Frecuencia absoluta | Porcentaje |
|---|---|---|
| **A** | **25** | **27,2%** |
| B | 33 | 35,9% |
| C | 1 | 1,1% |
| sin contestar | 34 | 37,0% |
| Total | 93 | 100,0% |

Todavía un más alto porcentaje que para el ítem último anterior no responde al ítem, (37%). La respuesta también requiere una comprensión de la interpretación (dar argumentos), para responder a la pregunta de investigación que se plantea en este diseño experimental, en particular se trata de analizar una prueba *t* para igualdad de medias. Una gran cantidad de alumnos 35,9% han decidido como correcto el distractor c), esto indicaría que no recuerda el criterio de decisión, pues está aceptando la hipótesis nula; es decir confunde el criterio de aceptación y rechazo. Sin embargo, un mínimo porcentaje de alumnos decide como correcto del distractor c) 1,1%, (un alumno, sobre un total de 93) lo que indicaría que está interpretando incorrectamente el test de homogeneidad de varianzas, cuyo resultado es que no puede rechazar la hipótesis de igualdad de varianzas. Por otro lado, incluso cuando las varianzas fuesen diferentes, el programa da como salida un contraste que es válido si las varianzas difieren.

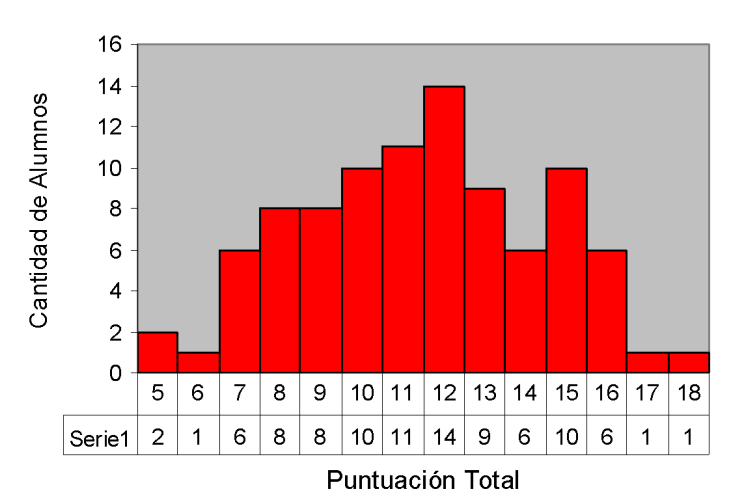
## 4.4. ANÁLISIS DE RESULTADOS GLOBALES DEL CUESTIONARIO

### 4.4.1 Puntuación total

Además de estudiarse la puntuación de cada ítem, es interesante estudiar el número de respuestas correctas de cada estudiante, que nos da una idea de la proporción de elementos de significado adquiridos en relación con los elementos pretendidos por la institución al elaborar el instrumento de evaluación. Se trata, por cada alumno, de la suma total de ítems contestados correctamente. Para estudiar esto, se puntuó con 1 cada respuesta correcta, sumando todas estas puntuaciones.

Mostramos a continuación un vector formado por 93 elementos, donde cada componente presenta la puntuación total de cada alumno para la prueba. Los valores para los mismos oscilaran entre 0 y 20; siendo el valor 0 adjudicado para aquellos que no contestaron correctamente algún ítem, y el valor 20 para aquellos que hubiesen contestado correctamente todos los ítems propuestos.

Fig.4.2. Histograma de la puntuación total



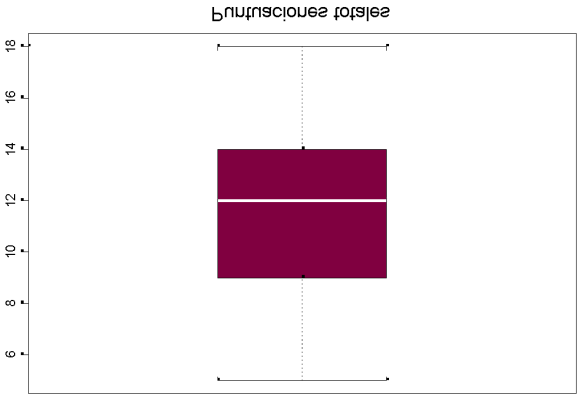

En la Fig. 4.2, se muestra la distribución de la puntuación en el cuestionario. Observamos que las puntuaciones efectivamente obtenidas por los alumnos varían de 5 a 18 puntos. Un 50% de la muestra contestó correctamente entre 9 y 14 ítems (observar primer y tercer cuartil, en Tabla 4.21). Este resultado es bastante bueno, ya que el 10 representa la mitad de los ítems, y significa que las dos terceras partes de los alumnos ha resuelto al menos casi la mitad de la prueba. La puntuación mínima obtenida en la prueba (5) significa que el total de los alumnos ha resuelto al menos una cuarta parte de la prueba. Un alumno típico resuelve el 60% de la prueba, ya que

la media es de 12 puntos.

Tabla 4.21. Resumen de estimación de estadísticos de la Puntuación

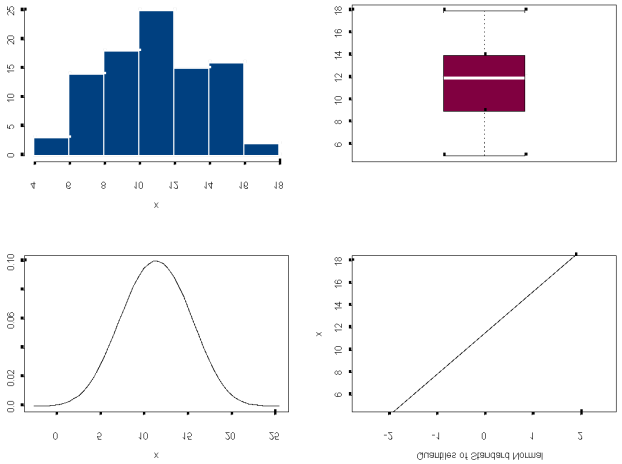
| Minino | 1er Cuartil | Mediana | Media | 3er Cuartil | Máximo | Desv.Estandar |
|---|---|---|---|---|---|---|
| 5 | 9 | 12 | 11,45 | 14 | 18 | 2.92 |

Fig. 4.3. Diagrama de Cajas de las puntuaciones totales



En el diagrama de caja que muestra la Fig. 4.3, se observa casi una perfecta simetría de la distribución de las puntuaciones, lo cual se completa observando que la mediana y la media de la distribución difieren muy poco (Tabla 4.21).

Figura 4.4. Análisis exploratorio de las puntuaciones



Los métodos clásicos de la inferencia estadística dependen en gran medida de ciertos supuestos, tales como: presencia de datos atípicos, ajuste a la distribución

normal, y del auto correlación. El análisis de datos exploratorio (AED) utiliza gráficas para ayudar a obtener una comprensión de si se
se cumplen ó no tales premisas. Por lo tanto, siempre deben realizarse algunas gráficas de análisis de datos exploratorio para responder a preguntas acerca del comportamiento de los datos. Hemos conseguido (Figura 4.4) cuatro muy buenas gráficas que nos proporcionan la forma de la distribución generada a través de la puntuación en el cuestionario, y también detectar la ausencia de valores atípicos y un excelente ajuste a la distribución Normal.

### 4.4.2 Índices de dificultad

En la Tabla 4.22 se presentan los índices de dificultad tomados para la muestra de estudiantes de psicología, junto con los intervalos de confianza correspondientes. Como el índice de dificultad se determina a través de la proporción de alumnos que superan el ítem en la muestra total, se calcularon adicionalmente los intervalos de confianza de dicha proporción utilizando la aproximación normal.

Un índice de dificultad cercano al 1, indica que el ítem ha resultado fácil pues la mayoría de los alumnos lo ha resuelto

Supongamos que con la letra $X$ representemos la cantidad de alumnos que han contestado correctamente el ítem, entonces se tiene que en forma aleatoria, $X$ tomará valores entre 0 y 93. Por ello se puede asumir entonces que $X$ distribuye en forma binomial con parámetros $n$=93, y probabilidad de éxito estimada por $\hat{p} = \frac{X}{n}$ , de donde cada intervalo de confianza se determina utilizando la aproximación normal como dijimos, y ocupando las siguientes fórmulas para los límites inferior y superior:

$$LI = \hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{93}} \qquad LS = \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{93}}$$

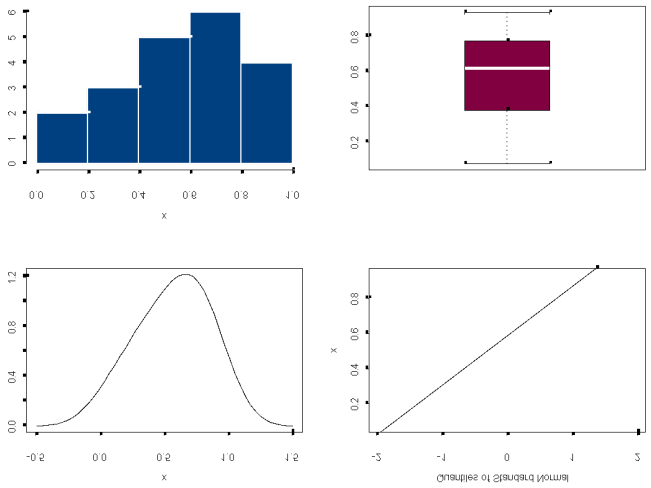
Hemos usado un nivel de confianza del 95%, lo que implica, un valor de $\alpha = 0.05$ . A continuación se muestra la tabla con los valores encontrados.

Tabla 4.22. Índice de Dificultad e Intervalo de Confianza para cada ID

| Ítem | ID | Int. Conf(ID) | Ítem | ID | Int. Conf(ID) |
|---|---|---|---|---|---|
| 1 | 0,924731 | (0,8479;1,0015) | 11 | 0,526882 | (0,3816;0,6722) |
| 2 | 0,774194 | (0,6525;0,8959) | 12 | 0,075269 | (-0,002;0,1521) |
| 3 | 0,107527 | (0,0174;0,1997) | 13 | 0,688172 | (0,5533;0,8230) |

| | | | | | |
|---|---|---|---|---|---|
| 4 | 0,752688 | (0,6271;0,8783) | 14 | 0,258065 | (0,1307;0,3854) |
| 5 | 0,817204 | (0,7047;0,9297) | 15 | 0,516129 | (0,3707;0,6616) |
| 6 | 0,645161 | (0,5059;0,7844) | 16 | 0,688172 | (0,5533;0,8230) |
| 7 | 0,462366 | (0,3172;0,6075) | 17 | 0,849462 | (0,7454;0,9536) |
| 8 | 0,774194 | (0,6525;0,8959) | 18 | 0,408602 | (0,2655;0,5517) |
| 9 | 0,935484 | (0,864;1,0070) | 19 | 0,290323 | (0,1582;0,4224) |
| 10 | 0,591398 | (0,4483;0,7345) | 20 | 0,354839 | (0,2165;0,4941) |

Figura 4.5. Análisis exploratorio del índice de dificultad



Los valores marcados en negrita en la Tabla 4.22, indican aquéllos ítems cuyo nivel de dificultad es demasiado alto ó bajo. Así consideramos dificultad baja aquellos que tengan un ID (Índice de Dificultad) mayor ó igual que 0.9. Aquellos que tienen un ID menor estricto que 0,2, los consideramos con alta dificultad; deberán ser revisadas las causas que han implicado tal dificultad.

En cuanto al análisis exploratorio del índice de dificultad (Figura 4.5), se observa claramente un cierto sesgo positivo de los datos, lo cual correspondería a afirmar que nuestro cuestionario posee una tendencia a ser resuelto con facilidad por la mayoría de los alumnos en este curso, reflejando el aprendizaje de la materia.

Tabla 4.23. Resumen de estimación de estadísticos del Índice de Dificultad

| Minino | 1er Cuartil | Mediana | Media | 3er Cuartil | Máximo | Desv.Estandar |
|---|---|---|---|---|---|---|
| 0.075 | 0.4 | 0.62 | 0.58 | 0.77 | 0.94 | 0.26 |

El 50% de los ítems poseen una dificultad entre los valores 0.4 y 0.77 (1er y 3er Cuartil en la Tabla 4.23) lo que correspondería a que la mitad de los ítem posen una dificultad de moderada a baja.

### 4.4.3 Índice de discriminación

Éste índice refleja la capacidad de ítem de separar (discriminar) los individuos que poseen una característica de los que no la poseen. Si se tratase de una prueba de conocimientos de alguna asignatura, un ítem tiene un alto poder de discriminación si lo aciertan casi todos los que han obtenido muchos puntos en la prueba, y fallan la mayoría de los que han obtenido pocos puntos. Para calcular este índice, se seleccionan dos grupos, en función de sus puntuaciones totales de acuerdo con el detalle que a continuación se da:

- Grupo inferior: Todas las puntuaciones inferiores o iguales al percentil 27 (se selecciona el 27% de las puntuaciones totales, escogiendo las mas bajas posibles).
- Grupo superior: Todas las puntuaciones superiores o iguales al percentil 73(se selecciona el 27% de las puntuaciones totales, escogiendo las mas altas posibles).

Se considera que el ítem "discrimina", si existe diferencia significativa entre la proporción de respuestas en ambos grupos (Tabla 4.24). Mediante la tabla última antes mencionada, pretendemos indicar cual de los ítems discrimina para esta Evaluación. Como ya hemos señalado antes, entendemos que un ítem tiene un alto poder de discriminación si lo aciertan casi todos los que han obtenido muchos puntos en la prueba, y fallan la mayoría de los que han obtenido pocos puntos; es lógico entonces pensar en comparar los valores medios de las puntuaciones por ítem entre el primer tercio con menos puntaje, y el último tercio de alumnos con el más alto puntaje. Las pruebas *t* para diferencias entre estos valores medios para cada ítem se presentan en la Tabla 4.24. Aquí se muestra una secuencia de cuarenta pruebas *t* para diferencias de medias.

Explicaremos por que se trata de cuarenta pruebas *t* y no veinte, si se debe realizar una por cada ítem. Primeramente se deben realizar pruebas *F* de igualdad de varianzas (primeras dos columnas de la Tabla 4.24), pues la prueba *t* ocupa distinto percentil, ya que cambian los grados de libertad, según el resultado de cada prueba *F*.

Por ejemplo, tomemos el ítem 1, la prueba *F* en este caso tiene un nivel de significación con valor 0.000, eso implica que las varianzas son estadísticamente distintas. Finalmente, debido a esta causa, debemos observar la significación de la prueba *t* cuando "no se han asumido varianzas iguales", de donde el *p*-valor de esta

prueba *t* es 0.043 < 0.05 (si decidimos tomar un nivel de significación para el test al 5%). Este resultado del valor *p* implica que se asume estadísticamente diferencia entre las medidas de los puntajes inferiores y superiores. Se decide que este ítem discrimina en la prueba. Asimismo, para el ítem 2, el *p*-valor de la prueba *t* resulta 0.001, de donde se decide que este sí discrimina en la prueba. De este modo, y en forma sucesiva podemos interpretar los resultados de la Tabla 4.24 que se resumen en la Tabla 4.25.

Tabla 4.24. Prueba de diferencias de medias en los dos grupos

| | | Prueba de Levene para la igualdad de varianzas | | Prueba T para la igualdad de medias | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | SIG. | T | GI | SIG. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| | | | | | | | | | Inferior | Superior |
| I1 | v. iguales[1] | 27,903 | 0,000 | 2,138 | 48 | 0,038 | 0,160 | 0,075 | 0,310 | 0,010 |
| | v. diferentes[2] | | | 2,138 | 24,000 | 0,043 | 0,160 | 0,075 | 0,314 | 0,006 |
| I2 | v. iguales | 576,000 | 0,000 | 4,000 | 48 | 0,000 | 0,400 | 0,100 | 0,601 | 0,199 |
| | v. diferentes | | | 4,000 | 24,000 | 0,001 | 0,400 | 0,100 | 0,606 | 0,194 |
| I3 | v. iguales | 15,755 | 0,000 | 1,760 | 48 | 0,085 | 0,160 | 0,091 | 0,343 | 0,023 |
| | v. idiferentes | | | 1,760 | 34,893 | 0,087* | 0,160 | 0,091 | 0,345 | 0,025 |
| I4 | v. iguales | 43,199 | 0,000 | 2,799 | 48 | 0,007 | 0,320 | 0,114 | 0,550 | 0,090 |
| | v. diferentes | | | 2,799 | 37,455 | 0,008 | 0,320 | 0,114 | 0,552 | 0,088 |
| I5 | v. iguales | 92,673 | 0,000 | 03,343 | 48 | 0,002 | 0,360 | 0,108 | 0,577 | 0,143 |
| | v. diferentes | | | 03,343 | 31,488 | 0,002 | 0,360 | 0,108 | 0,580 | 0,140 |
| I6 | v. iguales | 24,564 | 0,000 | 04,000 | 48 | 0,000 | 0,480 | 0,120 | 0,721 | 0,239 |
| | v. diferentes | | | 4,000 | 41,694 | 0,000 | 0,480 | 0,120 | 0,722 | 0,238 |
| I7 | v. iguales | ,340 | ,563 | 2,342 | 48 | 0,023 | 0,320 | 0,137 | 0,595 | 0,045 |
| | v. diferentes | | | 2,342 | 47,961 | 0,023 | 0,320 | 0,137 | 0,595 | 0,045 |
| I8 | v. iguales | 92,673 | ,000 | 3,343 | 48 | 0,002 | 0,360 | 0,108 | 0,577 | 0,143 |
| | v. diferentes | | | 3,343 | 31,488 | 0,002 | 0,360 | 0,108 | 0,580 | 0,140 |
| I9 | v. iguales | 17,551 | 0,000 | 1,809 | 48 | 0,077 | 0,120 | 0,066 | 0,253 | 0,013 |
| | v. diferentes | | | 1,809 | 24,000 | 0,083* | 0,120 | 0,066 | 0,257 | 0,017 |
| I10 | v. iguales | 3,437 | 0,070 | 1,433 | 48 | 0,158* | 0,200 | 0,140 | 0,481 | 0,081 |
| | v. diferentes | | | 1,433 | 47,776 | 0,158 | 0,200 | 0,140 | 0,481 | 0,081 |
| I11 | v. iguales | 9,143 | 0,004 | 3,098 | 48 | 0,003 | 0,400 | 0,129 | 0,660 | 0,140 |
| | v. diferentes | | | 3,098 | 46,154 | 0,003 | 0,400 | 0,129 | 0,660 | 0,140 |
| I12 | v. iguales | 0,000 | 1,000 | 0,000 | 48 | 1,000* | 0,000 | 0,094 | 0,189 | 0,189 |
| | v. diferentes | | | 0,000 | 48,000 | 1,000 | 0,000 | 0,094 | 0,189 | 0,189 |
| I13 | v. iguales | 69,988 | 0,000 | 5,669 | 48 | 0,000 | 0,600 | 0,106 | 0,813 | 0,387 |
| | v. diferentes | | | 5,669 | 31,784 | 0,000 | 0,600 | 0,106 | 0,816 | 0,384 |
| I14 | v. iguales | 56,824 | 0,000 | 3,792 | 48 | 0,000 | 0,440 | 0,116 | 0,673 | 0,207 |
| | v. diferentes | | | 3,792 | 37,022 | 0,001 | 0,440 | 0,116 | 0,675 | 0,205 |
| I15 | v. iguales | 0,000 | 1,000 | 8,102 | 48 | 0,000 | 0,760 | 0,094 | 0,949 | 0,571 |
| | v. diferentes | | | 8,102 | 48,000 | 0,000 | 0,760 | 0,094 | 0,949 | 0,571 |
| I16 | v. iguales | 24,842 | 0,000 | 5,447 | 48 | 0,000 | 0,600 | 0,110 | 0,821 | 0,379 |
| | v. diferentes | | | 5,447 | 38,569 | 0,000 | 0,600 | 0,110 | 0,823 | 0,377 |
| I17 | v. iguales | 161,185 | 0,000 | 3,361 | 48 | 0,002 | 0,320 | 0,095 | 0,511 | 0,129 |
| | v. diferentes | | | 3,361 | 24,000 | 0,003 | 0,320 | 0,095 | 0,517 | 0,123 |
| I18 | v. iguales | 5,133 | 0,028 | 5,842 | 48 | 0,000 | 0,640 | 0,110 | 0,860 | 0,420 |
| | v. diferentes | | | 5,842 | 44,813 | 0,000 | 0,640 | 0,110 | 0,861 | 0,419 |
| I19 | v. iguales | 29,487 | 0,000 | 2,642 | 48 | 0,011 | 0,320 | 0,121 | 0,563 | 0,077 |
| | v. diferentes | | | 2,642 | 41,379 | 0,012 | 0,320 | 0,121 | 0,565 | 0,075 |
| I20 | v. iguales | 29,487 | 0,000 | 3,633 | 48 | 0,001 | 0,440 | 0,121 | 0,683 | 0,197 |
| | v. diferentes | | | 3,633 | 41,379 | 0,001 | 0,440 | 0,121 | 0,685 | 0,195 |

1: Se asumen varianzas iguales; 2 Se asumen varianzas diferentes

Resultó que sólo los ítems 3, 9, 10 y 12 no discriminan en la prueba, por lo cual deberán ser revisados exhaustivamente, tanto la respuesta correcta como los distractores, para que puedan ser incluidos en una prueba definitiva en investigaciones futuras, y así lograr un cuestionario mas completo, válido y fiable.

Tabla 4.25. Discriminación de cada ítem.

| Ítem | p-valor | Discrimina Si/No | Ítem | p-valor | Discrimina Si/No |
|---|---|---|---|---|---|
| 1 | 0.043 | Si | 11 | 0.003 | Si |
| 2 | 0.001 | Si | 12 | 1.000 | No |
| 3 | 0.087 | No | 13 | 0.000 | Si |
| 4 | 0.008 | Si | 14 | 0.001 | Si |
| 5 | 0.002 | Si | 15 | 0.000 | Si |
| 6 | 0.000 | Si | 16 | 0.000 | Si |
| 7 | 0.023 | Si | 17 | 0.003 | Si |
| 8 | 0.002 | Si | 18 | 0.000 | Si |
| 9 | 0.083 | No | 19 | 0.012 | Si |
| 10 | 0.158 | No | 20 | 0.001 | Si |

### 4.4.4 Aproximación a la fiabilidad

En un proceso de medida en educación, el objetivo es realizar inferencias sobre conceptos abstractos a partir de indicadores empíricos, más precisamente, relacionar los significados personales de los alumnos sobre un objeto matemático con sus respuestas a los ítems del cuestionario (Meliá, 2001).

La medida siempre produce un cierto error aleatorio, siendo la fiabilidad la tendencia a la consistencia o precisión del instrumento en la población medida (Meliá, 2001). Se partió para ello de la teoría clásica de los tests donde la fiabilidad se define como correlación entre las puntuaciones verdadera y observada (Martínez Arias, 1995).

Se tomó el coeficiente Alfa de Cronbach (que se reduce al de Kuder- Richardson para ítems dicotómicos), como estimador del coeficiente de fiabilidad. Entre otras propiedades, el coeficiente Alfa de Cronbach: a) refleja el grado en el que covarían los ítems que constituyen el test; b) es el valor medio de todos los que se obtendrían con el método de las dos mitades si se utilizasen todas las combinaciones de ítems; c) es cota inferior de la que se obtendría por el método de la prueba repetida; y d) estima la fiabilidad en el acto (Martínez Arias, 1995). Se calculó este coeficiente mediante el programa correspondiente del paquete estadístico SPSS 15, analizando los estadísticos de cada ítem si se suprime del instrumento y estudiando el efecto sobre el coeficiente al

ir suprimiendo sucesivamente los ítems que presentan peores resultados (Tabla 4.26).

El valor obtenido con el total de la muestra es de 0,630, que corresponde a una correlación entre la puntuación observada y la puntuación verdadera de 0,794. Se considera que el valor es razonable, dado el tamaño de muestra usado para la estimación y teniendo en cuenta que el número de ítems es pequeño y el cuestionario abarca una variedad de objetos matemáticos.

Tabla 4.26 Resultados del análisis de fiabilidad

| | Media de la escala si se elimina el elemento | Varianza de la escala si se elimina el elemento | Correlación elemento-total corregida | Alfa de Cronbach si se elimina el elemento |
|---|---|---|---|---|
| I1 | 11,08 | 9,201 | ,156 | ,625 |
| I2 | 11,23 | 8,546 | ,325 | ,605 |
| I3 | 11,89 | 9,380 | ,024 | ,636 |
| I4 | 11,25 | 9,232 | ,039 | ,640 |
| I5 | 11,18 | 8,847 | ,227 | ,617 |
| I6 | 11,35 | 8,710 | ,204 | ,620 |
| I7 | 11,54 | 9,012 | ,086 | ,637 |
| I8 | 11,23 | 8,916 | ,171 | ,624 |
| I9 | 11,06 | 9,387 | ,049 | ,632 |
| I10 | 11,41 | 9,157 | ,040 | ,643 |
| I11 | 11,47 | 8,513 | ,258 | ,612 |
| I12 | 11,38 | 8,650 | ,221 | ,618 |
| I13 | 11,30 | 8,234 | ,406 | ,592 |
| I14 | 11,74 | 8,520 | ,315 | ,605 |
| I15 | 11,48 | 8,079 | ,417 | ,589 |
| I16 | 11,31 | 8,304 | ,373 | ,597 |
| I17 | 11,15 | 9,064 | ,152 | ,625 |
| I18 | 11,59 | 8,157 | ,397 | ,592 |
| I19 | 11,71 | 8,643 | ,250 | ,614 |
| I20 | 11,65 | 8,666 | ,220 | ,618 |

Un objetivo de la siguiente etapa será completar el cuestionario con nuevos ítems y aumentar la fiabilidad.

## 4.5. CONCLUSIONES

En este capítulo hemos comentado los resultados de nuestro estudio de evaluación que tenia una doble finalidad: a) Realizar una prueba empírica de un conjunto de ítems para seleccionar aquellos que pudieran posteriormente pasar a formar parte de un cuestionario más completo, válido y fiable; b) Dar una información del significado personal logrado por los estudiantes en los objetos

estadísticos elementales utilizados en el diseño experimental. Las principales conclusiones son las siguientes

Respecto a las pruebas empíricas de ítems, ha resultado con buenas propiedades psicometrías (índices adecuados de dificultad y discriminación) todos los ítems excepto el: 1 (excesivamente fácil), 3 (excesivamente difícil y no discrimina), 9 (excesivamente fácil y no discrimina), 10 y 12 (no discriminaron adecuadamente e ítem 12 muy difícil). Los otros 15 ítems muestran buenas propiedades psicometrías, lo que los hace adecuados para un futuro instrumento de evaluación. Aunque en algunos casos, debido al desequilibrio mostrado entre los distractores, será necesario revisar su redacción.

Respecto al significado personal logrado por los estudiantes, nuestro trabajo confirma las siguientes dificultades encontradas en investigaciones previas:

- *Conceptos-definición*: Confusión de hipótesis nula y alternativa (4.3% en ítem 1) confirmando los resultados de Vallecillos (1994). Confundir la definición de error II (18% en ítem 6) confirmando los resultados de Vallecillos (1994). Confundir errores Tipo I y Tipo II (10% en ítem 4), confirmando los resultados de Vallecillos (1994). Confundir $\beta$ y 1-$\beta$ (6% en ítem 4), confirmando los resultados de Vallecillos (1994).
- *Lenguaje:* Un 3,3% (ítem 9) no considera las hipótesis estadísticas complementarias. Aunado a una dificultad conceptual de acuerdo como está planteado este ítem, la respuesta incorrecta también supone una dificultad de lectura de la notación.
- *Proposiciones*: Los estudiantes dan una interpretación bayesiana del intervalo de confianza (38% en ítem 3), confirmando los resultados de Olivo (En prensa). Es decir no ponen correctamente en relación el nivel de confianza con los límites del intervalo, suponiendo estos límites constantes.
- *Procedimientos*: Un 28,3 % de alumnos olvida alguno de los criterios para que un contraste de un resultado significativo en ítem 10, es decir, tiene dificultad con el procedimiento de contraste, confirmando los resultados de Vallecillos (1994).
- *Argumentos:* No es capaz de argumentar con el cálculo de probabilidades el contraste de hipótesis (26%, ítem 10).

Además hemos encontrado las siguientes dificultades no descritas en trabajos previos:

- *Conceptos-definición:* Un 3,3% (ítem 5) confunde el concepto de independencia de muestras; y en el mismo ítem otro 3,3% desconoce los supuestos mínimos de ANOVA de un factor completamente aleatorizado. El 18,4% (ítem 15) confunde supuestos para aplicar el modelo ANOVA de dos factores completamente *aleatorizado de efectos fijos.*
- *Propiedades:* Confundir muestras independientes y relacionadas (3,3% en ítem 5)
- *.Procedimiento:* Cometen errores al calcular la puntuación típica en un test unilateral derecho (34,8% en ítem 7), 63% de los estudiantes no comprenden bien todos los pasos del cálculo en el análisis de la varianza (ítem 12). Un 31,5% (ítem 19) no comprende el procedimiento para comprobar el supuesto de esfericidad en un estudio con una variable independiente con medidas repetidas.
- *Lenguaje:* El 31,5% de estudiantes no asocian un modelo matemático dado por una ecuación con el correspondiente modelo de análisis de varianza descrito verbalmente (ítem 11). Un 31,5% (ítem 19) no es capaz de leer de una tabla producida por un ordenador para analizar pruebas independientes. Un 31,5% (ítem 17) no interpreta estadísticamente el gráfico de interacciones. Un 12,9% no traduce correctamente el enunciado verbal de un problema en el ítem 16.
- *Argumentos:* El 11,9% (ítem 14) no interpreta los resultados de la tabla ANOVA. El 35,9% (ítem 20) confunde los argumentos a tener en cuenta a la hora de rechazar ó no rechazar una hipótesis nula.

# 5. CONCLUSIONES

## 5.1. INTRODUCCIÓN

Tal como lo hemos indicado en la introducción de este Trabajo, hemos llevado a cabo un estudio exploratorio de evaluación de las posibles dificultades encontradas en una muestra de estudiantes de Psicología sobre las ideas elementales en el Diseño de Experimentos. La elección del tema se basa en que esta es una asignatura importante en la formación de especialistas en esta materia, dado el marcado carácter experimental de la disciplina.

Analizamos las respuestas de una muestra de 93 estudiantes de segundo curso de la Licenciatura en Psicología a un cuestionario de opciones múltiples completado al finalizar una de las asignaturas dedicadas parcialmente al Diseño Experimental en el plan de estudios. Las respuestas, recogidas y analizadas en el Capitulo 4, las hemos comparado con las dadas en las investigaciones previas (siempre que hemos contado con ellas) desarrolladas en el Capítulo 2. Nuestro análisis, tal como lo señalamos en la introducción, lo hemos restringido a algunos de los contenidos estadísticos elementales, base sobre la que se construye el resto de las ideas del diseño experimental, y cuya comprensión debe apoyar la aplicación posterior en la vida profesional.

Luego señalaremos las conclusiones respecto de los objetivos planteados en el apartado 1.5 del Capítulo 1. Así analizaremos las aportaciones más importantes de este trabajo.

Finalmente, describiremos algunas posibles líneas para continuar y completar este trabajo, en que creemos hemos iniciado un aporte a la didáctica del diseño experimental en la enseñanza universitaria.

## 5.2. CONCLUSIONES RESPECTO A LOS OBJETIVOS.

Como lo hemos indicado en esta introducción, en el apartado 1.5 del Capítulo 1 hemos enunciado dos objetivos centrales, perseguidos con este trabajo de fin de master, a saber: 1.) *Definición preliminar de la variable "Comprensión de objetos estadísticos básicos del diseño experimental"*, y 2.) *Puesta a prueba de algunos ítems que podrían constituir parte de un posible cuestionario y que evalúen los contenidos incluidos en la variable.*

En cuanto al objetivo 1.) creemos haberlo alcanzado a través de este trabajo. Para ello hemos seguido a Martínez Arias (1995) para iniciar la definición semántica de la variable (fase a)) delimitando los contenidos que se utilizarán en el proceso de elaboración de nuestro instrumento.

En la Tabla 3.2 presentada en el Capítulo 3, realizamos una especificación de contenidos, es decir hemos tratado de definir el significado institucional evaluado en la investigación. Con esta Tabla, obtenemos una primera aproximación a la definición semántica de la variable "comprensión de objetos estadísticos en diseño experimental". Los objetos estadísticos que inicialmente señalamos como elementales para ser contemplados como componentes de nuestra variable han sido: intervalos de confianza, contraste de hipótesis, modelo ANOVA con sus supuestos de independencia, de homogeneidad y normalidad, comprobación de tales supuestos, modelos matemáticos asociados, interpretación de tablas, interacciones, diferenciación de modelos de efectos fijos ó aleatorios; unifactoriales ó bifactoriales y modelos con medidas repetidas. También es cierto y señalamos aquí que la variable necesita refinarse, para conseguir un futuro cuestionario que cumpla con los requisitos de validez respecto al significado institucional implementado en los cursos de estadística en psicología.

En cuanto al objetivo 2), también lo estimamos alcanzado, pues hemos puesto a prueba una serie de ítems de respuesta múltiple, analizando también las respuestas de estudiantes a los mismos, así como también la comprensión de los contenidos señalados para la variable definida en el objetivo 1).

El análisis de los ítems se ha efectuado tanto a priori, como a posteriori; es decir, primero teóricamente desarrollando paso a paso su solución, así como los posibles razonamientos que llevan a elegir cada distractor. Asimismo hemos estudiado las respuestas después de pasar el cuestionario a nuestro grupo de alumnos de la asignatura Análisis de Datos II, en la carrera Psicología. Hemos estudiado además de la puntuación de cada ítem, el número de respuestas correctas de cada estudiante, para obtener una proporción de los elementos de significado adquiridos en relación con los pretendidos: Del estudio de esa puntuación hemos concluido que el 75% de la muestra ha contestado correctamente la mitad de la prueba, y un alumno que resuelve el 60% de la misma puede ser considerado como típico.

Finalmente las medidas psicométricas de dificultad y discriminación nos han

permitido seleccionar 15, de toda nuestra lista de 20 ítems puestos a prueba (75%); ya que cuatro de ellos debes ser revisados por no discriminar, es decir pueden llegar a confundir a los alumnos que saben, y facilitar una respuesta correcta a los que no saben. Se trata de los ítems: 3, 9, 10 y 12; además el 3, 9 y 10 también se encuentran en los extremos del índice de dificultad, el del medio (ítem 9) por demasiado fácil, y los de los extremos por demasiado difíciles. Además el ítem 1 ha resultado demasiado fácil.

Somos concientes que estos 15 ítems además deberán ser revisados para ser incluidos en trabajos futuros, ya que algunos de ellos presentan un cierto desequilibrio a la hora de elegir los distractores, tales como el ítem 6, 7, 11 y 20. Por otro lado, debemos completar las pruebas empíricas de ítems con un juicio de expertos en que personas que conozcan a fondo el tema evalúen independientemente estos ítems para elegir finalmente aquellos que alcancen un consenso respecto a su adecuación para nuestro estudio.

## 5.3. POSIBLES LÍNEAS PARA CONTINUAR EL TRABAJO.

Tanto los resultados como las conclusiones que presentamos en este trabajo, aunque limitados, por razones ya expuestas más arriba; sientan algunas bases como tema de investigación, y así abrir líneas para continuarla. Creemos que entre otras posibilidades, cabe completar el cuestionario con nuevos ítems, ponerlos a prueba y de este modo contribuir a aumentar la fiabilidad del instrumento.

Para poder completar el cuestionario será necesario cumplimentar algunos pasos comenzando por la mejora de la definición semántica de la variable. Para ello deberemos incluir un exhaustivo análisis de contenidos en los textos en uso sobre diseño experimental, en Psicología en las Universidades utilizando nuestro marco teórico como fundamento e identificando los campos de problemas, definiciones, propiedades, procedimientos, lenguaje y argumentos que finalmente se incluyan en la evaluación. Este análisis, nos permitirá mejorar la primera lista de contenidos, presentado con la Tabla 3.2 del Capítulo 3, pues sabemos que esto sólo ha marcado un punto de partida.

Por otro lado, será necesario aumentar el número de ítems puestos a prueba, seleccionando otros ítems que potencialmente puedan incorporarse al instrumento, bien de investigaciones previas o de los textos analizados y adaptarlos. Además de

ítems de opción múltiple se desea incluir algunos problemas de respuestas abiertas que proporcionarían una información más cualitativa y podría completar este trabajo, permitiéndonos anexar posibles conflictos semióticos que pudiesen encontrarse.

Todos estos ítems deberían someterse a pruebas empíricas similares a las descritas en el trabajo, a las que habría que añadir para el total de ítems el juicio de expertos, (Batanero, Díaz, 2005), localizando expertos en el diseño experimental que colaboren en la evaluación de los ítems, tanto nuevos, como los seleccionados en esta memoria, y así finalmente usar esa información para la selección final de los que conformen el cuestionario.

Preparado el cuestionario, sería necesario llevar a cabo pruebas piloto formales para analizar su validez (discriminante, de contenido, de constructo) fiabilidad (consistencia interna, prueba repetida) y generalizabilidad. Una vez que el cuestionario sea satisfactorio, cumpliendo con los requisitos se utilizaría para realizar un estudio de evaluación de las dificultades de los estudiantes.

También podríamos completar la investigación con un análisis de tipo histórico-epistemológico para localizar posibles causas de frecuentes errores y dificultades encontrados en los alumnos. Así como también tratar de ampliar la lista de antecedentes para preparar un estado de la cuestión.

Deseamos esta investigación nos sirva para seguir nuestro trabajo en una futura tesis doctoral, e incentive a otros para investigar en la comprensión de objetos estadísticos en el ámbito universitario.

# REFERENCIAS


Alvarado, H. (2007). *Significados del teorema central del limite en la enseñanza de la estadística en ingeniería.* Tesis Doctoral. Universidad de Granada.

American Psychological Association, American Educational Research Association y National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, 2(1-2), 75-98.

Batanero, C. y Díaz, C. (2005). Análisis del proceso de construcción de un cuestionario sobre la probabilidad condicional. Reflexiones desde el marco de la TSFS. En A. Contreras, L. Ordóñez y C. Batanero (Eds.). *Investigación en Didáctica de las Matemática. I Congreso Internacional sobre Aplicaciones y Desarrollos de la Teoría de las Funciones Semióticas.* (pp. 15-39). Jaén. Universidad de Jaén.

Behar, R. (2001). *Aportaciones para la mejora del proceso de enseñanza-aprendizaje de la estadística.* Tesis doctoral. Universidad Politécnica de Cataluña.

Belia, S., Fidler, F. y Cumming, G. (2005). Researchers misunderstand confidence intervals and standar error bars. *Psychological Methods, 4,* 389-396.

Birnbaum, I. (1982). Interpreting statistical significance. *Teaching Statistics, 4*, 24–27.

Box, G. E. P., Hunter, W. G. y Hunter J .S. (1989). *Estadística para experimentadores.* Ed. Reverté S.A. Barcelona.

Cumming, G., Williams J., y Fidler F. (2004). Replication, and researchers' *understanding of confidence intervals and standard error bars.* Understanding Statistics, 3*, 299-311*.

delMas, R. C., Garfield, J. B. y Chance, B. L. (1999). A model of classroom research in action: developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3). On line: http://www.amstat.org/publications/jse.

delMas, R. C., Garfield, J. B. y Chance, B. L. (2004). Using assessment to study the development of students' reasoning about sampling distributions. Trabajo presentado en el *American Educational Research Association. Annual Meeting*. California.

delMas, R.C., Garfield, J.B., Ooms, A. y Chance, B.L. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58. On line: www.stat.auckland.ac.nz/serj.

Díaz, C. (2004). *Elaboración de un instrumento de evaluación del razonamiento condicional. Un estudio preliminar.* Memoria DEA. Universidad de Granada.

Díaz, C. (2007). *Viabilidad de la enseñanza de la inferencia bayesiana en el análisis de datos en psicología*. Tesis Doctoral. Universidad de Granada.

Díaz, C. y Batanero, C. (2006).¿Cómo puede el método bayesiano contribuir a la investigación en psicología y educación? *Paradigmas*, (27)2, 35-53.

Díaz, C., Batanero, C. y Wilhelmi, M. R. (En prensa). Errores frecuentes en el análisis de datos en educación y Psicología. *Publicaciones.*

Dunn, O. J. y Clark, V.A. (1997). *Applied statistics: Analysis of variance and Regression*. New York: John Wiley.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9*, 83–96.

Ferré, J. y Rius, F. X. (2008). *Introducción al diseño estadístico de experimentos.* Departamento de Química Analítica y Química Orgánica Universitat Rovira i Virgili. Tarragona, On line http://www.quimica.urv.es/quimio/general/dis.pdf

Fidler, F. y Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th International Statistics Institute Session* CD-ROM. Sidney, Australia: International Statistical Institute.

García, R. M (2004). *Inferencia estadística y diseño de experimentos*. Buenos Aires: EUDEBA.

Godino, J. D. (2003). Teoría de las funciones semióticas. Un enfoque ontológico-semiótico de la cognición e instrucción matemática. *Departamento de Didáctica de la Matemática. Universidad de Granada*. On line: URL: http://www.ugr.es/local/jgodino/.

Godino, J. D. y Batanero, C. (1994). Significado institucional y personal de los objetos matemáticos. *Recherches en Didactique des Mathematiques*, *14*(3), 325-355.

Godino, J. D. Batanero, C. y Font, V. (2007). The onto-semiotic approach to research in mathematics education. *ZDM. The International Journal on Mathematics Education, 39* (1-2), 127-135.

Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (1997*). What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Hines, W. W., Montgomery, D. C., Goldsman, D. M. y Borror, C. M. (2006). *Probabilidad y Estadística para Ingeniería* (Trad. G. Nagore). México: CECSA. (Original en inglés, 2003).

Hodgson, T. (1996). The influence of hands-on activities on student´s understanding of selected statistical concepts. En E. Jacubowski, D. Watkins y H. Biske (Eds.), *Proceedings of Eighteenth Annual Meeting of North American Chapter of the International Group for the Psychology of Mathematics* Education (pp. 241- 246). Panamá City, Fl: PME-NA.

León, O. G. y Montero, I. (2002). *Métodos de investigación en psicología y educación.* Madrid: McGraw-Hill.

Martínez Arias, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Meliá, J. L. (2001). *Teoría de la fiabilidad y la validez*. Valencia: Cristóbal Serrano.

Montgomery, D. C. (2002). *Diseño y análisis de experimentos*. Segunda Edición. Versión en español. Universidad Estatal de Arizona. México: LIMUSA WILEY.

Moses, L. E. (1992). The reasoning of statistical inference. In D. C. Hoaglin y D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107-122). Washington, DC: Mathematical Association of America.

Olivo, E. (En prensa). *Significados del intervalo de confianza en la enseñanza de la ingeniería en México.* Tesis Doctoral. Universidad de Granada.

Osterlind, S. J. (1989). *Constructing test items*. Boston: Kluwer.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10*, 159-163.

Rubin, A. y Rosebery, A. S. (1990). Teachers' misunderstandings in statistical reasoning; evidence from a field test of innovative materials. In A. Hawkins (Ed.) *Training teachers to teach Statistics*. (Voorburg, The Netherlands: ISI), 7289.

Schuyten, G. (1991). Statistical thinking in psychology and education. En D. Vere-Jones (Ed.). *Proceeding of the Third International Conference on Teaching Statistics* (pp. 486-490).Voorburg, The Netherlands: International Statistical Institute.

Selvin, H. C. (1970). A critique of tests of significance in survey research. En. D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 94 – 106). Chicago: Aldine.

Skipper, J. K., Guenter, A. L., & Nass, G. (1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social sciences. In D. E.

Morrison & R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 155-160). Chicago: Aldine.

Tauber, L. (2001). *Significado y comprensión de la distribución normal a partir de actividades de análisis de datos.* Tesis Doctoral. Universidad de Sevilla.

Terán, T. (2006). Elements of meaning and its role in the interaction with a computacional program. En A. Rossman y B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics.* CD-ROM. Salvador (Bahia): International Association for Statistical Education.

Vallecillos, A. (1994). *Estudio teórico experimental de errores y concepciones sobre el contraste de hipótesis en estudiantes universitarios*. Tesis Doctoral. Universidad de Granada.

Vallecillos, A. (1996). *Inferencia estadística y enseñanza: Un análisis didáctico del contraste de hipótesis.* Granada: Comares.

Vallecillos, A. , Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticaw y su comprensión por estudiantes universitarios. *Recherches en Didactique des Mathématiques*, 17(1), 29-48.

Well, A. D., Pollastsek, A.y Boyce, S. J. (1990). Understanding the effects of the sample size on the variability of the means. *Organizational Behavior and Human Decision Processes, 47*, 289-312.

Yates, F. (1951). The influence of "Statistical methods for research workers" on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.

# ANEXO

**Publicaciones a partir del Trabajo de fin de Master**

Vera, O., y Díaz, C. (En prensa). Algunas dificultades de estudiantes de psicología en relación al contraste de hipótesis. Trabajo aceptado para presentación en el *I Encuentro de docentes e Investigadores de Estadística en Psicología.* Buenos Aires, Argentina: Facultad de Psicología. Universidad de Buenos Aires.

Vera, O., Olivo, E., Alvarado, H. y Batanero, C. (2007). Estadística y competencias en la formación del ingeniero En M. Molina, P. Pérez-Tyteca y M. A. Fresno (Eds.). *Jornadas de Investigación en el Aula de Matemáticas. Competencias matemáti*cas. Granada. Sociedad Thales y Departamento de Didáctica de las Matemáticas. CD- ROM.

Vera, O., Olivo, E., Díaz, .C. (En prensa). Intervalos de confianza. Interpretación por estudiantes universitarios. Trabajo aceptado para presentación en el *VI Congreso Iberoamericano de Educación Matemática.* Puerto Montt, Chile: Universidad de los Lagos.