Presentaciones de Estadística

Grado en Enfermería

Pedro Femia Marzo

curso 2022/23

Bioestadística – Facultad de Medicina

Universidad de Granada

ESTADÍSTICA

Grado en Enfermería

Contenidos

Actualización curso 2019/20

Pedro Femia Marzo

pfemia@ugr.es

Unidad de Bioestadística – Facultad de Medicina

Universidad de Granada

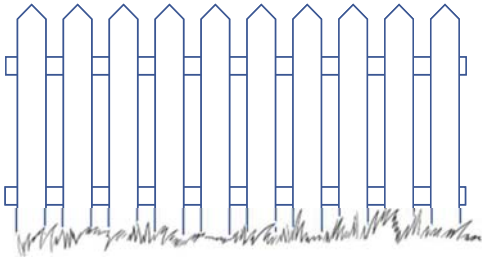
ESTADÍSTICA

Grado en Enfermería

Preámbulo

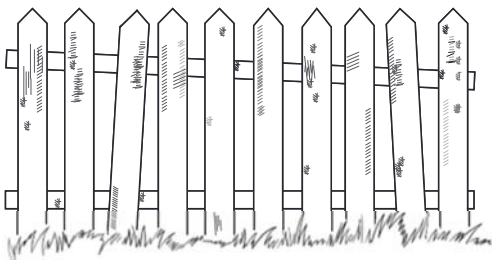
¡Variabilidad!

Pedro Femia Marzo
pfemia@ugr.es
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada
Actualización: 2019



La valla recién hecha:

- Todas las tablas son iguales, no tienen **variabilidad**
- No hace falta decir mucho para describir de forma pormenorizada cómo es la valla



La misma valla después de un periodo de tiempo

- Ya no son todas las tablas iguales, ha aumentado su **variabilidad**
- Ahora, la descripción pormenorizada de la valla requiere aportar mucha más información que antes

THE ZERO-FORCE EVOLUTIONARY LAW (ZFEL)
Brandon & McShea (2010)

Diversidad (variabilidad) ⇨

- Complejidad
- Aparición de niveles jerárquicos

ESTADÍSTICA

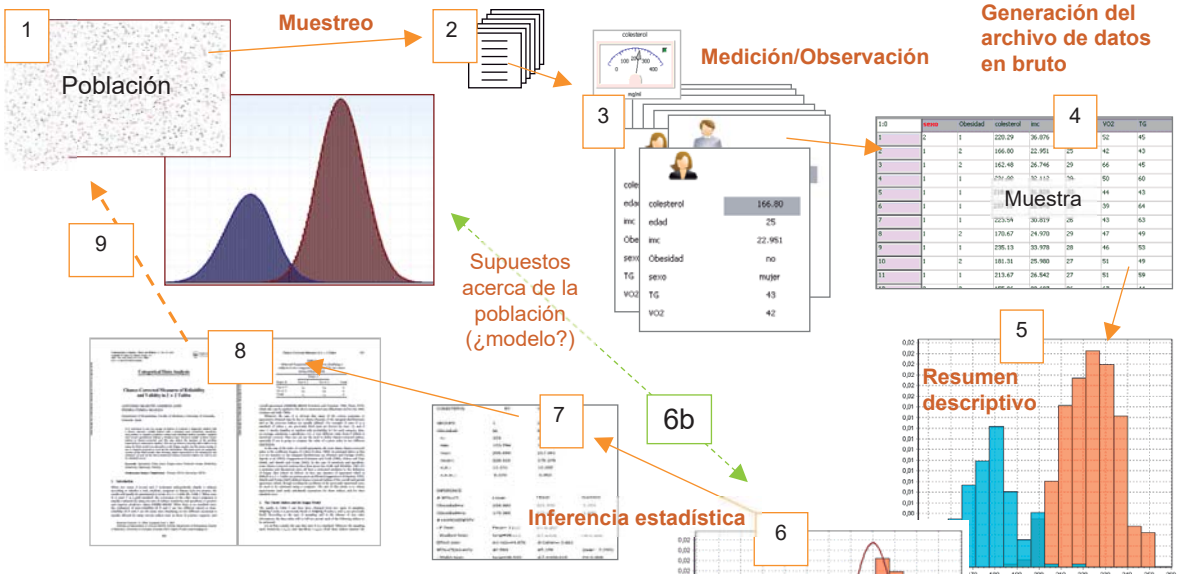
Grado en Enfermería

Tema I

Estadística descriptiva

Pedro Femia Marzo
pfemia@ugr.es
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada
Actualización: 2019

Preámbulo
La investigación empírica



1. Definición de la población objeto de estudio
2. Selección de la muestra → Representatividad de las observaciones
3. Medición → validez y fiabilidad del instrumento de medida
4. Formato admisible por un programa de análisis estadístico
5. Métodos descriptivos
6. Métodos inferenciales. (6b) ¿Son correctos los supuestos sobre la población? ¿es posible establecer un **modelo**?
7. Informe estadístico → evaluación de los resultados.
8. Publicación de resultados → Nueva información sobre la población (9)

¿En cuáles de esta etapas se puede cometer *error*? ¿de qué tipo?

Tabla de datos en bruto (casos × variables) → cómo son los datos que analiza un paquete informático

Id	Sujeto	Sexo	FNacimiento	Edad	Peso	TALLA	Grupo	Rehabilitacion	Nivel_AF	Fe	Na	K	urea	AcUrico	Colesterol	centro
1	Elias R. O.	Masculino	27-Nov-1987	17	63,2	180,0	Grupo control	.	Moderado	.	.	.	.	.	.	Granada
2	Manuel J G. O.	Masculino	14-Jan-1988	17	101,3	175,7	Grupo control	51	Alto	149	140	4,8	40	6,3	145	Málaga
3	Elena D. H.	Femenino	15-Jul-1987	18	52,0	166,7	Tratamiento intenso	71	Moderado	76	141	4,9	19	3,1	177	Málaga
4	J Ignacio F. E.	Masculino	09-Apr-1989	16	68,4	161,1	Tratamiento intenso	47	Alto	114	143	4,4	43	6,3	214	Granada
5	Laura A. R.	Femenino	22-Dec-1986	18	56,1	165,4	Tratamiento suave	53	Alto	53	141	4,3	46	4,8	177	Sevilla
6	Leticia R. U.	Femenino	16-Sep-1989	15	59,8	149,3	Grupo control	75	Moderado	124	137	4,5	29	5,2	211	Sevilla
7	Antonio Raul T. O.	Masculino	12-Jul-1989	16	62,8	164,8	Tratamiento suave	.	Alto	.	.	.	.	.	.	Granada
8	Julio M. A.	Masculino	05-May-1986	19	58,1	169,3	Tratamiento intenso	53	Alto	116	145	5,0	28	5,1	126	Málaga
9	Jose Manuel M. O.	Masculino	22-Aug-1987	18	71,1	174,5	Tratamiento suave	46	Bajo	73	142	4,4	21	4,8	152	Málaga
10	Luis B. E.	Masculino	08-Nov-1984	20	84,9	175,1	Grupo control	45	Bajo	125	143	4,2	29	6,9	161	Sevilla
11	Eva A. R.	Femenino	17-Aug-1987	18	58,7	154,2	Tratamiento suave	50	Moderado	74	142	4,2	37	4,9	159	Málaga
12	Enrique R. E.	Masculino	10-Nov-1985	19	84,6	177,4	Tratamiento intenso	36	Alto	147	136	4,8	24	3,6	150	Málaga
13	Andres J P. R.	Masculino	22-Jan-1989	16	54,4	173,2	Tratamiento intenso	47	Alto	68	139	4,1	30	5,7	172	Granada
14	Manuel J G. O.	Masculino	14-Jan-1988	17	104,3	176,4	Grupo control	48	Moderado	243	142	4,5	36	6,7	143	Málaga
15	Sebastián G. U.	Masculino	14-Sep-1989	15	91,3	170,4	Tratamiento intenso	49	Alto	80	140	4,4	43	5,1	160	Granada
16	Antonio Raul T. O.	Masculino	12-Jul-1989	16	62,8	163,4	Tratamiento suave	24	Alto	69	146	4,2	35	6,1	135	Granada
17	Angel M. A.	Masculino	04-Aug-1988	17	67,2	176,1	Tratamiento suave	.	Bajo	.	.	.	.	.	.	Granada
18	Patricia B. A.	Femenino	16-Jun-1988	17	61	[illegible]	[illegible]	[illegible]	Moderado	103	137	4,7	23	4,1	212	Sevilla
19	Raúl G. A.	Masculino	06-May-1987	18	58,1	175,5	Tratamiento intenso	.	.	.	.	.	.	.	.	Málaga
20	Carmen T. E.	Femenino	17-Jul-1988	17	73,4	167,1	Grupo control	39	Moderado	72	140	4,4	23	2,5	139	Málaga
21	Carmen T. E.	Femenino	17-Jul-1988	17	71,8	166,6	Grupo control	.	Bajo	.	.	.	.	.	.	Málaga
22	Javier P. E.	Masculino	16-Jul-1987	18	76,9	187,2	Grupo control	65	Moderado	81	140	4,4	30	4,4	148	Granada
23	Antonio D. I.	Masculino	13-May-1989	16	54,7	168,4	Tratamiento suave	.	Moderado	.	.	.	.	.	.	Granada
24	Carlos Fª C. I.	Masculino	08-Aug-1988	16	54,9	164,3	Grupo control	49	Bajo	92	139	4,7	19	5,4	183	Granada
25	Antonio Raul T. O.	Masculino	12-Jul-1989	16	62,5	164,7	Tratamiento suave	41	Alto	76	139	4,5	32	7,0	168	Sevilla
26	Miriam M. E.	Femenino	22-Sep-1989	15	52,3	154,9	Tratamiento intenso	53	Moderado	84	141	4,6	33	4,8	150	Sevilla
27	Enrique J A. N.	Masculino	15-Apr-1988	17	59,7	172,2	Tratamiento intenso	39	Moderado	118	143	4,4	21	6,5	122	Granada
28	Angel M. A.	Masculino	04-Aug-1988	17	68,8	176,3	Tratamiento suave	62	Alto	148	141	4,5	30	4,8	139	Granada
29	Pedro G. O.	Masculino	30-Jul-1988	17	66,3	167,3	Tratamiento suave	41	Alto	77	.	.	42	4,8	116	Granada
30	Oscar Manuel L. A.	Masculino	04-Feb-1986	19	60,1	168,0	Tratamiento suave	43	Alto	63	143	4,6	26	5,5	137	Málaga
31	Alba T. O.	Femenino	20-Oct-1989	15	85,2	154,9	Tratamiento intenso	48	Moderado	112	138	4,2	18	5,0	146	Sevilla

caso

Variable

Valor perdido o faltante

Los tipos de datos

Modalidad

Cada una de las maneras en las que se presenta un carácter.

Tipos de Datos (o tipos de variables)

Cualitativos: aquellos que se refieren a una cualidad, no son expresables de manera rigurosa por un número.

1. **Nominales:** las modalidades no son susceptibles de estar ordenadas. Si solo hay dos modalidades se habla de datos *Binarios* o *Dicotómicos*.
   → **Categorías** (sexo, grupo sanguíneo, ser o no seropositivo para VIH,…)
2. **Ordinales:** cuando las modalidades son susceptibles de estar ordenadas
   → **Escalas** (escala de dolor (nada/poco/mucho), escala de satisfacción (baja/media/alta),...)

Cuantitativos: aquellos que necesariamente requieren de un número para ser expresados de manera rigurosa.

3. **Discretos:** aquellos datos que sólo pueden tomar "valores numéricos aislados" (*números enteros*)
   → **Recuentos** (nº de recaídas tras un tratamiento, nº de hijos, nº de afectados,…)
4. **Continuos:** pueden tomar cualquier valor dentro de un intervalo, de modo que entre dos valores cualquiera siempre existe otro valor posible (*números reales*)
   → **Medidas** (peso, nivel de colesterol, presión arterial,…)

Pedro Femia
pfemia@ugr.es

Los tipos de datos

Cualitativos:

1. **Nominales** → Categorías
2. **Ordinales** → Escalas

↑ categorización

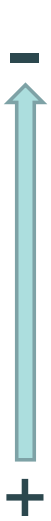
Cuantitativos:

3. **Discretos** → Recuentos

↑ redondeo

4. **Continuos:** → Medidas

Contenido informativo



? ¿Qué tipo de carácter es la edad de un paciente?

Edad > 35	si no
Grupos de Edad	20-30 30-40 40-50
Edad (forma habitual)	22 38 48 ....
Edad (exacta)	22.2 38.5 48.8 ....

Las escalas en Medicina y Ciencias de la Salud

Constructo (RAE: Construcción teórica para comprender un problema determinado)

Idealización de un concepto en el marco de una teoría (habitualmente psicológica). Desde el punto de vista físico, el constructo es una magnitud ***mal medible***, no hay un instrumento físico de medida que permita evaluar su magnitud o intensidad. En su lugar se utilizan ***cuestionarios*** (= ***instrumentos*** en el ámbito psicométrico). Habitualmente se define a través de múltiples aspectos (=variables = ***ítems*** del cuestionario).

Ejemplos de constructos: Inteligencia, personalidad, innovación, ansiedad,…

Por parte del profesional
- Nivel de Sobrecarga
- Nivel de Compromiso
- Nivel de Competencia
- … etc

¿Puede **definir** lo que es la **estatura** de un paciente?
¿Puede explicar cómo **cuantificar** la **estatura** de un paciente?

¿Puede definir lo que es el **nivel de ansiedad** de un paciente?
¿Puede explicar cómo **cuantificar el nivel de ansiedad** de un paciente?

Por parte del paciente
- Nivel de calidad de vida (Cuestionario SF36)
- Nivel de dependencia (Escala de Barthel)
- Capacidad funcional física de las actividades cotidianas (Test de Barthel, Escala de Lawton y Brody)
- Riesgo de deterioro de la integridad cutánea (Escala Norton)
- Nivel de impacto de la fibromialgia (FIQ, *Fibromyalgia Impact Questionnaire*)
- … etc

Cuestionario = Instrumento para *medir constructos*

Suele constar de varias preguntas = ***ítems*** cada uno de los cuales evalúa un aspecto determinado
Con frecuencia, los ítems se construyen en forma de un enunciado y una escala de intensidad (habitualmente con un número impar de opciones) = **escalas de Likert**

Las escalas de Likert son ***sumativas***, de manera que se obtiene una ***puntuación total*** del cuestionario obtenida por la suma de los ítems

El paciente con esquizofrenia debe permanecer aislado

	Codificaciones alternativas		
Niveles de la escala	A	B	C
Total desacuerdo	1	0	-2
Desacuerdo	2	1	-1
Neutral	3	2	0
De acuerdo	4	3	1
Totalmente de acuerdo	5	4	2

¿Da igual la codificación elegida?

Pedro Femia
pfemia@ugr.es

Ejemplo de variables de tipo ordinal (escalas)

Test de Zarit para evaluar el nivel de sobrecarga del cuidador

Ítem	Pregunta a realizar	Puntuación: 0 Nunca	1 Rara vez	2 Algunas veces	3 Bastantes veces	4 Casi siempre
1	¿Siente que su familiar solicita más ayuda de la que realmente necesita?					
2	¿Siente que debido al tiempo que dedica a su familiar ya no dispone de tiempo suficiente para usted?					
3	¿Se siente tenso cuando tiene que cuidar a su familiar y atender además otras responsabilidades?					
4	¿Se siente avergonzado por la conducta de su familiar?					
5	¿Se siente enfadado cuando está cerca de su familiar?					
6	¿Cree que la situación actual afecta de manera negativa a su relación con amigos y otros miembros de su familia?					
7	¿Siente temor por el futuro que le espera a su familiar?					
8	¿Siente que su familiar depende de usted?					
9	¿Se siente agobiado cuando tiene que estar junto a su familiar?					
10	¿Siente que su salud se ha resentido por cuidar a su familiar?					
11	¿Siente que no tiene la vida privada que desearía debido a su familiar?					
12	¿Cree que su vida social se ha visto afectada por tener que cuidar de su familiar?					
13	¿Se siente incómodo para invitar amigos a casa, a causa de su familiar?					
14	¿Cree que su familiar espera que usted le cuide, como si fuera la única persona con la que puede contar?					
15	¿Cree que no dispone de dinero suficiente para cuidar a su familiar además de sus otros gastos?					
16	¿Siente que será incapaz de cuidar a su familiar por mucho más tiempo?					
17	¿Siente que ha perdido el control sobre su vida desde que la enfermedad de su familiar se manifestó?					
18	¿Desearía poder encargar el cuidado de su familiar a otras personas?					
19	¿Se siente inseguro acerca de lo que debe hacer con su familiar?					
20	¿Siente que debería hacer más de lo que hace por su familiar?					
21	¿Cree que podría cuidar de su familiar mejor de lo que lo hace?					
22	En general: ¿Se siente muy sobrecargado por tener que cuidar de su familiar?					

La escala es **sumativa**. Para el diagnóstico suelen asumirse ***puntos de corte*** (=categorización de la variable numérica)

Un total de 22 ítems puntuados en una escala de 0 a 4 → Puntuación total 0-88 puntos →

< 46	No hay sobrecarga
46-56	Sobrecarga moderada
>56	Sobrecarga intensa

En España es más habitual puntuar los ítems de 1 a 5, entonces la escala iría de 22 a 110 puntos →

< 68	No hay sobrecarga
68-78	Sobrecarga moderada
>78	Sobrecarga intensa

Fiabilidad *test-retest* = 0.86

Ejemplo de variables de tipo ordinal (escalas)

Más ejemplos

Escala de Norton – Valoración del riesgo del deterioro de la integridad cutánea.

Consta de 14 ítems. Cada uno se puntúa de 1 (=mayor deterioro) a 4 (=menor deterioro)

Riesgo de formación de úlceras según la puntuación total:

Puntuación	Riesgo
1-11	Muy alto
12-14	Moderado
15-20	Mínimo

Escala de Barthel – Valoración de la capacidad para realizar las actividades básicas de la vida diaria

Consta de 10 ítems. No todos se puntúan igual. Valoración de cada ítem: 0 (=independencia), 5, 10 incluso 15 en algunos ítems (mayor puntuación = mayor independencia) →

Actividad	Puntuación
Comer	0 – 5 – 10
Empleo ducha/baño	0 – 5
Vestirse	0 – 5 – 10
Aseo personal	0 – 5
Control anal	0 – 5 – 10
Control vesical	0 – 5 – 10
Uso retrete	0 – 5 – 10
Trasladarse (sillón/cama)	0 – 5 – 10 – 15
Desplazamiento	0 – 5 – 10 – 15
Subir escaleras	0 – 5 – 10

Puntuación	Nivel de dependencia
0 – 19	Dependencia Total
20 – 39	Dependencia Grave
40 – 59	Dependencia Moderada
60 – 85	Dependencia Leve
86 – 100	Independiente

Construir un instrumento no es inmediato. Debe demostrarse su

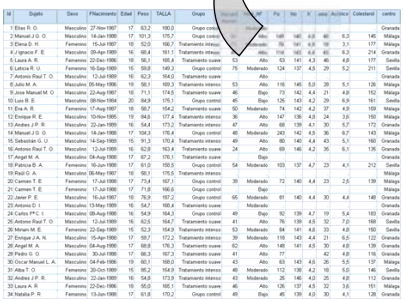
- **Validez**: que el instrumento realmente mide el constructo que se pretende medir
- **Fiabilidad**: si se mide repetidamente a los mismos pacientes ¿se obtiene siempre el mismo resultado?

El **análisis de cuestionarios** no se va a abordar en este curso. Requiere conocimientos un tanto más avanzados (para los que este curso constituye la base). Técnicas habituales son las propias del **Análisis Factorial** (tanto exploratorio como confirmatorio)

Pedro Femia
pfemia@ugr.es

Métodos descriptivos

Datos



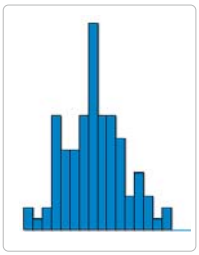
Resumen de la información

3 niveles descriptivos:

- **Tablas de frecuencias**
- **Representaciones gráficas**
- **Medidas descriptivas**

EDAD

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	2.00	6	10.7	10.7
	3.00	11	19.6	30.4
	4.00	9	16.1	46.4
	5.00	13	23.2	69.6
	6.00	7	12.5	82.1
	7.00	6	10.7	92.9
	8.00	4	7.1	100.0
	Total	56	100.0	



Se trata de sintetizar cómo es la ***distribución de la variable***

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
Peso (kg)	190	41,5	106,8	64,777	12,9230

Tablas de frecuencias // variables nominales

Grupo sanguíneo observado en los participantes del estudio (n=500)

	Frecuencia absoluta (f_i)	Frecuencia relativa (h_i)	Porcentaje
A	150	0.30	30.0
B	75	0.15	15.0
AB	25	0.05	5.0
0	250	0.50	50.0
Total	500		100.0

La **frecuencia relativa** (o el %) no es mas que la *estandarización* de la frecuencia absoluta. Este tipo de estandarización es muy conveniente, ya que permitirá hacer comparaciones coherentes, como veremos más adelante

Frecuencia absoluta (*recuento*): $f_1 + f_2 + ... + f_k = \sum_{i=1}^{k} f_i = n$

Frecuencia relativa (*porcentaje*): $h_i = f_i / n; \quad h_1 + h_2 + ... + h_k = \sum_{i=1}^{k} h_i = 1$

Sobre la notación y el símbolo sumatoria

Aquí $k = 4$: $n = \sum_{i=1}^{4} f_i = \underset{f_1}{150} + \underset{f_2}{75} + \underset{f_3}{25} + \underset{f_4}{250} = 500$

Propiedades: 1) $\sum_{i=1}^{k} a x_i = a \sum_{i=1}^{k} x_i$; 2) $\sum_{i=1}^{k} (a + x_i) = ka + \sum_{i=1}^{k} x_i$

Pedro Femia
pfemio@ugr.es

Tablas de frecuencias // variables ordinales

Efecto observado tras la aplicación de una terapia acuática en n=500 mujeres con fibromialgia

Estado	Frecuencia	Frecuencia acumulada	Porcentaje	Porcentaje acumulado
Mucho peor	25	25	5.0	5.0
Peor	98	25+98= 123	19.6	5+19.6= 24.6
Igual	159	123+159= 282	31.8	24.6+31.8= 56.4
Mejor	147	282+147= 429	29.4	85.8
Mucho mejor	71	429+71= 500	14.2	100
Total	500		100	

La posibilidad de ordenar de forma unívoca a las categorías de la variable permite poder definir un nuevo tipo de frecuencia: la **frecuencia acumulada** (que podrá ser relativa o absoluta)

La frecuencia acumulada es relativa al orden natural de las categorías

Tablas de frecuencias // variables discretas

Distribución de la edad (años) n=124

	Frec	%	% acum.
15	5	4.0	4.0
16	28	22.6	26.6
17	64	51.6	78.2
18	18	14.5	92.7
19	6	4.8	97.6
20	3	2.4	100.0
Total	124	100.0	

Con pocos valores distintos (k=6)

Distribución de la edad (años) n=124

	Frec	%	% acum.
menos de 21	5	4.0	4.0
21-25	28	22.6	26.6
26-30	64	51.6	78.2
31-35	18	14.5	92.7
36-40	6	4.8	97.6
más de 40	3	2.4	100.0
Total	124	100.0	

Con muchos valores distintos (k grande)

Agrupación en intervalos desde *el punto de vista estadístico*

Si la variable discreta puede tomar muchos valores distintos, su resumen como tabla de frecuencias suele hacerse agrupándola en intervalos.
Los intervalos deben ser:

- **Homogéneos** (todos con la misma amplitud)
- **Exhaustivos** (recorren todos los valores de la variable)
- **Excluyentes** (no pueden solaparse)

Pedro Femia
pfemia@ugr.es

Tablas de frecuencias // variables continuas

Con variables continuas, no queda más remedio que agrupar en intervalos

Nivel de colesterol (mg/dL) observado n=157

	Frec	%	% acum.
102 - 128	20	12.7	12.7
129 - 155	69	43.9	56.7
156 - 182	85	28.7	85.4
183 - 209	13	8.3	93.6
210 - 236	10	6.4	100.0
Total	157	100.0	

No se han observado (se han redondeado) los valores decimales, por tanto no hace falta solapar los intervalos

Nivel de colesterol (mg/dL) observado n=157

	Frec	%	% acum.
(101, 128]	20	12.7	12.7
(128, 155]	69	43.9	56.7
(155, 182]	45	28.7	85.4
(182, 209]	13	8.3	93.6
(209, 236]	10	6.4	100.0
Total	157	100.0	

Si que se han observado los valores decimales, es preciso establecer un criterio para los límites de los intervalos

$(209, 236]$ indica: $209 < x \leq 236$

Tipos de datos

Diagramas de frecuencias ← Qué diagrama es correcto depende del *tipo* de variable

Datos cualitativos:

1. **Nominales**
2. **Ordinales**

Cuantitativos

3. **Discretos**
4. **Continuos**

$k \downarrow$

$k \uparrow$

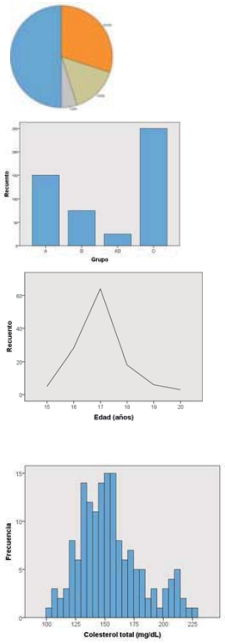


Diagrama de sectores

Diagrama de barras

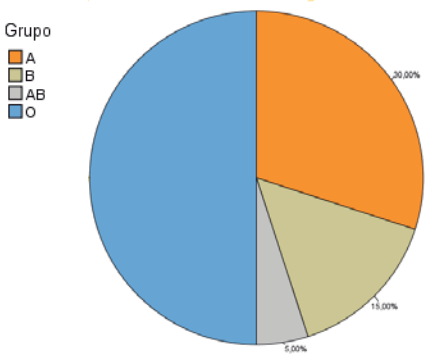
Polígono de frecuencias

Histograma

Pedro Femia
pfemia@ugr.es

Diagramas de frecuencias // variables cualitativas

Distribución de los grupos sanguíneos (n=500)



$$angulo_i^0 = 360^0 \times \frac{f_i}{n}$$

Diagrama de sectores

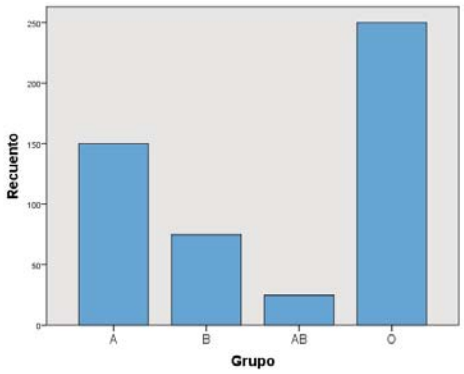
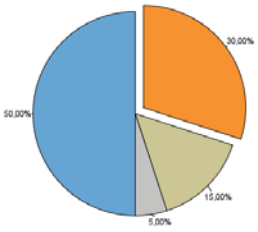
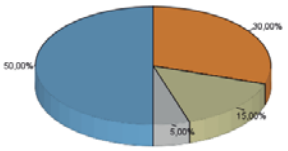


Diagrama de barras



Desgajado de un sector



¡menos adecuado!
(falsa tridimensionalidad)

Principio básico: para cada categoría de la variable, el área del objeto gráfico debe ser proporcional (o igual) a la frecuencia de dicha categoría.

Diagramas de frecuencias // variables ordinales y discretas

Distribución de la edad (n=124)

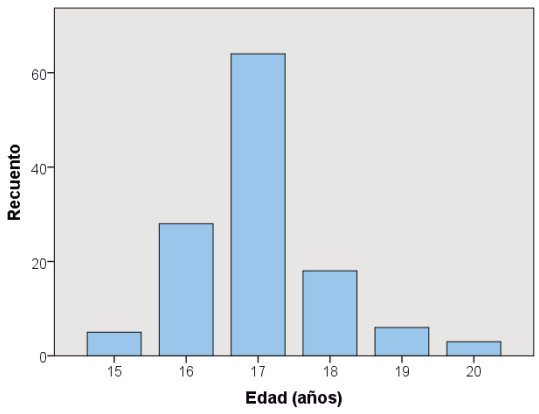
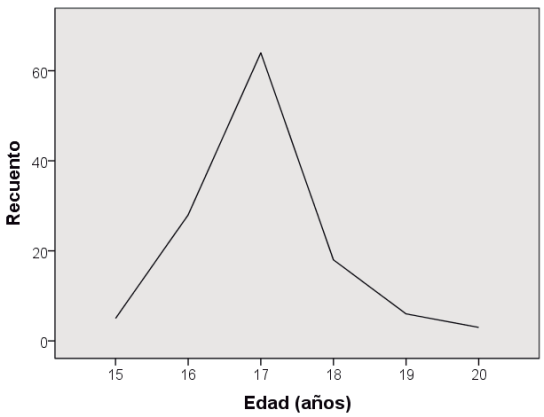


Diagrama de barras

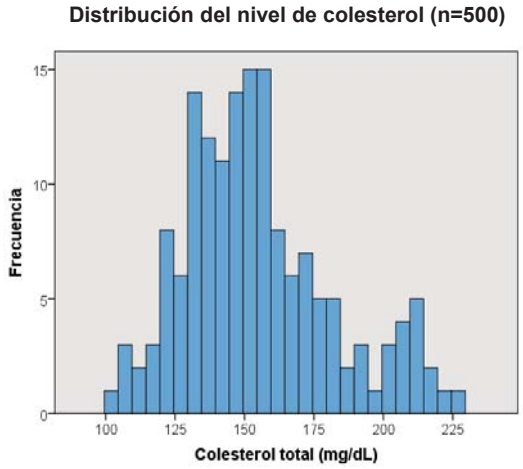
Distribución de la edad (n=124)

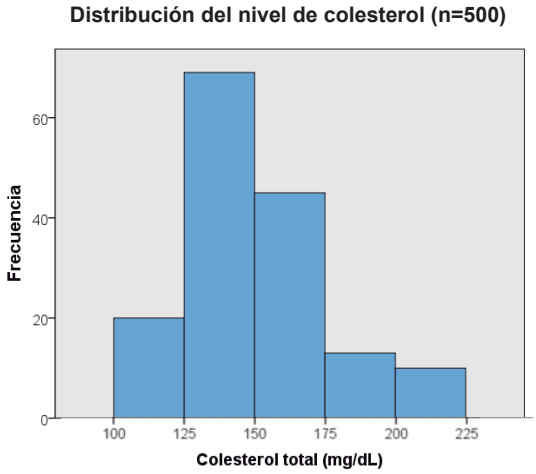


Polígono de frecuencias

Pedro Femia
pfemia@ugr.es

D. frecuencias // variables agrupadas en intervalos (discretas o continuas)

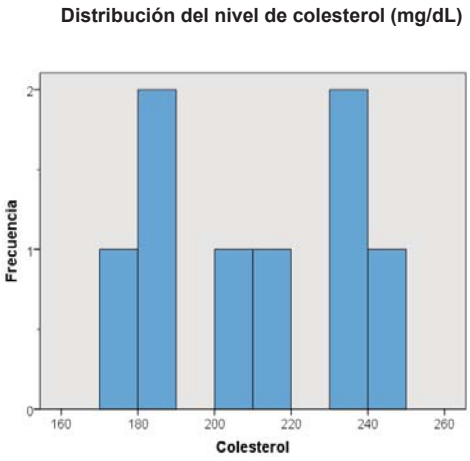


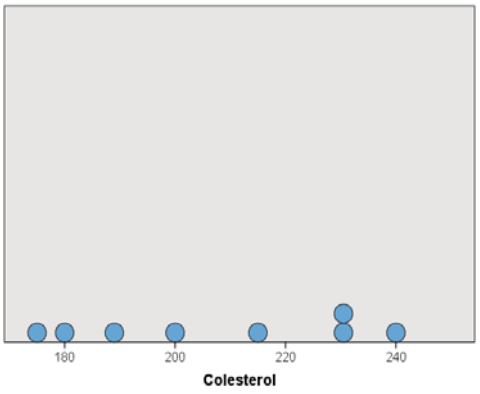


Histogramas

Las dos representaciones corresponden a los mismos datos. Pocos intervalos y muy anchos (dcha) pueden enmascarar la distribución real de la variable

Diagramas de frecuencias // muestras pequeñas





Muestras pequeñas de variables continuas

Cuando hay pocos datos, los histogramas (izda) carecen de sentido. En su lugar es mejor hacer un **diagrama de puntos** (dcha). Tales diagramas tienen mayor interés si se usan con fines comparativos.

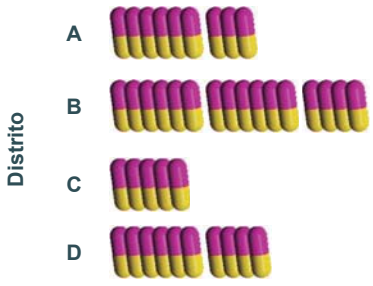
Pedro Femia
pfemia@ugr.es

Diagramas de frecuencias

Otros diagramas

Pictogramas de repetición

Consumo de medicamentos en los cuatro distritos de una ciudad



Pictogramas de amplificación

Enfermedades más comunes en España
(los datos son ficticios)

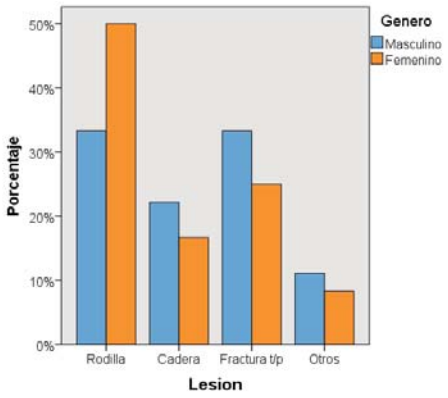
La ***nube de palabras*** es un tipo de pictograma de amplificación muy utilizado en el ámbito de internet

Principios básicos de los diagramas de frecuencias:

- El área del motivo gráfico asignado a cada categoría/valor/intervalo debe ser proporcional a su frecuencia
- Siempre deben tener un encabezado en donde aparezcan *datos básicos* como son el tamaño muestral, las unidades de medida si las hay, el año de referencia,…

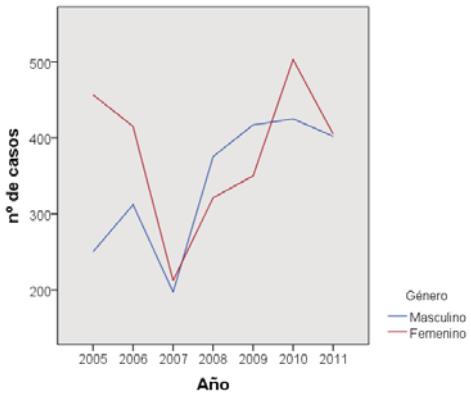
Diagramas de frecuencias // Gráficas comparativas

Barras solapadas

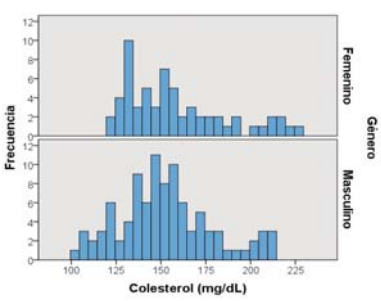


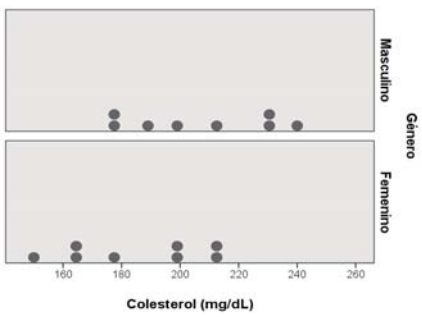
! Cuidado: En el eje de ordenadas debe aparecer la frecuencia relativa (%)

Polígonos de frecuencias



Uso de paneles



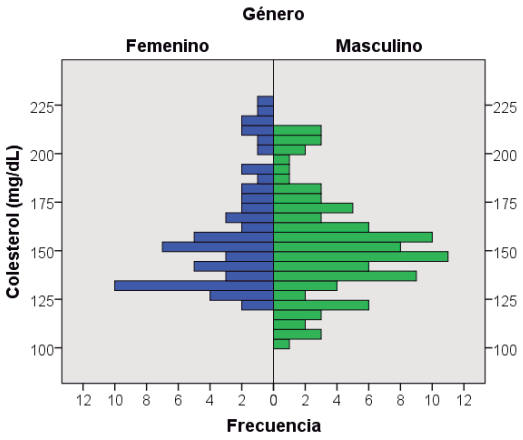


Pedro Femia
pfemia@ugr.es

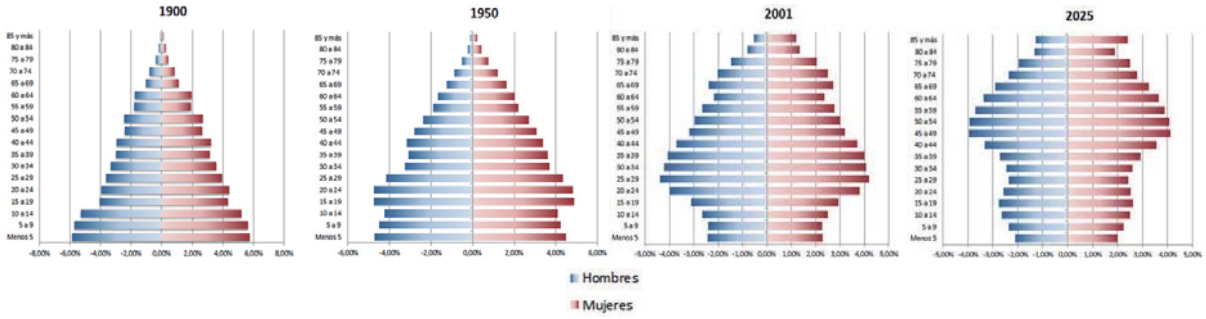
Diagramas de frecuencias // Gráficas comparativas

Pirámide de población

Muy utilizada en **demografía**. Se trata de un doble histograma en donde se representa la frecuencia de individuos en función de grupos de edad y sexo. No obstante, pueden representarse otras variables con fines comparativos (dcha)



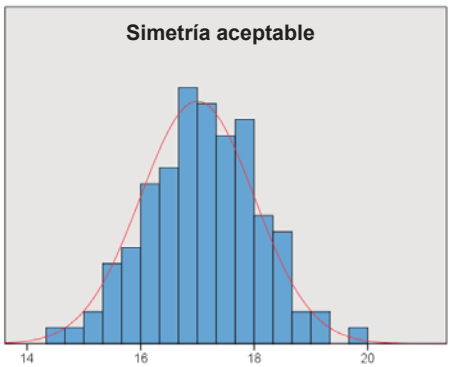
Evolución de la pirámide de población española con pronóstico para el año 2025 (fuente: INE)

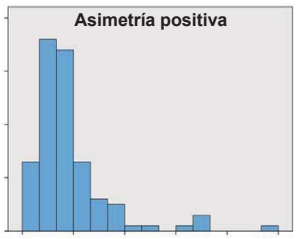


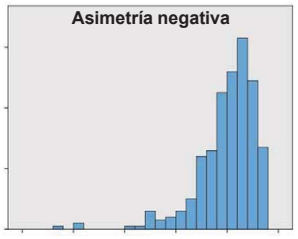
Diagramas de frecuencias // aspectos de las gráficas

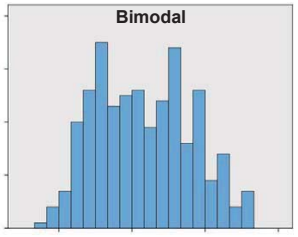
En qué hay que fijarse (histogramas):

- ***Simetría*** de la distribución
- Presencia de ***valores extremos o atípicos***
- Posible ***mezcla*** de grupos (*bimodalidad*)
- ***Equilibrio*** en las gráficas comparativas

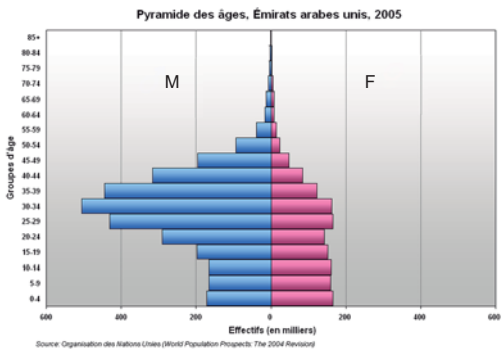








Analice esta imagen

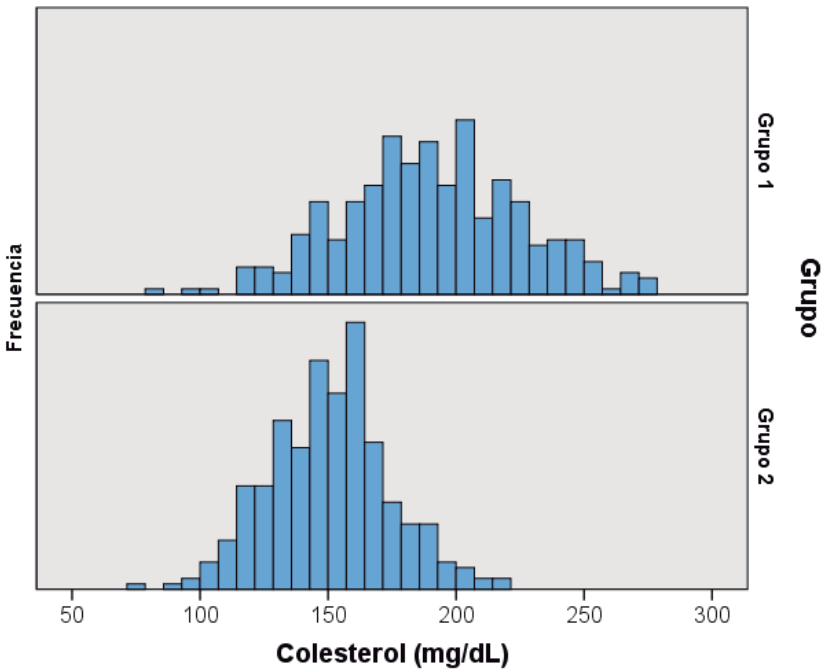


Pirámide de población de los Emiratos Árabes en 2005. (en ese año hubo mucha inmigración procedente, sobre todo, de Pakistán)

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas

¿En qué se diferencian estas dos distribuciones?



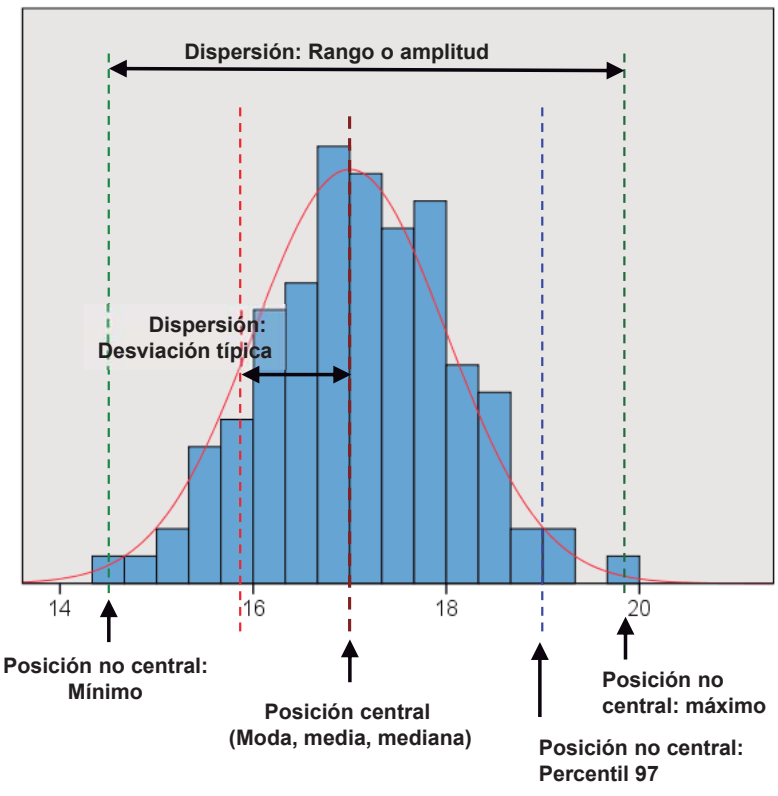
Síntesis de datos: medidas descriptivas

Medidas descriptivas

- De **posición,** con tendencia
  - **Central**: media, mediana, moda
  - **No central**: extremos, percentiles
- De **dispersión**: Rango, varianza, desviación típica, coeficiente de variación, rango intercuartílico

Dispersión = **Variabilidad**

Normalmente se **resumen** los datos proporcionando una medida de **posición central** junto a una de **dispersión**. Adicionalmente puede darse información relativa a otras medidas, como extremos y percentiles



Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas

Medidas descriptivas ← Qué medida se puede calcular depende del tipo de la variable

Datos cualitativos:

1. **Nominales** → Proporción (porcentajes), moda
2. **Ordinales** → Las anteriores y además medidas basadas en el orden de los datos: Mediana y Percentiles, Rango intercuartílico

Cuantitativos

3. **Discretos**
4. **Continuos**

→ Todas las anteriores y además
Medidas basadas en el valor numérico de los datos:
Media y varianza (desviación típica)

Síntesis de datos: medidas descriptivas de posición

Nominales

Proporción (de cada categoría): es la frecuencia relativa de dicha categoría: $p_i = \frac{f_i}{n}$
Cuando se informa de ella se suele dar como porcentaje: $(p_i \times 100)\%$

Moda: es el valor de la variable que tiene mayor frecuencia (puede no ser única).

Ordinales A las medidas descritas hay que añadir

Mediana: es el valor de la variable que divide a la muestra ordenada en dos partes iguales (es decir, deja tanto por debajo como por encima el 50% de las observaciones).

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

Ejemplos:

$a)\ \{3, 6, 7, 10, 15\} \rightarrow Me = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 7$

$b)\ \{3,6,10,15\} \rightarrow Me = x_{\left(\frac{4+1}{2}\right)} = x_{(2.5)} = \frac{6+10}{2} = 8$

$c)\ \{1,5,1,2,2,4,7,2,10\} \rightarrow Me = x_{(5)} = 2$

$d)\ \{peor, peor, peor, mejor, igual, mejor, peor\} \rightarrow Me = x_{\left(\frac{7+1}{2}\right)} = peor$

No confundir la *posición mediana* (n+1)/2, con la *mediana* en si, que es el valor que *ocupa* la posición mediana cuando los datos están ordenados de menor a mayor

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas de posición

Más ejemplos:

Edad (años)
n=124

	f_i	F_i	% acum.
15	5	5	4.0
16	28	33	26.6
17	64	97	78.2
18	18	115	92.7
19	6	121	97.6
20	3	124	100.0
Total	124		

$$Me = x_{\left(\frac{124+1}{2}\right)} = x_{(62.5)} = 17$$

La mediana es el primer valor de la variable cuya frecuencia acumulada es mayor o igual a 62.5 o, equivalentemente, cuyo porcentaje acumulado es mayor o igual al 50%

Edad (años)
n=124

	f_i	F_i	% acum.
menos de 21	5	5	4.0
21-25	28	33	26.6
26-30	64	97	78.2
31-35	18	115	92.7
36-40	6	121	97.6
más de 40	3	124	100.0
Total	124		

$$Me = x_{\left(\frac{124+1}{2}\right)} = 26 \le x_{(62.5)} \le 30$$

Cuando la variable está agrupada en intervalos, la mediana se puede aproximar por ***interpolación lineal*** (se ve en prácticas)

Síntesis de datos: medidas descriptivas de posición

Cuantitativas (Además de las medidas anteriores, al manejar números, ahora podemos hacer cálculos)

Media aritmética: dada una muestra de n observaciones numéricas $x_1, \ldots, x_n$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \cdots + x_n}{n};$$

si los datos están agrupados en una tabla de frecuencias: $\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{f_1 x_1 + \cdots + f_k x_k}{n}$

Ejemplos: **a)** $\{3, 6, 7, 10, 15\} \rightarrow \overline{x} = (3+6+7+10+15)/5 = 8.2$

b)

Edad (años)
n=124

x_i	f_i	$(x_i\, f_i)$
15	5	15×5=75
16	28	448
17	64	1088
18	18	324
19	6	114
20	3	60
Total	n=124	$\sum (x_i\, f_i)$ =2109

$\overline{x} = 2109 / 124 = 17.0$ años

c)

Edad (años)
n=124

	f_i	Marca de clase x_i	$(x_i\, f_i)$
menos de 21	5	18*	5×18=90
21-25	28	(21+25)/2=23	644
26-30	64	(26+30)/2=28	1792
31-35	18	33	594
36-40	6	38	228
más de 40	3	43*	129
Total	124		3477

$\overline{x} = 3477 / 124 = 28.0$ años

* En estos intervalos "mal definidos" se usa la misma regla que con los restantes de la tabla

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas de posición

Media ponderada: Debe utilizarse cuando no todas las observaciones tienen la misma *importancia* (= peso o ponderación: w_i)

Dados

x	x_1	…	x_n
w	w_1	…	w_n

$$\rightarrow \overline{x}_p = \frac{x_1 w_1 + \cdots + x_n w_n}{w_1 + \cdots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

Ejemplo:

Nota obtenida (x)	5	7	9
Créditos de la asignatura (w)	3	3	6

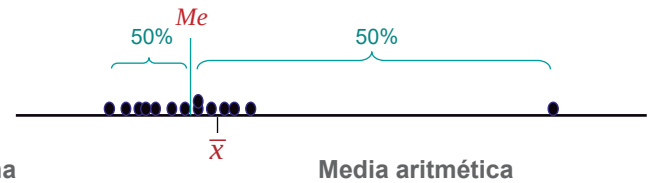
$$\rightarrow \overline{x}_p = \frac{5\times 3 + 7\times 3 + 9\times 6}{3+3+6} = \frac{90}{12} = 7.5 \quad \left(\neq \overline{x} = \frac{5+7+9}{3} = 7 \right)$$

Observe la similitud entre la **media aritmética** para datos agrupados y la **media ponderada**. Cuando la ponderación (w_i) de cada valor es su frecuencia de aparición (f_i), entonces tenemos la misma expresión

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \xrightarrow[\overline{x} \rightarrow \overline{x}_p]{f_i \rightarrow w_i} \overline{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

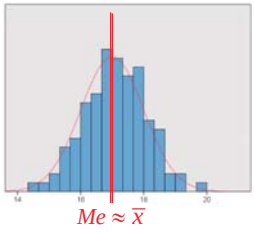
Síntesis de datos: comparativa media y mediana y simetría de la distribución

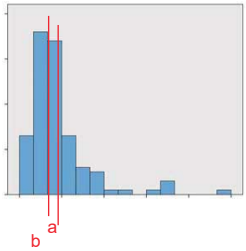
Mediana vs. Media aritmética

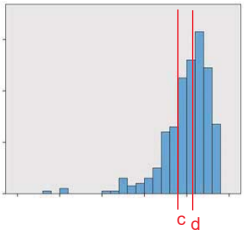


	Mediana	Media aritmética
Métrica de la variable	Se puede calcular con variables ordinales en adelante	Requiere que los datos sean cuantitativos
Criterio de centralidad	El *número de casos* con valores por debajo de la Me es el mismo que el número de casos con valores por encima de ella	La suma de las *distancias a la media* de los valores inferiores a $\overline{x}$ se compensa con la suma de distancias a la media de los valores por encima de ella: $-\sum_{x_i<\overline{x}} (x_i - \overline{x}) = \sum_{x_i>\overline{x}} (x_i - \overline{x})$
Robustez	Es ***robusta*** frente a valores extremos	Más ***sensible*** a los valores extremos, especialmente cuando el tamaño de muestra es pequeño (gana robustez al aumentar n)
Uso	Es menos frecuente como medida de posición en *inferencia*	Es la medida de posición más habitual en *inferencia*

Si la distribución es **simétrica**, $Me \approx \overline{x}$ lo cuál permite decir que en esta situación y de forma aproximada, la media viene a dividir a la muestra en dos partes iguales







¿Quiénes son cada caso la media y la mediana?

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas de posición no central

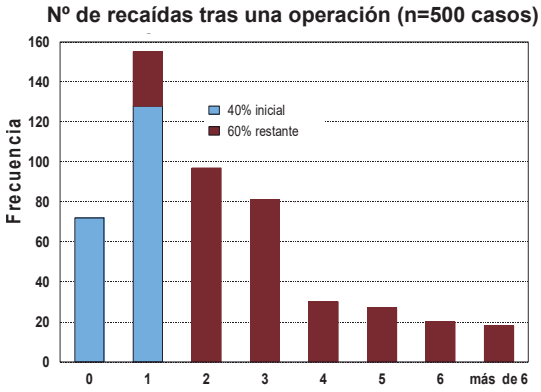
Medidas de posición no central para variables ordinales o cuantitativas

Extremos (mínimo y máximo): Pueden estar muy alejados del resto de los datos en cuyo caso deben estudiarse. Para caracterizar la distribución se usa *otro criterio de extremo*: los ***percentiles***

Percentiles: el percentil α, P_{α}, es el valor de la variable que divide a la muestra ordenada en dos partes, dejando por debajo el α% de las observaciones y por encima el $(1-\alpha)$ %. (generalización del concepto de mediana). Se habla de percentil 1 (P_1), ..., percentil 99 (P_{99}).

$$P_{\alpha} = x_{\left((n+1)\times\frac{\alpha}{100}\right)}$$

Ejemplo: Cálculo del P_{40} a partir de una tabla de frecuencias



x	fi	Fi	acum
0	72	72	14.4%
1	155	227	45.4%
2	97	324	64.8%
3	81	405	81.0%
4	30	435	87.0%
5	27	462	92.4%
6	20	482	96.4%
más de 6	18	500	100.0%
	500		

$$\text{Posición del } P_{40}\,(n=500): (500+1)\times 0.4 = 200.4 \longrightarrow P_{40} = x_{(200.4)} = 1$$

El 40% de los casos observados tienen una recaida o ninguna. El 60% restante tienen una recaida o más

Síntesis de datos: medidas descriptivas de posición no central

Observaciones respecto a los percentiles:

- Casos particulares: hay percentiles que reciben también otro nombre:
  - *Cuartiles*: $Q_1=P_{25}$, $Q_2=P_{50}=Me$, $Q_3=P_{75}$
  - *Deciles*: $D_1=P_{10}$, ..., $D_9=P_{90}$
- Obsérvese que si $\alpha_1 < \alpha_2$ entonces

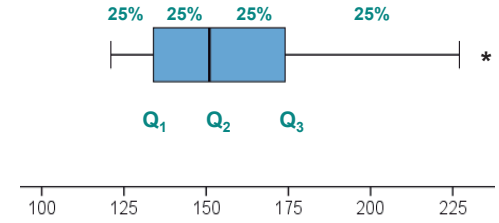
  $$P_{\alpha_1} \le P_{\alpha_2}$$

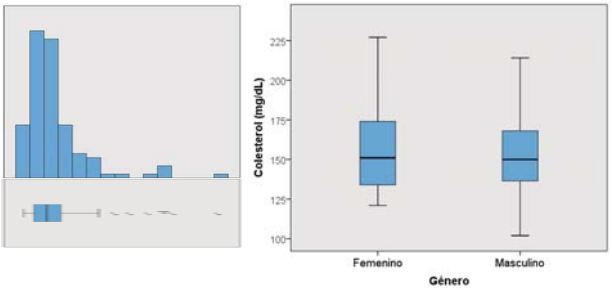
  Es decir, el P_{20} por ejemplo, no puede ser mayor que el P_{25} (aunque si que podría ser igual)
- Si la muestra es muy pequeña, no tiene sentido hablar de percentiles respecto a ella.
- Además de los cuartiles, se utilizan mucho los percentiles 5 y 95 ¿cómo se interpreta cada uno de ellos?

? ¿Encuentra alguna relación entre los percentiles y la frecuencia relativa acumulada?

Diagramas de caja

Un diagrama de caja es una interesante representación de la distribución de la variable que está basada en los cuartiles





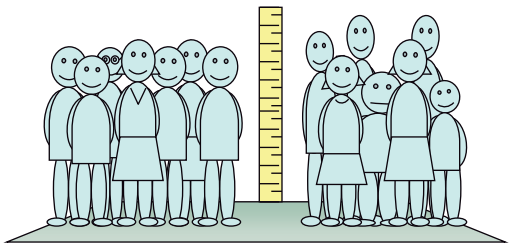
Relación con el histograma

Diagramas con fines comparativos

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas de dispersión

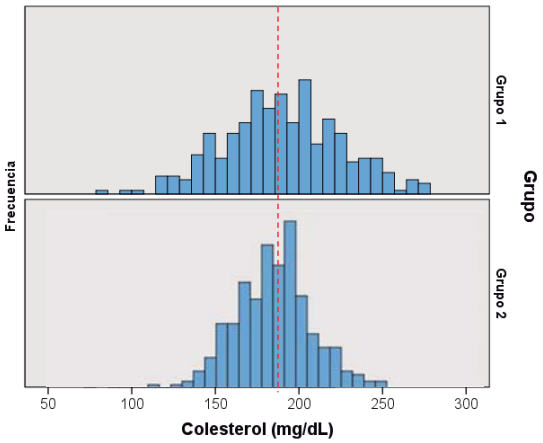
x = Estatura (cm)



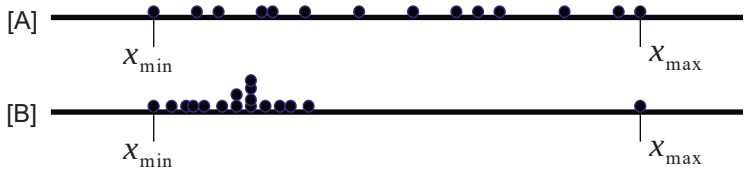
Grupo I $\quad \bar{x} = 170$ cm

Grupo II $\quad \bar{x} = 170$ cm

¿Son iguales en estatura los dos grupos?



¿Son "iguales" estas dos distribuciones con la misma media?



¿Qué se puede decir de estas dos distribuciones?

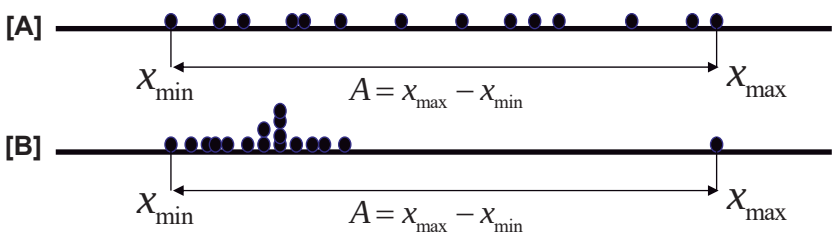
Síntesis de datos: medidas descriptivas de dispersión

Para caracterizar un conjunto de datos, no basta con dar solamente medidas de posición, también es necesario conocer su ***dispersión*** = ***heterogeneidad*** = ***variabilidad***

Medidas de dispersión (solo variables cuantitativas)

Rango (R) o amplitud (A): $\quad A = x_{max} - x_{min}$

Es una medida fácil de calcular, pero muy pobre, ya que solo tiene en cuenta a dos observaciones de toda la muestra (los extremos)



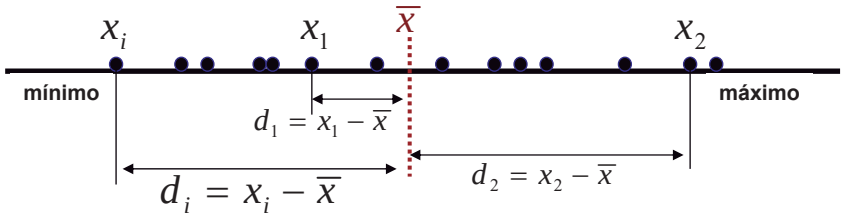
A y B tienen el mismo rango o amplitud

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas de dispersión

¿Cómo se puede definir una buena medida de dispersión que tenga en cuenta a todas las observaciones?

- Es necesario utilizar el concepto de **distancia** de cada observación respecto a un valor de referencia
- La mejor referencia es una medida de posición central: optamos por la media aritmética como referencia
- La idea es calcular todas las distancias de cada observación a la media y hacer la media de dichas distancias:



Ahora la media de las distancias (o distancia media respecto a la media aritmética) será

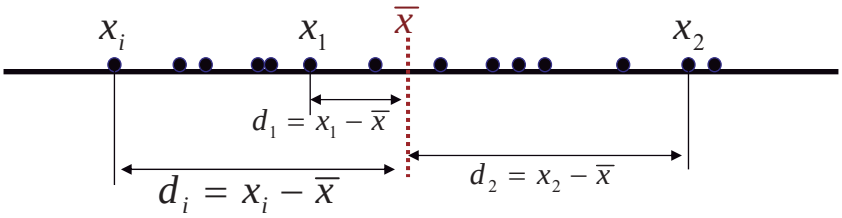
$$\overline{d} = \frac{\sum d_i}{n} = \frac{\sum (x_i - \overline{x})}{n} = \frac{\sum x_i}{n} - \frac{\sum \overline{x}}{n} = \overline{x} - \frac{n\overline{x}}{n} = 0$$

¡¡Siempre cero!!

El problema es que debido al carácter central de la media aritmética, las distancias negativas se compensan siempre con las positivas. En la imagen d_1<0 y d_2>0.
De hecho, el problema es permitir que haya ¡¡distancias negativas!! (eso no tiene sentido)

Síntesis de datos: medidas descriptivas de dispersión

La estrategia que se usa para impedir que las distancias sean negativas es una muy utilizada en las ciencias en general. Se trata de elevar cada distancia al cuadrado. Esto mantiene la información sobre su magnitud pero evita el problema del signo negativo.



El promedio (de las distancias al cuadrado respecto a la media) entonces es:

$$\frac{\sum d_i^2}{n} = \frac{\sum (x_i - \overline{x})^2}{n} = S^2$$

Por cuestiones que veremos mas adelante, se considera un promedio no respecto al tamaño de muestra n, sino respecto a una cantidad que se denomina grados de libertad y que es n-1.
Se define así la

Varianza muestral

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

Pedro Femia
pfemia@ugr.es

Síntesis de datos: medidas descriptivas de dispersión

Cálculo práctico de la varianza

Si los datos no están agrupados en una tabla de frecuencias

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

Si los datos están agrupados en una tabla de frecuencias

$$s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{n} f_i\, x_i^2 - \frac{\left(\sum_{i=1}^{n} f_i\, x_i\right)^2}{n}\right)$$

Ejemplos

a) $x_i : \{4,6,8,10\} \rightarrow \overline{x} = 7; s^2 = \frac{1}{3}\left(\left(4^2+6^2+8^2+10^2\right) - \frac{(4+6+8+10)^2}{4}\right) = 6.667$

b)

x_i	f_i	$x_i f_i$	$x^2_i f_i$
15	5	75	1125
16	28	448	7168
17	64	1088	18496
18	18	324	5832
19	6	114	2166
20	3	60	1200
Suma	124	2109	35987

$$\overline{x} = \frac{2109}{124} = 17.0$$

$$s^2 = \frac{1}{124-1}\left(35987 - \frac{2109^2}{124}\right) = 0.941$$

Obsérvese que la varianza **depende de la media.**

Síntesis de datos: medidas descriptivas de dispersión

La varianza tiene unidades de medida que son aquellas que tenga la variable pero al cuadrado.
Esto supone que si la variable es por ejemplo la estatura en cm (unidad de longitud) su varianza viene dada en cm^2 (¡¡unidad de superficie!!)
Para evitar este problema de dimensionalidad se define la

Desviación típica (o estándar)

$$s = \sqrt{s^2}$$

Esto debe entenderse en el sentido de que la d.t. se obtiene una vez calculada la varianza al considerar su raíz cuadrada

- Es una medida de dispersión que tiene en cuenta a todas las observaciones
- Se expresa en las mismas unidades que la variable
- La desviación típica de forma aislada no dice nada, debe de ir acompañada de la media. Normalmente se expresa

$$\overline{x} \pm s$$

por ejemplo: "…*el nivel de colesterol observado en la muestra fue* 175.3±12.5*.…*"

- Cuando la distribución de la variable es simétrica (mas concretamente *normal**) el intervalo

$$\overline{x} \pm 2s$$

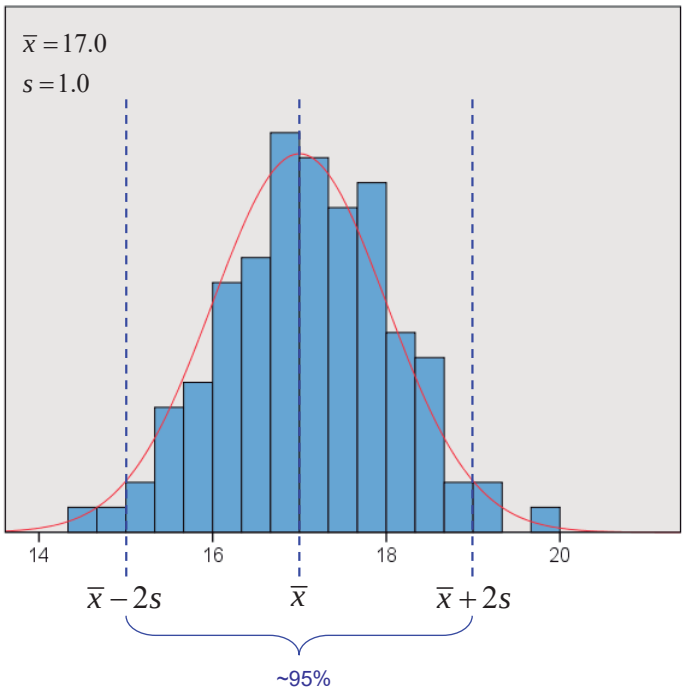
contiene aproximadamente al 95% de las observaciones.
En el ejemplo anterior, aproximadamente el 95% de los casos tienen un nivel de colesterol comprendido en el intervalo (150.5, 200.5) si la distribución del colesterol es simétrica (*normal**)

* Más adelante se verá qué quiere decir exactamente que una distribución sea *normal*

Pedro Femia
pfemia@ugr.es

Posición y dispersión sobre el diagrama de frecuencias

Indicar $\bar{x} \pm s$ define muy bien la distribución, especialmente si es **simétrica** (*normal**)



* Más adelante se verá qué quiere decir exactamente que una distribución sea *normal*

Síntesis de datos: medidas descriptivas de dispersión

Otras medidas de dispersión

Coeficiente de variación

$$CV = \frac{s}{\bar{x}}(\times 100\%)$$

- Es una medida de dispersión adimensional (se multiplica por 100 y es un %)
- Expresa la dispersión dada por la desviación típica en términos de la media (estandarización)
- No debe calcularse si la media es inferior a 1
- Permite comparar la variabilidad de muestras con diferente media e incluso con diferentes unidades
- Muy utilizada para expresar la precisión de los aparatos de medida

Rango intercuartílico (RIQ)

$$RIQ = Q_3 - Q_1$$

Es una medida de dispersión basada en percentiles, por tanto
- Se usa cuando la medida de posición considerada es la mediana
- No tiene sentido con muestras de pequeño tamaño
- Debe interpretarse como que en un margen de RIQ unidades (entre Q1 y Q3) está el 50% (central) de las observaciones

- Una variación de esta medida es el **rango semi intercuartílico (RSI)** o **desviación cuartil**

$$RSI = \frac{Q_3 - Q_1}{2}$$

Pedro Femia
pfemia@ugr.es

Síntesis de datos: ejercicio

Ejercicio

Calcule la media, la desviación típica y la mediana de las 7 muestras que aparecen a continuación ¿Puede decirse que las muestras 3 y 6 tienen la misma dispersión?

La media suele presentarse con un decimal más que los valores de la variable y la desviación típica con un decimal más que la media

Muestra										
1)	10.2	11.4	11.0	10.2	11.9	11.1	10.8	10.1	10.8	11.1
2)	102	109	107	106	103	103	105	102	104	106
3)	2.9	3.0	3.1	3.5	2.7	2.9	3.1	3.4	2.6	
4)	1.0	1.1	2.6	4.0	0.0	1.3	3.5	-1.0	1.0	1.5
5)	-0.6	-1.6	-1.6	-2.7	0.9	-0.3	-0.6	-2.9	-0.9	-2.2
6)	5.9	6.0	6.1	6.5	5.7	5.9	6.1	6.4	5.6	
7)	0.5030	0.4961	0.5001	0.4961	0.5003	0.5023	0.4972			

Solución

n	Media	s	Me
10	10.86	0.574	10.9
10	104.7	2.31	104.5
9	3.02	0.295	3.0
10	1.50	1.515	1.2
10	-1.25	1.177	-1.25
9	6.02	0.295	6.0
7	0.49930	0.00286	0.5001

La respuesta a la pregunta es que no. La media (dt) en la muestra 3 es 3.02 (0.295) mientras que en la 6 resulta ser 6.02 (0.295). Las desviaciones típicas, aunque numéricamente presentan el mismo valor, no son comparables, ya que ambas muestras tienen diferente media (la de 6 es el doble que la de 3). Si obtenemos los coeficientes de variación resulta que $CV_{(3)}$= 9.8% y $CV_{(6)}$=4.9% lo que pone de manifiesto que la dispersión relativa a la media de la muestra 3 es el doble que la de la muestra 6.

Pedro Femia
pfemia@ugr.es

ESTADÍSTICA

Grado en Enfermería

Tema II

Nociones sobre probabilidad, variable aleatoria y muestreo

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Introducción

Determinismo y aleatoriedad

Tipos de fenómenos (sucesos)

Fenómeno determinista: es aquel que se puede reproducir siempre que se desee estableciendo unas mismas condiciones de partida: iguales condiciones → iguales resultados (determinismo ←→ certidumbre)

Fenómeno aleatorio: es aquel cuya aparición viene condicionada por el azar. No se puede reproducir siempre que se desee estableciendo unas mismas condiciones de partida: iguales condiciones → resultado incierto (aleatoriedad ←→ incertidumbre).

Fenómenos en la naturaleza

Fenómenos totalmente deterministas: el determinismo está relacionado con las Teorías (por ejemplo, el *Determinismo Biológico*), pero en la naturaleza el determinismo absoluto es más que cuestionable

Fenómenos totalmente aleatorios: los resultados de los juegos de azar (lanzar una moneda, un dado, …)

Fenómenos naturales: al observar el mundo físico siempre hay un mayor o menor grado de aleatoriedad implícito en las observaciones (el *error de medida* es un primer responsable de esto)

Fenómeno observado = componente determinista + componente aleatorio

Peso = f **(talla, sexo,…)** (*componente determinista*) **+ variabilidad** (*componente aleatorio*)

El Peso de un individuo depende de su talla, sexo, etc…

No todos los hombres que miden 1.80m pesan lo mismo

***Modelo* estadístico**

Un papel fundamental de la **Inferencia Estadística** consiste en "separar" la partes determinista y aleatoria

Pedro Femia
pfemia@ugr.es

Concepto de probabilidad

La probabilidad es una medida

El estudio de los fenómenos aleatorios requiere cubrir la necesidad de poder **medir con qué frecuencia se presentan** → Concepto de **Probabilidad**

Probabilidad de un suceso A, $P(A)$ = medida de lo frecuente o lo extraño que resulta observar ese suceso A

No hay una definición única del concepto de Probabilidad

La **noción de probabilidad** se presta a *diferentes aproximaciones* y es objeto de particular atención en al ámbito de la *Epistemología*. Esto genera diferentes enfoques:

- Basada en la regla de Laplace
- Definición frecuencista
- Definición axiomática (de Kolmogorov, es la más formal)
- (Definición subjetiva de probabilidad)

Regla de Laplace

$$P(A) = \frac{\text{nº de casos favorables a A}}{\text{nº de casos posibles}}$$

Independientemente del enfoque, la **probabilidad de un suceso es un número positivo comprendido entre 0 y 1** (que se puede multiplicar por cien y darlo como porcentaje)

Que ocurra A es:

- Posible, si $P(A) > 0$
  - Seguro, si $P(A) = 1$
  - Probable, si $0 < P(A) < 1$
    - si $P(A) \nearrow 1$ A es frecuente
    - si $P(A) \searrow 0$ A es raro
- Imposible, si $P(A) = 0$

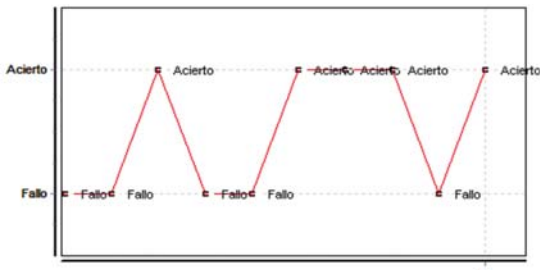
Es importante tener presente que la **probabilidad es una medida** (como si habláramos de la longitud o del peso) de lo ***frecuente*** o ***infrecuente*** (=raro) que resultar observar un determinado suceso

Concepto de probabilidad

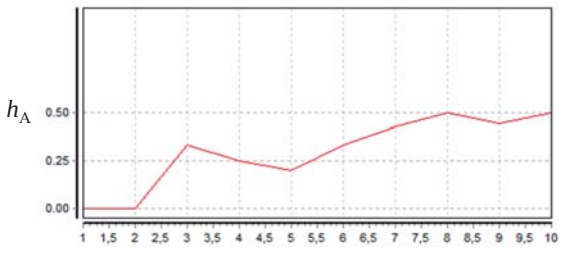
Perspectiva frecuencista: la Ley de Azar

Cada prueba es la observación de un suceso binario (resultado = Acierto/fallo)

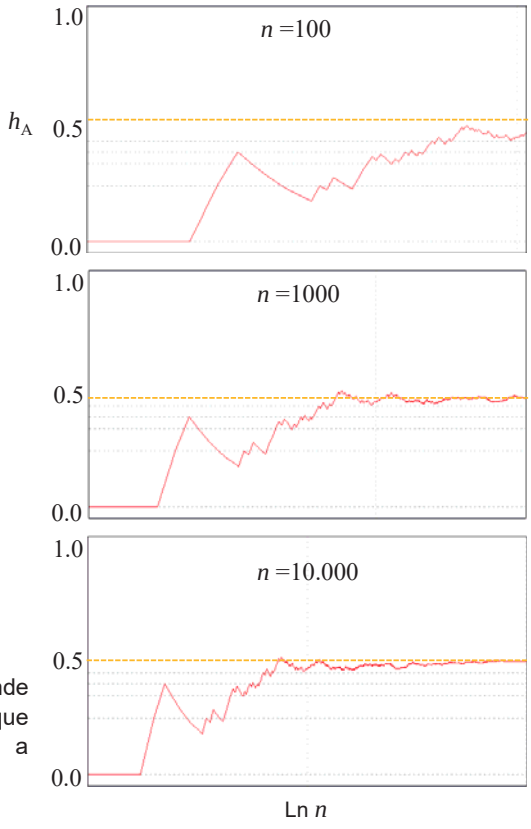
n=10 pruebas



Proporción (frecuencia relativa) de aciertos: h_A



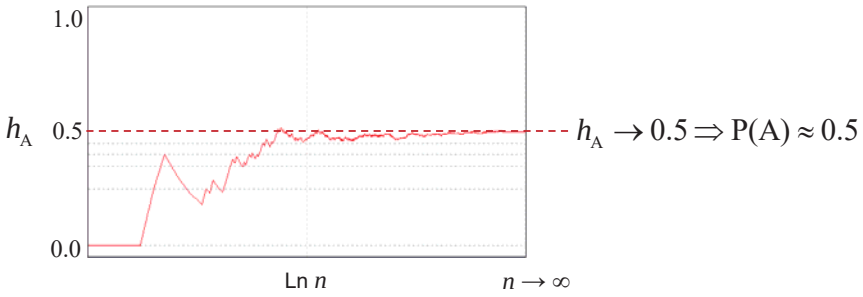
Al repetir un experimento un número n suficientemente grande de veces (se dice que $n \to \infty$) la frecuencia relativa con la que aparece cada posible modalidad del resultado tiende a **estabilizarse** en un valor (*Ley de los grandes números*)



Pedro Femia
pfemia@ugr.es

Perspectiva frecuencista: la Ley de Azar

Definición frecuencista de probabilidad: La **probabilidad** con la que se presenta un fenómeno es el valor al que se aproxima la **frecuencia relativa** con la que aparece dicho fenómeno cuando el **número de realizaciones** de la experiencia que permite observarlo es "**suficientemente**" **grande**



$$h_A = \frac{\text{nº de veces que aparece la modalidad A (casos favorables a A)}}{\text{nº total de pruebas realizadas (casos posibles)}} \xrightarrow{n\to\infty} P(A)$$

- Formalmente se indica $P(A) = \lim_{n\to\infty} h_A$
- Obsérvese que bajo esta definición $0 \le P(A) \le 1$ $\begin{cases} P(A) \leftarrow h_A = 0 & \text{Si nunca se observa A} \\ P(A) \leftarrow h_A = 1 & \text{Si siempre se observa A} \end{cases}$
- Por otra parte, si un fenómeno puede aparecer en forma de solo una de k modalidades posibles: $A_1, A_2, \ldots, A_k$ entonces

$$P(A_1) + P(A_2) + \ldots + P(A_k) = \sum P(A_i) = 1 \qquad \text{(Probabilidad total)}$$

Concepto de variable aleatoria (VA)

- Por ***variable*** entendemos aquí la acepción matemática, es decir, se trata de un **valor numérico** que puede ser cambiante de una observación a otra

  Como los valores numéricos pueden ser discretos (=recuentos) o continuos (=medidas) hablaremos de **variables discretas** y de **variables continuas**
- Por ***aleatoria*** entendemos que dicho valor numérico **depende en todo o en parte del azar**
- Entonces definir una **variable aleatoria** consiste en asignar un **valor numérico** al resultado de un **experimento aleatorio**

  Hablaremos de

  • Variable aleatoria ***discreta*** (VAD) ← recuentos cuyo valor esta afectado por el azar (ej: nº de hijos,…)

  • Variable aleatoria ***continua*** (VAC) ← medidas cuyo valor está afectado por el azar (ej: peso,…)

- Los **modelos estadísticos** (o matemáticos en general) solo pueden desarrollarse con valores numéricos (no con categorías nominales), por lo tanto, la forma de estudiar los **datos de tipo nominal** es transformarlos en variables discretas, y esto se hace **contando** cuantas veces aparece cada categoría (es decir, su frecuencia)

  Por ejemplo: Para estudiar la distribución por sexos (sexo=dato de **tipo nominal**) de los afectados por una lesión medular se estudia el **número de hombres** (o equivalentemente el número de mujeres) que están afectados por dicha lesión, obteniendo así una ***variable aleatoria discreta***.

- Indicar qué valores puede tomar la VA y la probabilidad con que toma cada uno de ellos supone indicar la **distribución de probabilidad** de dicha variable

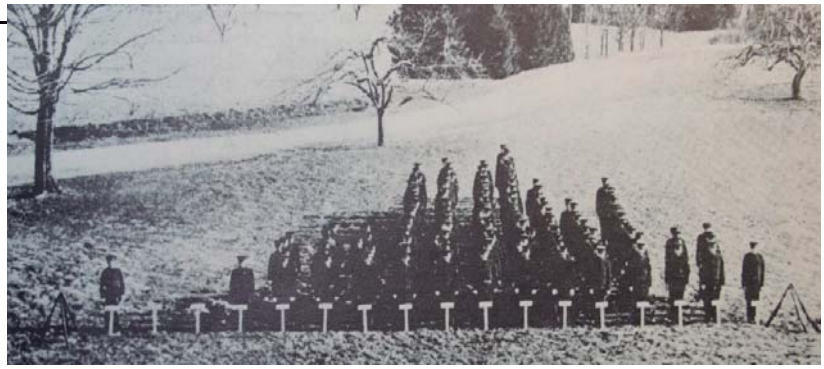
Vamos a profundizar en el concepto de modelo de **distribución de probabilidad** de variables aleatorias estudiando algunos patrones concretos (=*modelos*): **Normal** (para VAC), **Binomial** y **Poisson** (para VAD)

Pedro Femia
pfemia@ugr.es

Histogramas vivientes

Distribución de la estatura de 175 hombres reclutados por el ejército inglés a finales del siglo XIX

Tomado de Blakeslee & Hered (1914) por Ayala & Kiger (1984) *Modern Genetics*



Ambas imágenes representan la ***distribución* de la estatura** de un colectivo diferente en épocas diferentes. (A la izda. están los más bajos y a la dcha los individuos más altos)

¿Tienen algo en común estas dos ***distribuciones* de frecuencias** de la estatura?

¿Qué aspecto puede esperarse de la distribución si se hiciera con los compañeros de clase?

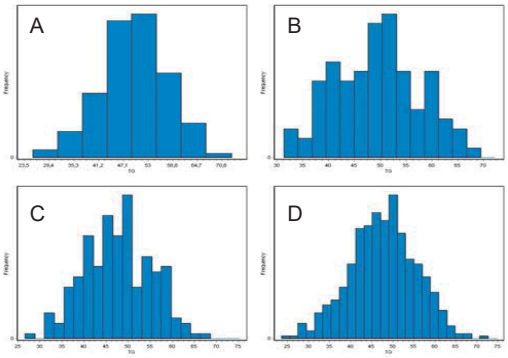
Solución:

Brian Joiner "Living Histograms" International Statistical Review, 43 (1975)

Qué es un modelo de distribución de probabilidad

Consideremos cuatro muestras de una VAC (como la *estatura* o el *nivel de colesterol*)

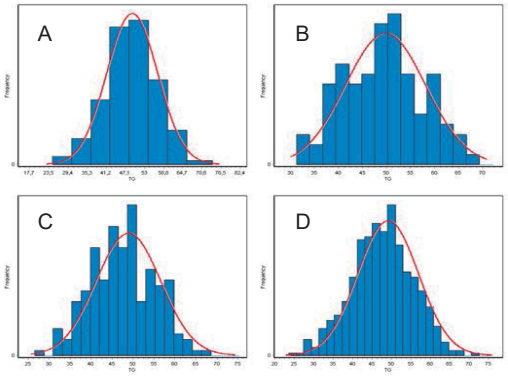
¿Tienen algo en común estas distribuciones de frecuencias correspondientes a cuatro **muestras aleatorias** tomadas de la misma población?



¿Cómo es el **patrón poblacional** que genera estos datos?

¿Corresponden a un mismo **modelo**

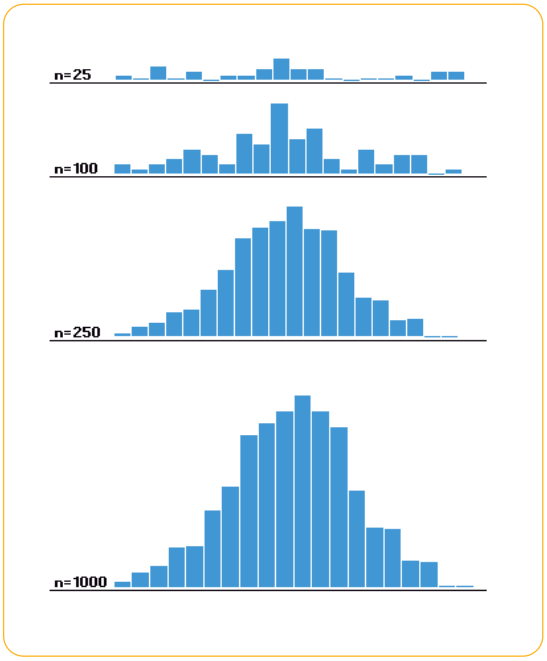
de distribución de la variable?



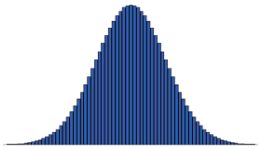
Pedro Femia
pfemia@ugr.es

Qué es un modelo de distribución de probabilidad

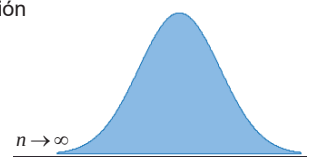
1. ¿Qué ocurre con la forma de la distribución si cada vez observamos más datos?



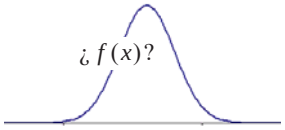
2. Conforme aumentamos el número de observaciones es posible considerar cada vez **más intervalos más estrechos**.



3. La cuestión es si en el **límite** (muchos intervalos, muy estrechos) se perfila siempre **un mismo patrón** (suave) de la distribución



4. ¿Es posible establecer un **modelo matemático** $f(x)$ (una expresión matemática manejable) que represente bien la forma como se **distribuye** la VA en la **población**?



Si existe $f(x)$, entonces dicho **modelo caracterizará** a la **distribución de toda la población** y resulta una **herramienta valiosísima** para hacer **inferencia**

$n \to \infty$ indica que estaríamos aproximándonos al tamaño de toda la población.

Qué es un modelo de distribución de probabilidad

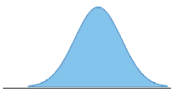
Utilidad de los modelos de distribución de probabilidad

- Los modelos caracterizan –cuando son correctos– el comportamiento de la VA estudiada en la población
- Un modelo de distribución de probabilidad permite determinar la probabilidad de que la VA estudiada tome determinados valores.

Por ejemplo: P(*talla* < 170 cm); P(180 ≤ *nivel de colesterol* ≤ 210),
P(*nº de seropositivos al observar a 300 individuos sea* > 2), etc

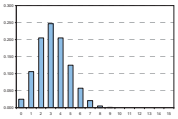
Caracterización de los modelos de distribución de probabilidad

- Un modelo de distribución de probabilidad será una expresión matemática en la que intervendrá la VA X que está siendo caracterizada y unos valores constantes que son **los parámetros del modelo.**
- Cuando un modelo es adecuado para representar la distribución de una población, sus parámetros suelen tener un **significado *físico*** en relación a dicha población.



Tipos de modelos de distribución de probabilidad

- Modelos de distribución para VAC: el modelo fundamental es el **modelo de distribución normal**
- Modelos de distribución para VAD: son de interés los modelos **Binomial** y de **Poisson**



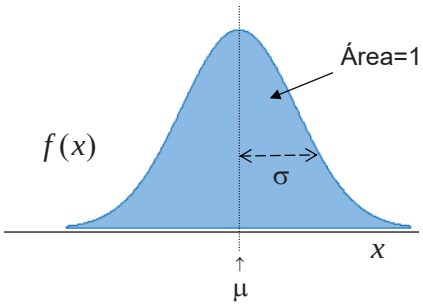
Además de los indicados hay muchos mas modelos que sirven para caracterizar fenómenos concretos (distribución uniforme, exponencial,…)

Veremos algunos modelos que se dan en llamar **distribuciones en el muestreo** y que son herramientas (más abstractas) para la *inferencia*, como son las distribuciones *t*-student, χ^2, F de Snédecor, etc… que usaremos más adelante en este curso

Pedro Femia
pfemia@ugr.es

Modelos de distribución de probabilidad de VAC

El modelo de distribución normal



Cuando el ***modelo de distribución Normal*** se utiliza para representar la **distribución de probabilidad** de una población (en términos de la variable aleatoria X), los **parámetros** del modelo tienen una relación directa con parámetros descriptivos de tal población:

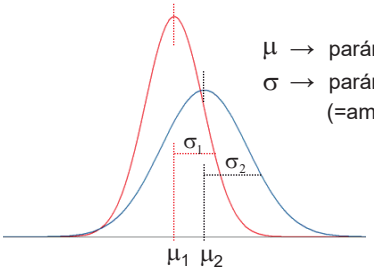
$\mu \rightarrow$ Media* de la variable aleatoria X
$\sigma \rightarrow$ Desviación típica de X
$\sigma^2 \rightarrow$ Varianza de X

Se indica: $X \rightarrow N(\mu;\sigma)$

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

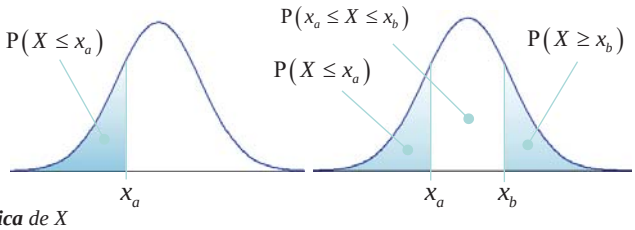
$f(x)$ es la ***función de densidad*** de la *distribución normal* o *gaussiana* (= *Campana de Gauss*)

El ***modelo Normal*** contempla 2 parámetros:



$\mu \rightarrow$ parámetro de posición
$\sigma \rightarrow$ parámetro de dispersión (=amplitud de la curva)

La probabilidad asociada a dicho modelo es **un área bajo la curva** definida por el mismo



* Es frecuente hablar también en términos de ***Esperanza Matemática*** *de X*

Modelos de distribución de probabilidad de VAC

Modelo de distribución normal

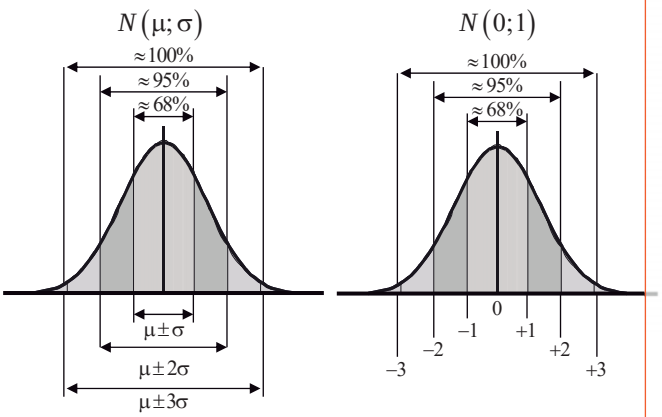
El **cálculo de probabilidades** = **cálculo de áreas** bajo la curva $f(x)$ se hace mediante **cálculo integral**, por tanto la probabilidad de que la variable se encuentre entre dos valores concretos x_a y x_b viene dada por:

$$P(x_a \le x \le x_b)=\int_{x_a}^{x_b} f(x)dx$$

¡¡ Difícil !!

Para facilitar esto (sin resolver integrales) se utilizan las **tablas de la distribución normal**

Algunas áreas de especial interés son (de forma aproximada):



Tipificación

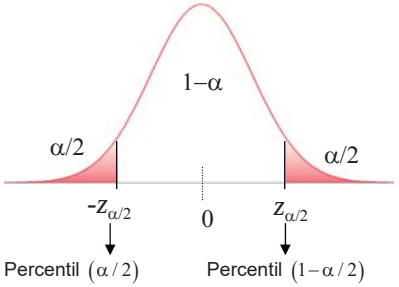
Tipificar una observación es una forma de **estandarizarla** en relación a un **grupo de referencia**

Si $X \rightarrow N(\mu;\sigma)$ entonces $Z=\frac{X-\mu}{\sigma}$ se dice que es **la tipificación de** X y verifica

$$Z \rightarrow N(0;1) \qquad \begin{cases} \mu_Z = 0 \\ \sigma_Z = 1 \end{cases}$$

Interés de la tipificación:

- Permite requerir solamente una tabla de la distribución normal, la de la distribución ***normal estándar*** o ***N*(0,1)**
- **Estandarizar** las observaciones eliminando sus dimensiones y de acuerdo a un grupo de referencia, permitiendo localizar a la observación en el grupo (análogo a lo que hacen los **percentiles**) y comparar con valores tipificados de otras variables



Pedro Femia
pfemia@ugr.es

Dos distribuciones discretas de especial relevancia:

Distribución Binomial

- Distribución asociada a las **proporciones**

  Ejemplo de uso:

  Si la prevalencia del síndrome metabólico (SM) en diabéticos es del 52.4% ¿qué número de afectados del por el SM se espera encontrar al observar una muestra de 200 individuos diabéticos? ¿Cuál es la probabilidad de encontrar en dicha muestra más afectados por el SM que los esperados?

Distribución de Poisson

- Distribución asociada a los **recuentos** por unidad de tiempo, volumen, área,…

  Ejemplo de uso

  Se viene observando que por término medio hay 3 lesiones de rodilla por semana durante la temporada de esquí en Sierra Nevada. ¿Cuál es la probabilidad de que el número de lesiones de rodilla en una semana de la temporada sea superior a la media observada?

Distribución Binomial

Definición

Consideramos la presencia de una característica A en una población de tamaño *N* individuos (se trata entonces de una variable binaria o dicotómica con modalidades presente/ausente respecto a A). En ella una proporción del $p\times100\%$ presentan la característica A y el resto, el 1-p = q ×100% no la presentan. Si de ella tomamos una muestra de tamaño n, y anotamos el número de individuos de la muestra que presentan la característica A, entonces la VA discreta X es ***Binomial*** siempre que se cumpla una de estas dos condiciones:

- $N \to \infty$ Es decir, la población es suficientemente grande
- $N > 40$ y n/N (*fracción de muestreo)* **≤ 0,10.** Es decir, la población no es tan grande pero la muestra es de tamaño despreciable respecto a ella

Se indica $X \to B(n, p)$

Parámetros: la distribución tiene dos parámetros

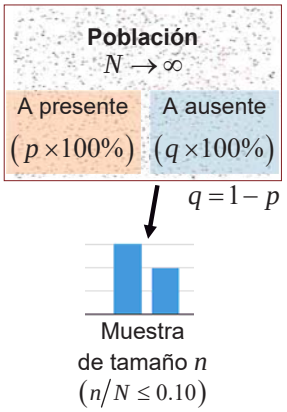
n: el número de casos observados o de pruebas realizadas
p: la proporción de éxitos en la población

Valor esperado (media) y varianza de X

Valor esperado*: np

Varianza: npq

Modelo Binomial



X = nº de observaciones en la muestra con A presente

$$\Downarrow$$

$$X \to B(n, p)$$

Función de probabilidad de X

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x=0,1,2,...,n$$

* Se habla de *esperanza matemática*

Pedro Femia
pfemia@ugr.es

Distribución Binomial. Ejemplo

complemento

¡¡No se exigen estos cálculos. Se trata de ver solo el razonamiento!!

Según los datos de la Encuesta Nacional de Salud de 2012, el 21.7% de la población española de 15 a 24 años fuma diariamente (independientemente del sexo). Se dispone de una muestra de n=15 personas de tal edad. Observe que:

X = número de fumadores; $X \to \mathrm{B}(15; 0.217) \to \mathrm{P}(X = x) = \binom{15}{x} 0.217^x (1-0.217)^{15-x}; \quad x = 0,1,2,...,15$

Número esperado de fumadores en la muestra: $np = 15 \times 0.217 = 3.25$ (aproximadamente tres)

Probabilidad de no encontrar **ninguno**: $\mathrm{P}(X = 0) = \binom{15}{0} 0.217^0 (1-0.217)^{15} = 0.025 \quad (2.5\%)$

Probabilidad de encontrar **alguno**:

$\mathrm{P}(X > 0) = 1 - \mathrm{P}(X = 0) = 1 - 0.025 = 0.974 \quad (97.4\%)$

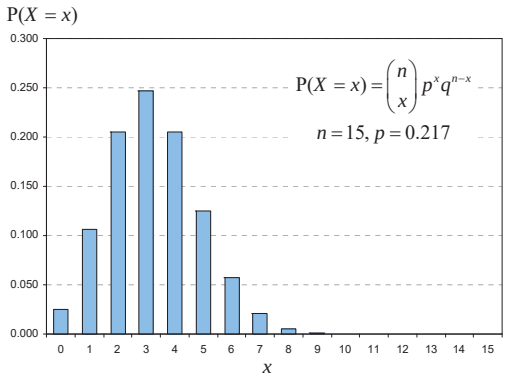
Probabilidad de encontrar **2 o 3**: $\mathrm{P}(X = 2) + \mathrm{P}(X = 3) =$

$\binom{15}{2} 0.217^2 (1-0.217)^{15-2} + \binom{15}{3} 0.217^3 (1-0.217)^{15-3} =$

$0.206 + 0.247 = 0.452 \quad (45.2\%)$

Probabilidad de encontrar entre cero y 15 fumadores =1 (=100%) ¿por qué?

Diagrama de probabilidad de la B(15,0.217)



Distribución de Poisson

Definición

Si X es una VAD que representa alguno de los siguientes recuentos

- El número de sucesos que ocurren independientemente y de forma no simultánea en el tiempo
- El número de partículas distribuidas al azar en una gran cantidad de medio
- El número de casos con una **rara característica** A cuando se observa un número grande de sujetos (= **ley de los sucesos raros**, una binomial con p pequeño, menor a 0.05, y n grande, mayor a 20)

entonces X tiene una distribución de Poisson

Se indica $X \to P(\lambda)$

Parámetros: la distribución tiene un solo parámetro

λ: el número medio de ocurrencias por unidad de tiempo, volumen, etc.

Valor esperado (media) y varianza de X

Valor esperado*: λ
Varianza: λ

En la distribución de Poisson se da la circunstancia particular de que ¡la media y la varianza de la VA coinciden!

Modelo de Poisson

Nº de lesiones medulares al año

Nº de salidas del 061 al día

Nº de hematíes por ml de sangre

X = nº de sucesos independientes que ocurren por unidad de tiempo, área, volumen,....

$\Downarrow$

$X \to P(\lambda)$

Función de probabilidad de X

$$\mathrm{P}(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x=0,1,2,3,...$$

* Se habla de *esperanza matemática*

Pedro Femia
pfemia@ugr.es

Distribución de Poisson. Ejemplo

complemento

¡¡No se exigen estos cálculos. Se trata de ver solo el razonamiento!!

Por término medio se registran en España 1.5 lesiones medulares al día.
Si asumimos que el número de lesiones medulares al día tiene es una VAD que se ajusta a una distribución de Poisson con parámetro λ=1.5, observe que:

X = número de lesiones/día $\quad X \rightarrow \mathrm{P}(1.5) \rightarrow \mathrm{P}(X=x) = \dfrac{e^{-1.5}1.5^x}{x!}; \quad x = 0,1,2,...$

Número esperado de lesiones en un día cualquiera: $\lambda = 1.5$ (entre una y dos)

Probabilidad de que un día no se observe **ninguna** lesión: $\mathrm{P}(X=0) = \dfrac{e^{-1.5}1.5^0}{0!} = 0.223 \quad (22.3\%)$

Probabilidad de que un día se observe **alguna** lesión:

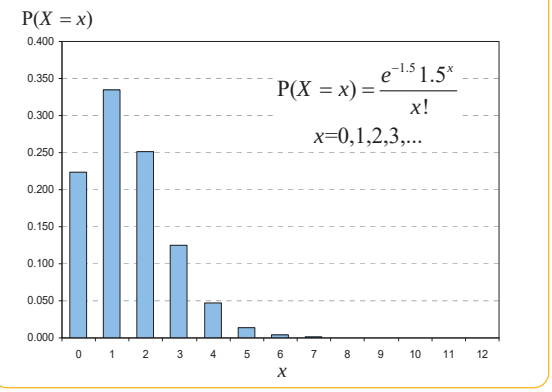
$\mathrm{P}(X>0) = 1 - \mathrm{P}(X=0) = 1 - 0.223 = 0.777 \quad (77.7\%)$

Probabilidad de observar **2 o 3**: $\mathrm{P}(X=2)+\mathrm{P}(X=3) =$

$\dfrac{e^{-1.5}1.5^2}{2!} + \dfrac{e^{-1.5}1.5^3}{3!} = 0.251 + 0.125 = 0.376 \quad (37.6\%)$

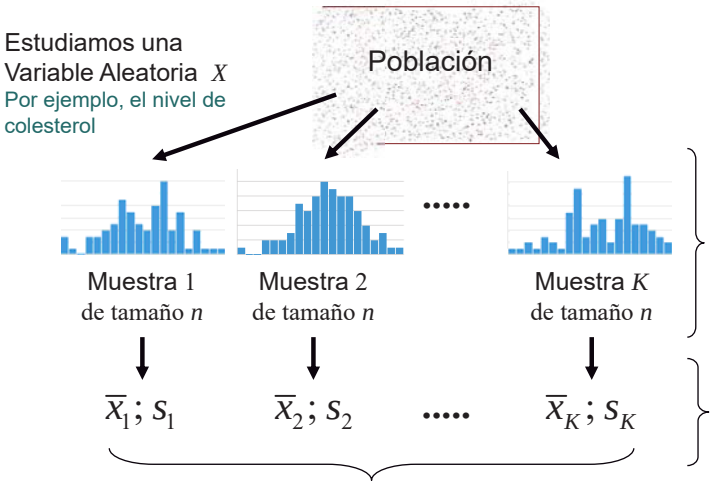
Probabilidad de encontrar cero o más lesiones =1 (=100%) ¿por qué?

Diagrama de probabilidad de la P(1.5)



La media muestral como VAC y su distribución

IMPORTANTE

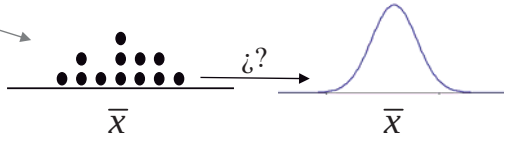


(1) Las muestras se eligen **al azar** (las K muestras tienen la misma probabilidad de aparecer)

(2) Todas las **medidas descriptivas** son **variables aleatorias** (además continuas)

Muestra	tamaño	media
1	n	$\overline{x}_1$
2	n	$\overline{x}_2$
...	...	...
K	n	$\overline{x}_K$

(3) ¿Se puede decir algo de la **distribución** que tienen las **medias muestrales**?



Pedro Femia
pfemia@ugr.es

La media muestral como VAC y su distribución

IMPORTANTE

Teorema del Límite Central (TLC)

- Resultado de importancia fundamental en Inferencia Estadística
- Justifica por qué la campana de Gauss se suele llamar distribución "**normal**" (en relación a su abundancia)
- En esencia se trata de lo siguiente:

1. Si una variable aleatoria surge como la **suma** de variables aleatorias **independientes** de magnitud similar **con distribución normal** entonces el resultado también tiene **distribución normal**

2. Si una variable aleatoria surge como la **suma** de variables aleatorias **independientes** de magnitud similar con **distribución desconocida** entonces el resultado tiene distribución **aproximadamente normal** siempre que haya un número de sumandos suficientemente grande (se considera "suficiente" a partir de 60)

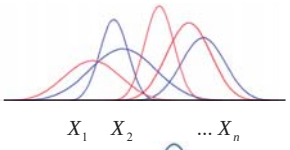
En la naturaleza muchas variables toman valores que son la suma de muchos factores

- Los errores de medida son suma de muchos factores, se consideran normales
- Las variables fisiológicas (en individuos sanos) toman valores que son fruto de la suma de muchos factores (herencia, dieta, modo de vida,…), se suele considerar que estas variables tienen distribución normal

1

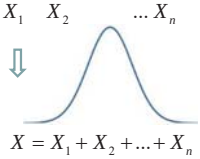
Si

$X_i \to N(\mu_i, \sigma_i)$



entonces

$X_i \to N(\mu_i, \sigma_i)$

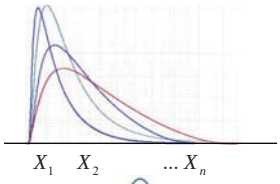


$X = X_1 + X_2 + ... + X_n$

2

Si

$X_i \to ??$

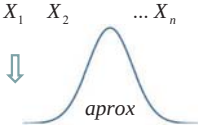


entonces

$X = \sum X_i \overset{aprox}{\to} N(\mu, \sigma)$

Siempre que sea

$n > 60$



$X = X_1 + X_2 + ... + X_n$

La media muestral como VAC y su distribución

IMPORTANTE

Teorema del límite central

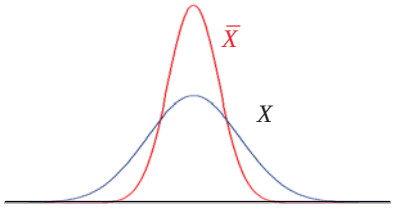
- La media muestral es una variable aleatoria continua que se obtiene como la suma de n valores de magnitud similar

$$\bar{X} = \frac{\sum x_i}{n}$$

Entonces:

1. Si la variable es normal, la media también.
   ¿Con qué parámetros?

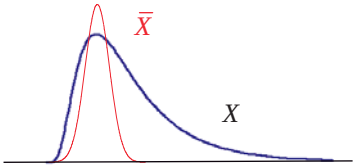
$$X \to N(\mu;\sigma) \Rightarrow \bar{X} \to N\left(\mu;\frac{\sigma}{\sqrt{n}}\right)$$



   - La media de las medias, es también la media poblacional, **la posición no cambia**
   - La **dispersión** de la media se reduce en la medida en que el tamaño de muestra aumenta; la media gana ***robustez***, ahora estamos viendo cómo lo hace

2. Si la variable no tiene distribución normal, la media muestral tiene distribución aproximadamente normal cuando el tamaño muestral (n) *es suficientemente grande* (mayor a 60).

$$X \to ?? \Rightarrow \bar{X} \overset{\text{aprox.}}{\underset{n\geq 60}{\to}} N\left(\mu;\frac{\sigma}{\sqrt{n}}\right)$$



Este resultado es muy importante. En inferencia, lo que más vamos a usar es la media. Si se puede considerar que la distribución de la media se adapta a un modelo normal, entonces se dispone de una herramienta muy potente. En breve ***toda inferencia basada en la normalidad es fácil de realizar***

Pedro Femia
pfemia@ugr.es

Inferencia Estadística. Población y muestra

Población: Conjunto de *individuos* -o más genéricamente *unidades*- sobre los que se desea hacer alguna afirmación (el objeto de nuestro estudio).

¿Cómo definirlas?

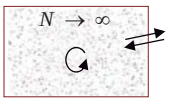
La población objetivo del estudio debe **definirse** mediante

- Condiciones de ***inclusión***: qué requisitos se deben cumplir los sujetos para pertenecer a la misma Ej. Personas de ambos sexos con edad superior a 80 años
- Condiciones de ***exclusión*****:** Características que no se deben cumplir**.** Ej. Padecer algún tipo de traumatismo discapacitante

Las poblaciones, en su totalidad, son inaccesibles

¿Cómo estudiarlas?

El concepto de población implica la connotación de ***gran tamaño*** lo que supone un gran esfuerzo estudiarlas en su totalidad (imposibilidad física y económica)



En general, las poblaciones son **dinámicas**:

- Hay individuos que se incorporan (inmigración, nacimiento)
- Hay individuos que desaparecen de ella (emigración, muerte)
- Hay individuos que cambian su estatus (pueden dejar de verificar las condiciones de inclusión o exclusión)

Las poblaciones, en su totalidad, son difícilmente descriptibles

¿Cómo caracterizarlas?

Las poblaciones suelen caracterizarse en términos de determinados **parámetros**, en función de ellos será posible responder a determinadas preguntas

Ejemplo

Pregunta	Parámetro de interés
¿Es la población infantil más obesa?	← **Media** del IMC
¿Ha disminuido el número de seropositivos del VIH?	← **Prevalencia** del VIH

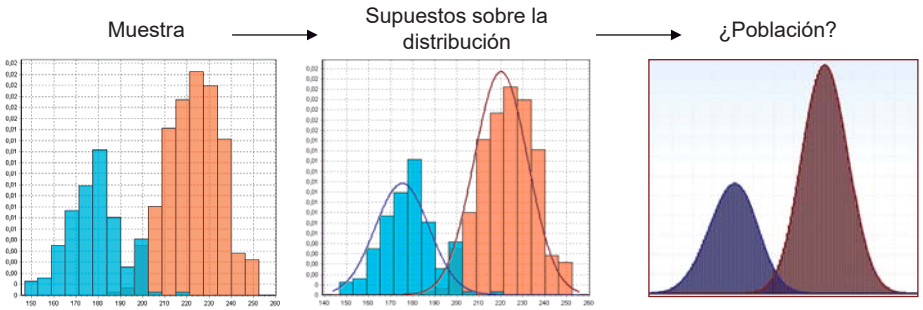
Inferencia Estadística. Población y muestra

Inferencia

Ante la imposibilidad de estudiar a toda la población, se debe considerar un subconjunto de ella: una **muestra**

Para generalizar (extrapolar) a toda la población la información suministrada por la muestra obtenida a partir de ella es necesaria una forma de **razonamiento inductivo**: la **inferencia estadística**

El cálculo de probabilidades y los modelos de distribución de probabilidad constituyen una herramienta fundamental para poder realizar la inferencia. A menudo hay que establecer determinados **supuestos** acerca de cómo se distribuye la variable estudiada en la población



La muestra debe estar bien escogida, debe **representar** a toda la ***población objetivo*** de forma que sea una imagen lo mas fiel posible de la misma.

El **tamaño de la muestra NO** garantiza su representatividad. Lo que si es cierto es que una muestra mayor tendrá más contenido informativo que una muestra pequeña. Un aspecto fundamental en lo que sigue es conocer qué aspectos están condicionados por el tamaño muestral

Pedro Femia
pfemia@ugr.es

Formas de obtener los datos de la población

Dos formas de obtener la información

- **Muestreo**: Se ***observan*** individuos de la población sin que el investigador establezca de antemano valores de ninguna variable.
  - Un estudio cuyos datos se obtienen por muestreo es un **estudio observacional**
  - Los estudios observacionales solo permiten hablar de **asociación** entre variables, nunca de relaciones **causales**
    Ejemplo: Se podría decir que la prevalencia del cáncer de pulmón está asociada con el consumo de tabaco, pero no que fumar provoca cáncer
- **Experimentación**: el investigador fija de antemano valores de una o varias variables (=variables ***explicativas*** o ***tratamientos*** o ***factores***) y observa el nivel de otras (=variables ***respuesta***).
  - La asignación de individuos a cada nivel del factor se debe hacer de forma ***aleatorizada***
  - Los ***ensayos clínicos*** son estudios de este tipo: un grupo de individuos recibe un tratamiento y otro grupo actúa como **grupo control**
  - La forma de planificar la experiencia es el ***Diseño experimental***. (toda una disciplina estadística)

Sesgo = Tendencia de la muestra a diferir de la población de la que se extrae

Algunos sesgos peligrosos

- **Sesgo de selección**. Debido a la exclusión sistemática de una parte de la población
- **Sesgo de medida o de respuesta**. Uso de métodos de medida inadecuados (ej. Cuestionario no válido)
- **Sesgo por falta de respuesta.** Debido a la falta de información de individuos que fueron seleccionados para pertenecer a la muestra

Definición de muestra aleatoria

Para que la muestra sea **representativa** de la población objeto de estudio, es preciso que sea extraída de ella de modo que:

- Todos los individuos de la población tengan la misma probabilidad de ser seleccionados e incluidos en la muestra (**igual probabilidad**).
- La selección de un individuo no influya para nada en la selección o no de otro individuo cualquiera (**independencia**)

Una muestra elegida así se dice que es una ***muestra aleatoria simple***

Forma de escoger una muestra aleatoria

1. Tener determinada la población objeto del estudio. (con condiciones explícitas de *inclusión* y de *exclusión*).
2. Tener identificados a cada uno de los individuos de esa población y asignado a cada uno de ellos un número.
3. Elegir, por un *mecanismo que represente suficientemente bien al azar*, una muestra de los individuos de la población.

   La reproducción del azar se hace mediante la generación de **números aleatorios**
   - Tablas de números aleatorios
   - Generación de números aleatorios por ordenador o calculadora de bolsillo

Las tablas de números aleatorios.

Una tabla de números aleatorios es en una lista de dígitos del 0 al 9 con las siguientes características:

1. Cada dígito tiene la misma probabilidad de aparecer en cada posición de la tabla (**igual probabilidad**)
2. La presencia o ausencia de un dígito dado en determinada posición no condiciona para nada a los dígitos que le siguen o que le preceden (**independencia**)

Pedro Femia
pfemia@ugr.es

Tabla de números aleatorios

Selección aleatoria de n =10 individuos de una población de tamaño N =800

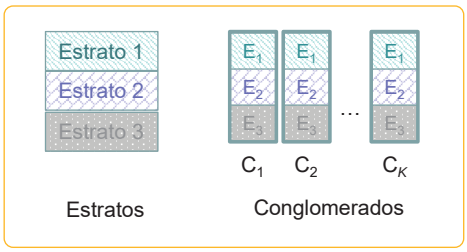
Tabla de números aleatorios

✓	0347437386	3696473661	4698637162	3326168045	6011141095
	9774246762	4281145720	4253323732	2707360751	2451798973
✓	1676622766	5650267107	3290797853	1355385859	8897541410
✓	1256859926	9696682731	0503729315	5712101421	8826498176
✓	5559563564	3854824622	3162430990	0618443253	2383013030
✓	1622779439	4954435482	1737932378	8735209643	8426349164
	8442175331	5724550688	7704744767	2176335025	8392120676
✓	6301637859	1695556719	9810507175	1286735807	4439523879
✓	3321123429	7864560782	5242074438	1551001342	9966027954
✓	5760863244	0947279654	4917460962	9052847727	0802734328
✓	1818079246	4417165809	7983861962	0676500310	5523640505
✓	2662389775	8416074499	8311463224	2014858845	1093728871
	2342406474	8297777781	0745321408	3298940772	9385791075
	5236281995	5092261197	0056763138	8022025353	8660420453
	3785943512	8339500830	4234079688	5442068798	3585294839
	7029171213	4033203826	1389510374	1776371304	0774211930
	5662183735	9683508775	9712259347	7033240354	9777464480
	9949572277	8842954572	1664361600	0443186679	9477242190
	1608150472	3327143409	4559346849	1272073445	9927729514
	3116933243	5027898719	2015370049	5285666044	3868881180
	6834301370	5574307740	4422788426	0433460952	6807970657
	7457256576	5929976860	7191386754	1358182476	1554559552
	2742378653	4855906572	9657693610	9646924245	9760490491
	0039682961	6637322030	7784570329	1045650426	1104966724

Otros esquemas de muestreo

complemento

A menudo la población se subdivide en **grupos** antes de realizar el muestreo aleatorio simple.
Dos tipos de agrupaciones:



- **Estratos**: Subconjuntos **homogéneos** de la población que son **internamente heterogéneos** algún criterio (variable de estratificación)

  Ejemplo: son estratos los grupos definidos por sexo, tramos de edad,…

- **Conglomerados**: Subconjuntos **heterogéneos** de la población que son **internamente homogéneos** bajo algún criterio (variable de estratificación)

  Ejemplo: Son conglomerados los grupos definidos por ciudades, centros hospitalarios, colegios,…

Idea intuitiva de un posible esquema de ***muestreo polietápico***:

Conocidos los conglomerados y los estratos de la población

1. Selección aleatoria de conglomerados
2. División en estratos dentro de cada conglomerado
3. Selección de sujetos dentro de cada estrato de cada conglomerado

1. Selección aleatoria de hospitales
2. En cada hospital hay Auxiliares, Personal de Enfermería y Médicos
3. Selección aleatoria de un nº determinado de Auxiliares, de Enfermeros/as y de Médicos

Pedro Femia
pfemia@ugr.es

Dos facetas de la **Inferencia Estadística:**

1. Estimación de parámetros

Se trata de asignar valores a determinados parámetros poblacionales de interés.

A menudo los parámetros poblacionales se corresponden con parámetros de una distribución de probabilidad

¿Cuál es el nivel medio de colesterol en la población infantil? → ¿μ? de una $N(\mu,\sigma)$
¿Cuál es la prevalencia del síndrome metabólico en la población? → ¿ p? de una $B(n,p)$
¿Cuál es el número medio de llamadas al 061 al día? → ¿λ? de $P(\lambda)$

2. Contraste (test) de hipótesis

Se trata de decidir si un enunciado (hipótesis estadística) es soportado por la evidencia empírica o no.

Las hipótesis estadísticas se formulan en términos de los parámetros poblacionales, que a menudo se corresponden con parámetros de una distribución de probabilidad

¿Esta relacionada la capacidad aeróbica con el sexo? → ¿$\mu_h=\mu_m$?
¿Ha variado la prevalencia del síndrome metabólico en la población en la última década? →¿$p_1=p_2$?

Pedro Femia
pfemia@ugr.es

ESTADÍSTICA

Grado en Enfermería

Tema III

Teoría de la estimación

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Recordemos que:

- Lo que interesa estudiar es a la **población** ← la población en su conjunto es inaccesible
- La población ha de caracterizarse de alguna manera "operativa" ← parámetros poblacionales
- Se necesita algún tipo de técnica que permita extender las conclusiones extraídas a partir de una o varias muestras a toda la población ← **Inferencia Estadística.**

Veremos **dos tipos** de inferencia que permiten dar respuesta a dos tipos de problemas diferentes:

- Teoría de la **Estimación** → ¿Cuál es el tiempo medio que tarda en hacer efecto un fármaco?
- Teoría de los **Contrastes de Hipótesis** → El tratamiento A ¿es igual de efectivo que el tratamiento B?

En ambos
- Se estudia a la población a través de sus **parámetros**
- El **tamaño de muestra** jugará un papel relevante que siempre habrá que tener en cuenta

La **Teoría de la Estimación** es la parte de la **Inferencia Estadística** que trata de determinar el valor de los parámetros poblacionales.

Distinguiremos **dos formas de estimación**:

- Estimación **puntual** ← lo mejor que se puede decir del parámetro a partir de la muestra reduciendo la información a **un solo valor**
- Estimación por **intervalo** ← la mejor manera de explotar la información contenida en la muestra de cara a conocer entre qué valores debe encontrarse el parámetro. En lugar de un solo valor, se da un conjunto de ellos.

Pedro Femia
pfemia@ugr.es

En la estimación puntual se trata de asignar al parámetro poblacional un único valor; que será un valor aproximado y que depende de la muestra

Conceptos de estimador y de estimación

Un estimador de un parámetro es una ***función*** de las observaciones de la muestra que permite dar valores apropiados para ese parámetro.

Notación: se utiliza el acento circunflejo sobre el símbolo del parámetro poblacional para aludir a su estimador

$\hat{\mu}$ es el estimador puntual de μ

Cuando se haya obtenido la muestra y determinado un valor para μ a partir de ella, se dice entonces que ese valor es una estimación de μ.

En la práctica los estimadores habituales son (*método analógico*)**:**

Hay más métodos de estimación, como el método de *mínimos cuadrados* que veremos en el último tema

Parámetros poblacionales de interés			**Estimadores puntuales**	
Media	μ	→	Media muestral	$\overline{x} = \hat{\mu}$
Desviación típica	σ	→	Desviación típica muestral	$s = \hat{\sigma}$
Proporción	π	→	Proporción muestral*	$p = \hat{\pi}$

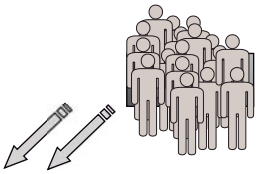
* En adelante usaremos p en lugar de π para mantener la misma notación que en los resúmenes

Problema que presenta el uso de estimadores puntuales:

El problema de los estimadores puntuales es que solo dan una idea de lo que puede valer el parámetro que estimamos, sin conocer como de buena es la aproximación; es decir, simplemente proporcionan un valor (de los muchos posibles) que puede proponerse como valor del parámetro.

Por ejemplo:

Considere la población de estudiantes de la Universidad de Granada. Estamos interesados en investigar el valor del parámetro μ=‘***peso medio de los estudiantes de la UGR***.’. Para ello seleccionamos aleatoriamente **dos muestras** de 20 estudiantes cada una y obtenemos el valor del peso medio en cada caso. ¿**coincidirán las dos medias muestrales**? ¿cuál es **mejor como estimación**? ¿qué **error** se comete al asumir como valor de μ el ofrecido por una de estas medias?



$$\left.\begin{array}{l} \text{Muestra } M_1: \ \overline{x}_1 = 67Kg \\ \text{Muestra } M_2: \ \overline{x}_2 = 71Kg \end{array}\right\} \Rightarrow ¿\mu?$$

La **estimación puntual** supone **reducir toda la información muestral a un solo valor**. Es obvio que dicho valor difícilmente coincidirá de forma exacta con el del parámetro poblacional

Intuitivamente **una muestra de tamaño grande** es **preferible a otra de tamaño menor** ¿por qué? (en la estimación puntual no se aprecia esto)

Pedro Femia
pfemia@ugr.es

Estimación por intervalo de confianza (IC)

La **estimación por IC** implica asumir que
- Es prácticamente imposible (por improbable) que valor único dado por el estimador puntual coincida con el valor del parámetro poblacional
- Es mejor dar un **conjunto de valores probables** para el parámetro, en lugar de uno solo
- Existe un **error de estimación** que en la estimación puntual no se pone de manifiesto

La estimación por IC consiste en asignar al parámetro poblacional desconocido un intervalo de valores, digamos (*a, b*) entre los cuales está dicho parámetro con una cierta probabilidad que denominaremos **nivel de confianza** y que se representa por $(1-\alpha)$. De aquí que α sea el **nivel de error** o probabilidad de que el intervalo obtenido no contenga al parámetro.

Diremos entonces que (*a, b*) es un **intervalo de confianza** para el parámetro μ construido al $(1-\alpha)\%$ de confianza (o al $\alpha\%$ de error) si se verifica que;

$$P(a \le \mu \le b) = 1-\alpha$$

La **confianza de un intervalo** debe **interpretarse** en el sentido siguiente: Por cada 100 intervalos que construyamos para estimar un mismo parámetro (a partir de otras tantas muestras aleatorias y para un valor α prefijado), en promedio el $(1-\alpha)\%$ de los intervalos obtenidos recogerán en su interior al verdadero valor del parámetro, mientras que el $\alpha\%$ restante, por cosas del azar, pueden resultar 'equivocados'. En la práctica se construye un único IC, así que si elegimos un nivel de confianza $(1-\alpha)\%$ grande, por ejemplo del 95%, entonces nuestra esperanza es que dicho intervalo sea uno de los 95 de cada 100 'acertados' y no uno de los 5 de cada 100 (en promedio) que no contienen al valor del parámetro.

Estimación por intervalo de confianza (IC)

Ejemplo: Se trata de estudiar el nivel de glucosa en sangre en la población cuyos valores aparecen en el recuadro sombreado (esta población es pequeñísima, pero nos vale como ejemplo). Seleccionamos **de forma aleatoria** 5 muestras de tamaño n = 5 y elaboramos, en cada caso, el intervalo de confianza para el nivel medio de glucemia (consideraremos un nivel de confianza del 95%). Observemos los resultados:

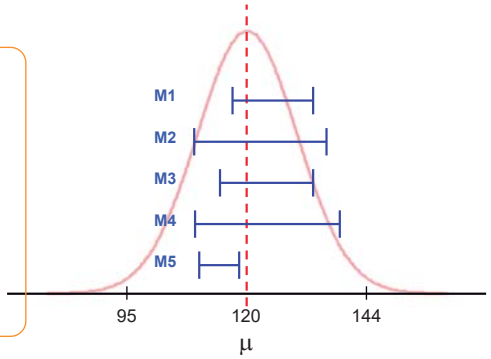
Población

108	118
112	120
123	133
109	127
121	125
136	115
124	117
113	125
118	117
129	110

$\mu = 120$

Muestra	Datos	$\bar{x}$	s	Intervalo (95% conf.)
M1	123 125 118 125 113	120.80	5.215	**(114.325 ; 127.275)**
M2	124 110 115 133 112	118.80	9.576	**(106.912 ; 130.688)**
M3	125 113 117 123 124	120.40	5.177	**(113.973 ; 126.827)**
M4	133 110 136 125 110	122.80	12.357	**(107.459 ; 138.141)**
M5	118 113 117 110 112	114.00	3.391	**(109.790 ; 118.210)** !

- Las estimaciones puntuales varían de muestra a muestra
- Los 5 intervalos tienen diferente amplitud ¿por qué?
- Los cuatro primeros intervalos contienen al verdadero valor de la media (que, excepcionalmente, por conocer a toda la población sabemos que vale 120), sin embargo, en la 5ª muestra los valores obtenidos son, por azar, mas bajos de la cuenta y dan lugar a un intervalo que ¡no contiene a μ!



Pedro Femia
pfemia@ugr.es

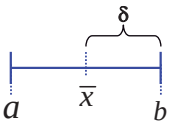
Estimación por intervalo de confianza (IC)

Intervalos construidos a nivel de confianza del 95% → $P(a \le \mu \le b) = 1-\alpha = 0.95$

Cuidado: El parámetro no "*cae*" en el intervalo. Es el intervalo el que intenta **contener al parámetro**

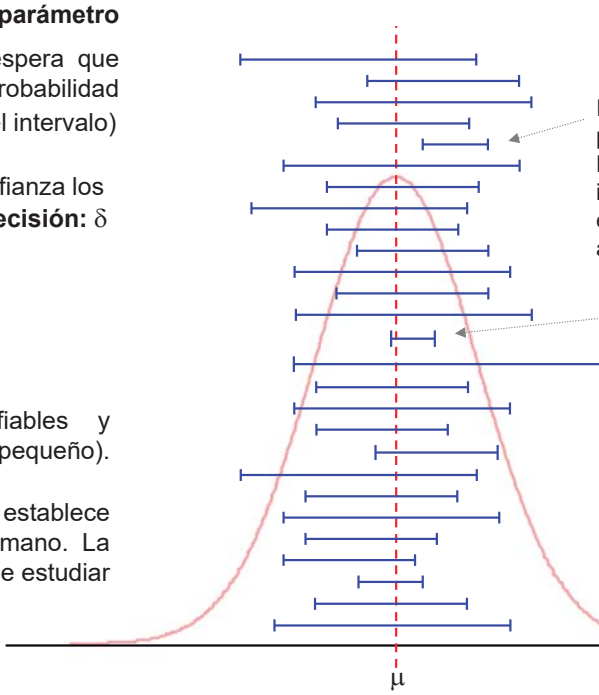
Para un IC dado, se espera que esto ocurra con una probabilidad de $1-\alpha$ (la **confianza** del intervalo)

Además del nivel de confianza los IC tienen un **nivel de precisión:** δ (su radio)



Interesan intervalos fiables y precisos (estrechos → δ pequeño).

El nivel de confianza lo establece el investigador de antemano. La precisión no → habrá que estudiar de qué depende



Intervalo que **no contiene** al parámetro poblacional
En promedio, por cada 100 intervalos, el α% de ellos no contienen al parámetro (por azar)

Intervalo **muy preciso** (estrecho)

Intervalo **poco preciso** (ancho)

Todos los IC están elaborados al mismo nivel de confianza $1-\alpha$

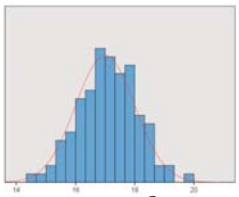
Estimación por intervalo de confianza (IC)

Intervalos de confianza a estudiar:

- IC para la media μ de una variable cuantitativa X

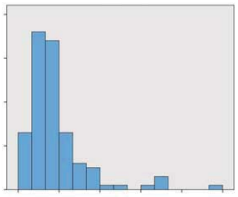
  Distinguimos dos casos:

  - X es una VA con distribución normal → $X \to N(\mu;\sigma)$

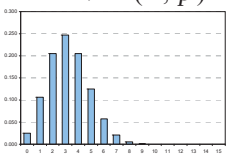
    ¿ μ ?

  - X es una VA con distribución desconocida → $X \not\to N(\mu;\sigma)$

    ¿ μ ?

- IC para el parámetro p asociado a la distribución binomial (x es ahora una VAD) → $X \to B(n;p)$

  ¿ p ?

Los métodos estadísticos permiten construir IC para cualquier parámetro (en el adelante se verá alguno más). Siempre que se hable de **estimar** un parámetro debe pensarse en **obtener su IC** (no basta con el estimador puntual)

Pedro Femia
pfemia@ugr.es

1. ¿Qué se puede decir de la **normalidad** de la variable estudiada X ?

- La variable X tiene distribución normal → Su media $\overline{x}$ tiene entonces distribución normal (lo que sigue vale siempre)
- La variable X no tiene distribución normal → Su media $\overline{x}$ tiene distribución normal si $n \geq 60$
  - Si $n \geq 60$ → lo que sigue vale de forma aproximada (mejor cuanto mayor sea n)
  - Si $n < 60$ → no se puede estimar μ con los métodos que se exponen aquí

2. **Información muestral**:

- n (tamaño muestral)
- $\overline{x}$ (media muestral)
- s (deviación típica muestral) ← Observe el detalle de que s es el estimador puntual de σ, que normalmente es un parámetro desconocido (si no se conoce μ, es muy raro conocer σ)

3. **Nivel de confianza** $1-\alpha$

Lo fija el investigador de antemano. NO depende del tamaño muestral

En general se establece al 95% (α=0.05). Otros valores habituales son el 90% (α=0.10) y el 99% (α=0.01)

4. Expresión del **IC** para estimar μ

$$IC(\mu) = \overline{x} \pm t_{\alpha;n-1} \frac{s}{\sqrt{n}}$$

Equivalentemente

$$P\left(\overline{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n}} \right) = 1-\alpha$$

Condición de **validez**

Si X normal, vale siempre, si no es normal solo es válido como aproximación si $n \geq 60$ (la aproximación que mejora cuanto mayor sea n)

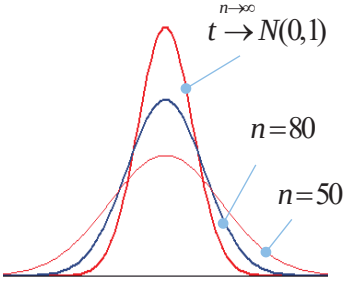
¿Quién es $t_{\alpha;n-1}$?

La distribución t-student está vinculada a la distribución normal. Se trata de una normal corregida por el hecho de no conocer σ y utilizar s (su estimador puntual) en su lugar.

El efecto de la corrección es tanto mayor cuanto menor sea el tamaño de muestra. Para muestras grandes, la t y la normal coinciden

n-1 son los grados de libertad de la distribución

Para un α prefijado ¿quién es mayor z_a de la normal o t_α de la t-student?



Distribución t de student

Pedro Femia
pfemia@ugr.es

Distribución t de student

¿Cómo obtener un valor $t_{\alpha;n-1}$?

Supongamos una muestra de 10 observaciones:

$$n=10 \quad \Rightarrow \quad n-1=9=g.l.$$

Para una confianza del 95% :

$$(1-\alpha)=0.95$$

Entonces

$$\alpha = 0.05$$

De aquí se obtiene

$$t_{0.05;9}=2.262$$

Si no aparecen los gl que buscamos, lo correcto es interpolar, pero aquí bastará con tomar los anteriores de la tabla (nunca más de los que tengamos)

$n \to \infty \Rightarrow t = N(0,1)$

g.l \ α	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.929
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
35	0.682	0.852	1.052	1.306	1.690	2.030	2.438	2.724	3.592
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.705	3.551
45	0.680	0.850	1.049	1.301	1.679	2.014	2.412	2.690	3.521
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.497
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.461
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.417
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.391
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Intervalos de confianza para la media de una variable cuantitativa

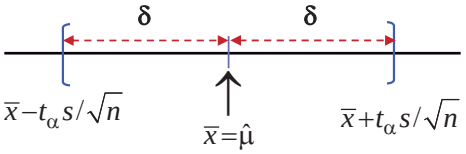
Fijar α implica, para un tamaño de muestra dado, fijar $t_{\alpha;n-1}$

Si $\overline{x} \pm t_{\alpha;n-1}\dfrac{s}{\sqrt{n}}$ es el intervalo para **μ**, $\delta = t_{\alpha;n-1}\dfrac{s}{\sqrt{n}}$ es su **precisión**

De qué depende la precisión

Si $\alpha \uparrow \quad \Rightarrow t_\alpha \downarrow \quad \Rightarrow \delta \downarrow$

→ Si aumenta el error, **aumenta** la precisión



Si $n \uparrow \quad \Rightarrow t_\alpha \downarrow$ y $\delta \downarrow \quad \Rightarrow \delta \downarrow$

→ Si aumenta el tamaño de muestra **aumenta** la precisión

→ Por otra parte, cuanto mayor sea la variabilidad, menor será la precisión (esto no es controlable)

5. Como garantizar una precisión $\delta_{deseada}$

Es necesario disponer de información previa = **Muestra piloto** que proporcione una estimación de la variabilidad que se puede encontrar en la población

Tamaño mínimo de muestra para estimar μ: $n \geq \left(\dfrac{t_{\alpha;n_{piloto}-1}\, s_{piloto}}{\delta_{deseada}} \right)^2$

Como la información dada por la muestra piloto es una variable aleatoria (s) la fórmula es una aproximación cuyo resultado se debe comprobar después

Pedro Femia
pfemia@ugr.es

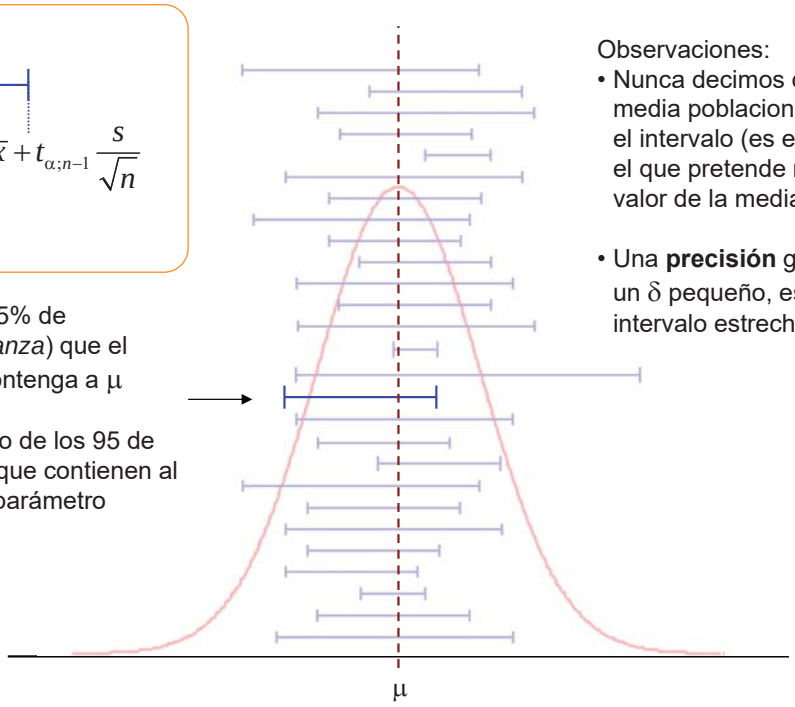
Intervalos de confianza para la media de una variable cuantitativa

Intervalos construidos a nivel de confianza del 95% $\rightarrow P(a \le \mu \le b) = 1 - \alpha = 0.95$

$$\bar{x} - t_{\alpha;n-1}\frac{s}{\sqrt{n}} \qquad \bar{x} \qquad \bar{x} + t_{\alpha;n-1}\frac{s}{\sqrt{n}}$$

Se espera, con un 95% de probabilidad (=*confianza*) que el intervalo obtenido contenga a μ

Es decir, que sea uno de los 95 de cada 100 intervalos que contienen al verdadero valor del parámetro poblacional



Observaciones:

- Nunca decimos que la media poblacional "*cae*" en el intervalo (es el intervalo el que pretende recoger al valor de la media)
- Una **precisión** grande es un δ pequeño, es decir, un intervalo estrecho

Intervalos de confianza para la media de una variable cuantitativa

Ejemplo:

Para determinar el nivel medio de colesterol en la población de estudiantes universitarios de primer año en la UGR se tomó una muestra al azar de 10 de ellos, en la que se obtuvieron los valores siguientes (en mg/dl):

162, 176, 169, 165, 171, 169, 172, 168, 167, 175

1) Suponemos que el nivel de colesterol es una VAC con distribución normal
2) Información muestral: $n = 10; \quad \bar{x} = 169.40; \quad s = 4.30$
3) Fijamos una confianza del 95% $\Rightarrow \quad \alpha = 0.05 \quad$ y $\quad t_{0.05;9} = 2.262$
4) Entonces, con una confianza del 95% $\mu \in \bar{x} \pm t_{\alpha;n-1}\frac{s}{\sqrt{n}} \quad \Rightarrow \quad \mu \in 169.40 \pm 2.262\frac{4.30}{\sqrt{10}}$

$$\mu \in 169.4 \pm 3.08 \text{ mg/dl} \quad \Rightarrow \quad \mu \in (166.32; 172.48) \text{ mg/dl}$$

5) lo que **se interpreta** diciendo que el nivel medio de colesterol en los estudiantes de primer curso de la UGR debe de ser un valor comprendido entre 166.32 y 172.48 mg/dl con una probabilidad o nivel de confianza del 95%. La precisión de la estimación es de 3.08 mg/dl

¿Qué hay que hacer para tener una precisión de 5 mg/dl? → ¡Nada! La precisión obtenida es mejor (mayor)

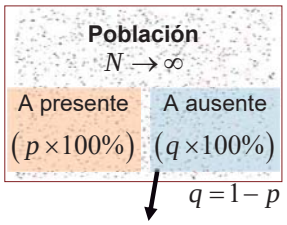
¿Qué hay que hacer para tener una precisión de 1 mg/dl? → Aumentar el tamaño de muestra. Veamos a cuanto:

$$n \ge \left(\frac{t_{\alpha;n_{piloto}-1}\, s_{piloto}}{\delta_{deseada}}\right)^2 = \left(\frac{2.262\ 4.30}{1.0}\right)^2 \rightarrow 95$$

Se necesitarían 85 individuos más. Esta fórmula es aproximada. Una vez aumentada la muestra habría que comprobar que la precisión es la deseada

Pedro Femia
pfemia@ugr.es

Recordemos:
Modelo Binomial



Muestra de tamaño n $(n/N \le 0.10)$

X = nº de observaciones en la muestra con A presente

$\Downarrow$

$X \to B(n, p)$

Pero, ¿cuánto vale p?

Información muestral:

$$\begin{cases} \text{Tamaño de muestra} = n, \text{ de los cuales} \\ x \text{ casos cumplen con la característica de interés} \\ n - x \text{ casos no cumplen con la característica de interés} \end{cases}$$

el **estimador puntual** de la proporción de casos que verifican la característica de interés es

$$\hat{p} = x / n$$

¿Qué relación tiene esta expresión con la frecuencia relativa?

Obviamente $\hat{q} = 1 - \hat{p} = (n - x)/n$ estima puntualmente a la proporción complementaria.

En estos problemas, el uso de la distribución binomial puede resultar complicado (se habla de **métodos exactos**). Lo habitual es que si x y n-x son valores suficientemente grandes se construya el intervalo de confianza para p utilizando la distribución normal (se habla de **métodos aproximados** o **asintóticos**)

Fijado α de antemano consideramos tres métodos (en orden de preferencia)

• Método de Wilson: Si $x > 5$ y $(n - x) > 5$ ← Cuando se pueda usar es el mejor

$$p \in \frac{1}{n + z_\alpha^2}\left((x \pm 0.5) + \frac{z_\alpha^2}{2} \pm z_\alpha \sqrt{\frac{z_\alpha^2}{4} + (x \pm 0.5)\left(1 - \frac{x \pm 0.5}{n}\right)} \right)$$

• Wald ajustado: Válido siempre, pero es preferible el de Wilson cuando se puede aplicar

$$\mathrm{p} \in \frac{1}{n+4}\left[(x+2) \pm \left(z_\alpha \sqrt{\frac{(x+2)(n-x+2)}{n+4}} \right) \right]$$

• Método de Wald: Si $x > 20$ y $(n - x) > 20$

$$p \in \hat{p} \pm \left(z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{1}{2n}} \right)$$

← Cuando se dan estas condiciones es una aproximación que funciona tanto mejor cuanto mayores sean x y n-x. Es el peor de los tres expuestos y uno de los que más se usan tradicionalmente

Si hay **condiciones de validez** deben ser comprobarlas **siempre**

Siempre se debe indicar cuál ha sido el método utilizado

En estas expresiones z_α es el valor de la distribución normal correspondiente al (1-α)% de confianza (observe que ahora no hablamos de grados de libertad).

Pedro Femia
pfemia@ugr.es

Distribución normal estándar

¿Cómo buscar un valor Z_α?

Para una confianza del 95% :

$$(1-\alpha)=0.95$$

Entonces

$$\alpha=0.05$$

Así que tenemos que

$$z_{0.05}=1.96$$

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.751	1.695
0.1	1.645	1.598	1.555	1.514	1.476	1.440	1.405	1.372	1.341	1.311
0.2	1.282	1.254	1.227	1.200	1.175	1.150	1.126	1.103	1.080	1.058
0.3	1.036	1.015	0.994	0.974	0.954	0.935	0.915	0.896	0.878	0.860
0.4	0.842	0.824	0.806	0.789	0.772	0.755	0.739	0.722	0.706	0.690
0.5	0.674	0.659	0.643	0.628	0.613	0.598	0.583	0.568	0.553	0.539
0.6	0.524	0.510	0.496	0.482	0.468	0.454	0.440	0.426	0.412	0.399
0.7	0.385	0.372	0.358	0.345	0.332	0.319	0.305	0.292	0.279	0.266
0.8	0.253	0.240	0.228	0.215	0.202	0.189	0.176	0.164	0.151	0.138
0.9	0.126	0.113	0.100	0.088	0.075	0.063	0.050	0.038	0.025	0.013

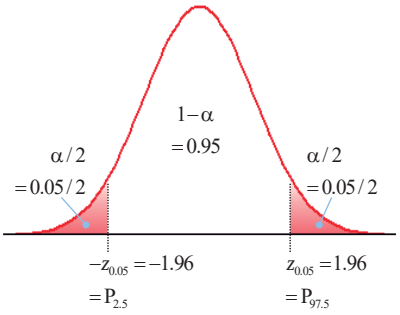
Tabla para los pequeños valores de α

α	0.002	0.001	0.000 1	0.000 01	0.000 001	0.000 000 1
z_α	3.090	3.291	3.891	4.417	4.892	5.327

$$\alpha=0.05=0.0+0.05 \Rightarrow Z_{0.05}=1.96$$

Para buscar un valor de α con dos decimales se debe sumar la primera columna (primer decimal) y la primera fila (segundo decimal)

Observe que z_α es el percentil 1-α/2 de la distribución:

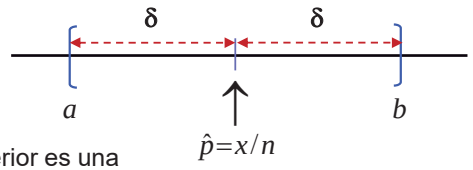


Intervalo de confianza para la proporción binomial

Precisión del intervalo

Obtenido el intervalo, que tendrá la forma (*a, b*), la precisión de la estimación puede obtenerse como sigue si el intervalo es simétrico (métodos de Wald y Wald ajustado):

$$\delta=\frac{(b-a)}{2}$$



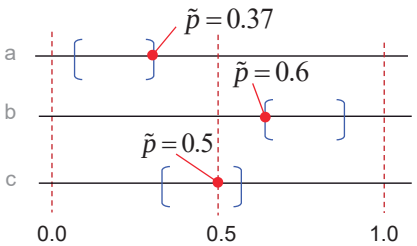
Observación: El IC de Wilson no es simétrico y el valor anterior es una aproximación. En cualquier caso, la precisión es un valor expresable porcentualmente.

Tamaño de muestra

El **tamaño mínimo de muestra** para obtener una precisión deseada δ, para α prefijado, puede determinarse de forma general o considerando la información dada por una muestra piloto:

- Sin información previa (tamaño necesario en el *peor* de los casos): $n \geq \frac{z_\alpha^2 \times 0.25}{\delta^2_{deseada}}$
- Con información previa (muestra piloto): $n \geq \frac{z_\alpha^2 \times \tilde{p}(1-\tilde{p})}{\delta^2_{deseada}}$

La expresión sin información es exacta por no depender de ninguna muestra. El precio a pagar es que normalmente da lugar a un *n* muy grande. Si se usa la información de una muestra piloto, la estimación de *n* no es exacta y hay que comprobar que se obtiene la precisión deseada. **Si el IC contiene a 0.5, las dos expresiones coinciden,** y la información no aporta ninguna ventaja para reducir *n*



$\tilde{p}$ es el valor del IC *mas próximo* a 0.5

Pedro Femia
pfemia@ugr.es

Ejemplo:
De 150 estudiantes de primer curso de grado universitario que se han encuestado, 33 fuman habitualmente. ¿Cuál es la prevalencia del tabaquismo en este colectivo?. Se desea obtener la estimación con una precisión del 10%

1) Valores muestrales: $n=150;\quad x=33;\quad n-x=150-33=117$

La estimación puntual la prevalencia del tabaquismo vendrá dada por p

$$\hat{p}=x/n=33/150=0.220\quad(22.0\%)$$

De modo que el porcentaje de no fumadores esta estimado por

$$\hat{q}=1-\hat{p}=0.780\quad(=78.0\%)$$

2) Fijamos el nivel de confianza al $95\%\ \Rightarrow\ \alpha=0.05$ y $Z_{0.05}=1.96$

3) Validez y elección del método: Como x=33 y n-x =150 son mayores a 5, se puede utilizar el método de Wilson

$$p\in\frac{1}{150+1.96^2}\left((33\pm0.5)+\frac{1.96^2}{2}\pm1.96\sqrt{\frac{1.96^2}{4}+(33\pm0.5)\left(1-\frac{33\pm0.5}{150}\right)}\right)$$

Límite inferior: $$\frac{1}{150+1.96^2}\left((33-0.5)+\frac{1.96^2}{2}-1.96\sqrt{\frac{1.96^2}{4}+(33-0.5)\left(1-\frac{33-0.5}{150}\right)}\right)=0.15826$$

Límite superior: $$\frac{1}{150+1.96^2}\left((33+0.5)+\frac{1.96^2}{2}+1.96\sqrt{\frac{1.96^2}{4}+(33+0.5)\left(1-\frac{33+0.5}{150}\right)}\right)=0.29642$$

Solución por los otros métodos

$95\%\text{-IC}_{\text{Wald ajust}}(p)=(16.1\%;29.3\%)$

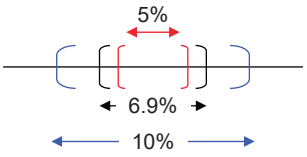
$95\%\text{-IC}_{\text{Wald}}(p)=(15.0\%;29.0\%)$

Ejemplo (continuación):

$$95\%\text{-IC}_{\text{Wilson}}(p)=(15.8\%;29.6\%)$$

4) Interpretación de los resultados: La prevalencia del tabaquismo entre los estudiantes de primer curso universitario es, con un 95% de probabilidad, un valor que debe estar comprendido entre el 15.8% y el 29.6%. La precisión de la estimación es del 6.9% ← (0.29642-0.15826)/2

5) Respecto a la precisión pedida del δ=10%, nuestros resultados tienen una precisión **mayor**, (6.9% es mayor precisión que 10%) por lo tanto no hay que considerar ningún aumento del tamaño de muestra. Si quisiéramos una precisión del 5%, entonces si que habría que considerar un aumento del tamaño de muestra



Sin considerar la información previa:

$$n\geq\frac{z_\alpha^2\times0.25}{\delta^2_{deseada}}=\frac{1.96^2\times0.25}{0.05^2_{deseada}}\rightarrow385$$

Considerando la información previa: $\begin{cases}\tilde{p}=0.296\\1-\tilde{p}=0.704\end{cases}$ $\tilde{p}$ Es el valor del IC=(0.158; 0.296) más próximo a 0.5

$$n\geq\frac{z_\alpha^2\times\tilde{p}(1-\tilde{p})}{\delta^2_{deseada}}=\frac{1.96^2\times0.296(0.704)}{0.05^2}\rightarrow317$$

Sin aprovechar la información dada por la muestra piloto

Pedro Femia
pfemia@ugr.es

Un posible esquema de trabajo

Identifique claramente **qué parámetro** se trata de estimar.

Identifique claramente la **información muestral** y el **estimador puntual** correspondiente. Proporcione la estimación puntual (e interprétela).

Establezca el **nivel de confianza** $(1-\alpha)$ (si no hay motivos para hacer otra cosa consideramos el 95%)

Compruebe si se cumplen las **condiciones de validez** que pudiera haber, por ejemplo, al estimar una media, la normalidad de la variable o si no es normal que $n>60$. Al estimar una proporción que x y n-x son mayores al valor que es requisito de la fórmula que se vaya a utilizar

Escriba la expresión del intervalo de confianza y **realice los cálculos** necesarios para obtenerlo.

Interprete los resultados obtenidos, indicando claramente
- cuál es el **intervalo** (no olvide las **unidades de medida**)
- cuál es la **precisión** obtenida (también tiene unidades de medida)
- el nivel de **confianza** con que se ha construido.

Si se trata de obtener una precisión dada de antemano $\delta_{deseada}$ compruebe si la precisión obtenida es mayor: $d<\delta_{deseada}$ de ser así, no hay que hacer nada más (la muestra actual es suficiente). En caso contrario (si $d \geq \delta_{deseada}$) se debe determinar el tamaño de muestra necesario para cumplir con esta especificación

Ejercicios

Estimación de medias (suponer que la v.a. es normal)

Muestra	Datos										n	media	desv.	t	Lim inf	Lim. Sup	precisión Observada	para δ =		n =
M1	2.6	2.4	2.7	3.2	3.4	3.5	3.2	3.6	3.4	3.5	10	3.150	0.428	2.262	2.844	3.456	0.306	0.2	23.4	24
M2	12	21	15	11	13	12	18				7	14.571	3.690	2.447	11.158	17.984	3.413	1.7	28.2	29
M3	3.75	7.75	3.88	3.67	9.2	0.77	8.61	2.46	2.68		9	4.752	3.000	2.306	2.446	7.058	2.306	1.2	33.2	34
M4	75.53	58.7	52.32	18.05	23.41	71.21	62.85	46.2	21.58	0.38	10	43.023	25.546	2.262	24.749	61.297	18.274	9.1	40.3	41
M5	6.93	6.34	4.69	7.79	3.26	0.98	8.86	5.29	8.35		9	5.832	2.563	2.306	3.862	7.803	1.970	1.0	34.9	35
M6	0.59	0.28	0.46	0.37	0.23	0.11	0.27	0.6			8	0.364	0.175	2.365	0.217	0.510	0.146	0.1	17.1	18
M7	0.54	0.98	0.74	0.69	0.2	0.98	0.51	0.54			8	0.648	0.260	2.365	0.430	0.865	0.217	0.1	37.8	38
M8	1	2	3.6	1.4	8.1	1.3	7.6				7	3.571	3.047	2.447	0.754	6.389	2.818	1.4	28.4	29
M9	0.04	0.08	0.4	0.17	0.21	0.86	0.38	0.96			8	0.388	0.348	2.365	0.097	0.678	0.291	0.1	67.5	68
M10	3.49	64.5	19.08	65.67	8.46	52.36	27.6	37.3	11.44		9	32.212	24.002	2.306	13.763	50.662	18.450	9.2	36.2	37
M11	0.31	0.92	0.29	0.91	0.94	0.55	0.64	0.63			8	0.649	0.262	2.365	0.430	0.868	0.219	0.1	38.4	39
M12	0.73	0.61	0.82	0.11	0.51	0.57	0.69				7	0.577	0.231	2.447	0.364	0.790	0.213	0.1	31.8	32

Estimación de proporciones

Ej:	Muestra n	x	Método	Validez	IC(-)	IC(+)	d	% IC(-)	% IC(+)	% d	Precisión deseada (%)	Tamaño necesario Con inf	Tamaño necesario Sin inf
1	113	42	Wilson	Si	0.284	0.468	0.092	28.41%	46.81%	9.20%	5%	383	385
			Wald	Si	0.278	0.465	0.094	27.82%	46.52%	9.35%			
			Wald Ajustado	Si	0.288	0.464	0.088	28.83%	46.38%	8.78%			
2	11	6	Wilson	No	0.246	0.819	0.286	24.56%	81.86%	28.65%	7%	196	196
			Wald	No	0.206	0.885	0.340	20.57%	88.52%	33.97%			
			Wald Ajustado	Si	0.281	0.786	0.252	28.09%	78.58%	25.25%			
3	30	24	Wilson	Si	0.609	0.916	0.154	60.87%	91.60%	15.36%	7%	187	196
			Wald	No	0.640	0.960	0.160	64.02%	95.98%	15.98%			
			Wald Ajustado	Si	0.622	0.907	0.143	62.21%	90.73%	14.26%			
4	150	103	Wilson	Si	0.605	0.758	0.077	60.51%	75.85%	7.67%	10%	92	97
			Wald	Si	0.609	0.764	0.078	60.91%	76.42%	7.76%	No es necesario aumentar n		
			Wald Ajustado	Si	0.608	0.755	0.074	60.83%	75.54%	7.36%			
5	63	32	Wilson	Si	0.380	0.635	0.127	38.01%	63.48%	12.73%	3%	1068	1068
			Wald	Si	0.377	0.639	0.131	37.65%	63.93%	13.14%			
			Wald Ajustado	Si	0.388	0.627	0.120	38.78%	62.72%	11.97%			

Pedro Femia
pfemia@ugr.es

ESTADÍSTICA

Grado en Enfermería

Tema IV

Introducción al contraste de hipótesis

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Fundamentos de los test de hipótesis

Hipótesis estadísticas

Un **contraste** o **test de hipótesis** es un procedimiento propio de la **Inferencia Estadística** encaminado a tomar una **decisión** a favor o en contra de una determinada **hipótesis estadística** que afecta a la población bajo estudio en términos de la información muestral.

Las **hipótesis estadísticas** son enunciados que se construyen en términos de los **parámetros que caracterizan a la población**. Se tratará de comprobar si el supuesto considerado por la hipótesis acerca de esos parámetros es **compatible** con la **información empírica** (la dada por la o las muestras tomadas a tal efecto)

Las hipótesis estadísticas son siempre dos: la hipótesis nula (H_0) y su alternativa (H_1). La forma que toma cada una no depende del interés particular del investigador.

La hipótesis nula H_0
- Es lo que asumimos como cierto hasta que se demuestre lo contrario
- Su elaboración siempre se hace sobre la idea de *homogeneidad*, *igualdad*, *independencia*

La hipótesis alternativa H_1
- Es la negación de la hipótesis nula
- Es la hipótesis que aceptamos si la evidencia experimental nos lleva a rechazar la hipótesis nula
- Su elaboración siempre se hace bajo la idea de *heterogeneidad*, *diferencia*, *asociación*

Algunos ejemplos:

Prueba de normalidad	Tests para un parámetro		Tests de homogeneidad entre dos muestras (estudios comparativos)		
$H_0: X \to N(\mu,\sigma)$ $H_1: X \not\to N(\mu,\sigma)$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: p = p_0$ $H_1: p \neq p_0$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$
Estudios con una muestra			Estudios con dos muestras		

Pedro Femia
pfemia@ugr.es

Hipótesis estadísticas

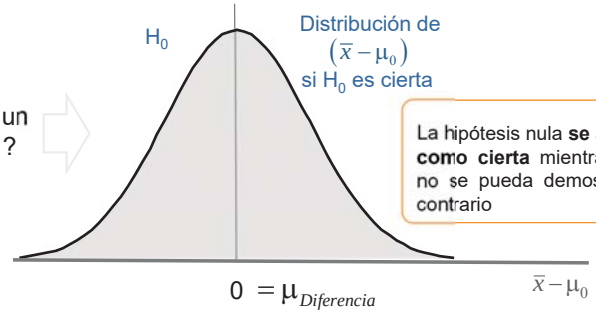
Desarrollaremos la teoría de los test de hipótesis sobre un caso concreto:

Test para una media μ de una VA normal

Premisa: asumimos que $X \to N(\mu;\sigma)$

como consecuencia $\bar{x} \to N\left(\mu;\sigma/\sqrt{n}\right)$

Se trata de decidir si la media poblacional adopta un valor concreto μ_0: ¿es asumible que sea $\mu = \mu_0$?



La hipótesis nula **se asume como cierta** mientras que no se pueda demostrar lo contrario

Hay **dos posibilidades**:

(1) Si $\mu = \mu_0 \Rightarrow \mu - \mu_0 = \mu_{Diferencia} = 0$

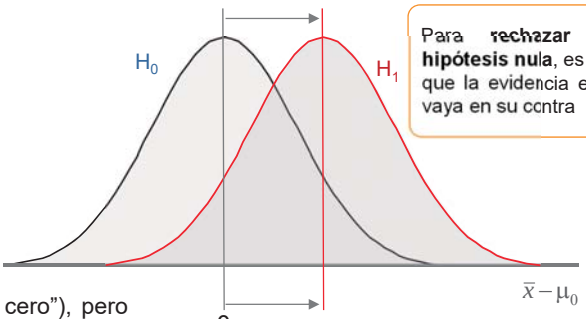
$\Rightarrow \left(\bar{X} - \mu_0\right) \to N\left(0;\sigma/\sqrt{n}\right)$

Hipótesis nula H_0: $\mu = \mu_0 \to \mu - \mu_0 = 0$

(2) Si $\mu \neq \mu_0 \Rightarrow \mu - \mu_0 = \mu_{diferencia} \neq 0$

$\Rightarrow \left(\bar{X} - \mu_0\right) \to N\left(\mu_{Diferencia};\sigma/\sqrt{n}\right)$

Hipótesis alternativa H_1: $\mu \neq \mu_0 \to \mu - \mu_0 \neq 0$



Para **rechazar a la hipótesis nula**, es preciso que la evidencia empírica vaya en su contra

Obsérvese que **H_0** *concreta algo* (la diferencia "vale cero"), pero **H_1 no** (solo dice que la diferencia "no vale cero", pero nada más), no sabemos en qué punto se centra la distribución dada por H_1 (solo que no es en el valor cero). Se trata de decidir cuál es la hipótesis a la que apoyan los datos empíricos

$\mu_{Diferencia} = \mu - \mu_0 \neq 0$

$\mu_{Diferencia} = \mu - \mu_0 = ?$

Para poder cuantificar la magnitud de la diferencia será necesario *estandarizarla* → *tipificarla*

¡repasar la tipificación y la distribución de la media muestral!

Hipótesis estadísticas

Evidencia muestral para tomar una decisión: el estadístico de contraste

1. Si la variable X tiene distribución normal — Si $X \to N(\mu;\sigma)$

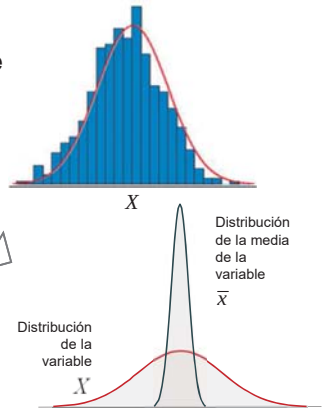
   entonces ⇓

2. Su media también tiene distribución normal $\bar{x} \to N\left(\mu;\sigma/\sqrt{n}\right)$

   ⇓

3. Tipificando: $\dfrac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} \to N(0;1)$ ⇐ $(\bar{x}-\mu) \to N\left(0;\dfrac{\sigma}{\sqrt{n}}\right)$

   Tipificación de la media muestral



Ahora interviene la **información dada por la hipótesis nula H_0**

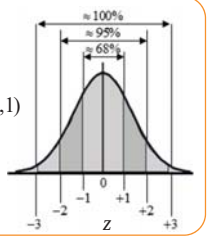
4. Si H_0 es cierta $\mu = \mu_0 \Rightarrow \dfrac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}} \to N(0;1)$ de este modo tenemos un criterio para decidir si la diferencia entre la información muestral ($\bar{x}$) y la propuesta de la hipótesis nula ($\mu = \mu_0$) son muy discrepantes o no (un valor tipificado es "*auto informativo*")

5. Solo falta eliminar al *parámetro perturbador* σ (la desviación típica poblacional, que es desconocida)

   Si sustituimos σ por su estimador puntual $\hat{\sigma} = s$ obtenemos*

$$t = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}} \to t\text{-}Student$$

$$z = \frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}} \to N(0,1)$$



Esta cantidad es una función de los valores muestrales que nos permite decidir en favor o en contra de la hipótesis nula a raíz de la evidencia muestral. Se denomina **estadístico de contraste** o **cantidad experimental.** Habitualmente se indica t_{exp} (o F_{exp}, χ^2_{exp},... según la distribución usada)

* Recordar que la distribución *t-Student* constituye una "corrección" de la distribución normal necesaria para muestras pequeñas (n<100) cuando σ es desconocido (siempre, en gral.)

Pedro Femia
pfemia@ugr.es

Evidencia muestral para tomar una decisión: el estadístico de contraste

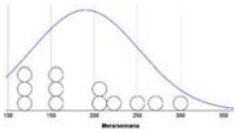
Ejemplo: Se desea comprobar si el nivel medio de actividad física semanal (medido en METs/semana) en la población de estudiantes de primer curso de la UGR puede asumirse que toma un valor concreto μ_0. **Vamos a ilustrarlo estudiando dos casos: a)** considerando μ_0=200; y **b)** considerando μ_0=230

1. **Muestra** e **información muestral**

Sujeto	1	2	3	4	5	6	7	8	9	10	11	12
METs/Semana	124	161	125	202	250	210	150	271	300	113	150	220

n	$\bar{x}$	s
12	189.76	62.019

2. **Supuestos**: asumimos que X = "METs / semana" $\rightarrow N(\mu;\sigma)$



Con muestras tan pequeñas, la normalidad de la variable es prácticamente imposible de evaluar

3. **Hipótesis** $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases} \longrightarrow$ Caso a) $\begin{cases} H_0: \mu = 200 \\ H_1: \mu \neq 200 \end{cases}$ Caso b) $\begin{cases} H_0: \mu = 230 \\ H_1: \mu \neq 230 \end{cases}$

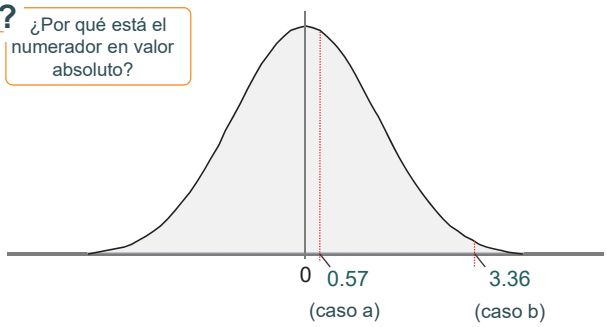
4. Obtenemos una **medida resumen** de la **información muestral** que permita **decidir** por H_0 o H_1,

el **Estadístico de Contraste** $t_{exp} = \dfrac{|\bar{x} - \mu_0|}{s/\sqrt{n}}$

? ¿Por qué está el numerador en valor absoluto?

Caso a) $t_{exp} = \dfrac{|189.76 - 200|}{62.019/\sqrt{12}} = 0.57$

Caso b) $t_{exp} = \dfrac{|189.76 - 230|}{62.019/\sqrt{12}} = 3.36$



Evidencia muestral para tomar una decisión: el estadístico de contraste

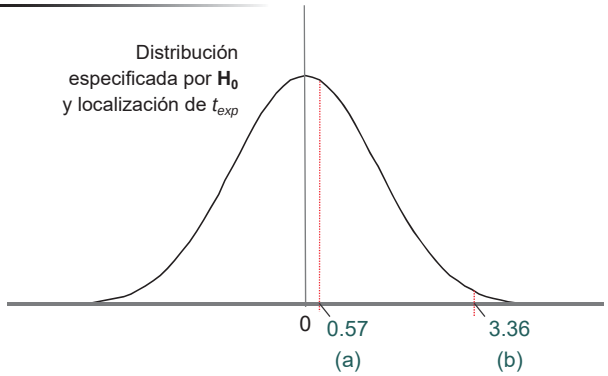
La información muestral se ha reducido a un indicador, **el estadístico de contraste**, construido suponiendo que es cierta la hipótesis nula:

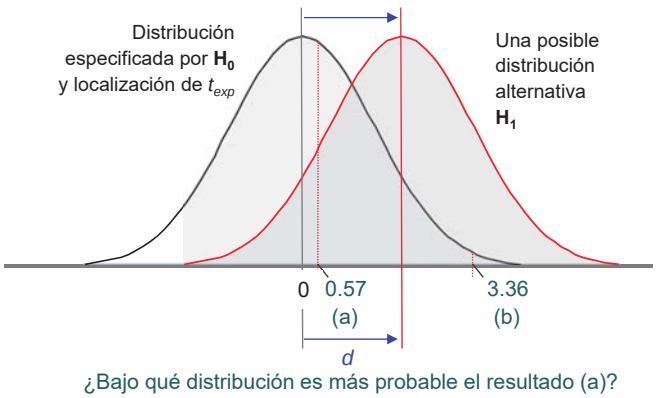
$$t_{exp} = \frac{|\hat{\mu} - \mu_0|}{\hat{\sigma}/\sqrt{n}} = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}}$$

Ahora se trata de comprobar si la información observada (el estadístico de contraste) es ***compatible*** con ella (en el sentido de ***probable*** bajo dicha distribución).

Caso a)
¿Es probable el resultado $t_{exp} = 0.57$
bajo la distribución de t_{exp} dada por H_0?

Caso b)
¿Es probable el resultado $t_{exp} = 3.36$
bajo la distribución de t_{exp} dada por H_0?





¿Bajo qué distribución es más probable el resultado (a)?
¿y el resultado (b)?
¿y si el desplazamiento *d* fuera mayor/menor?

Pedro Femia
pfemia@ugr.es

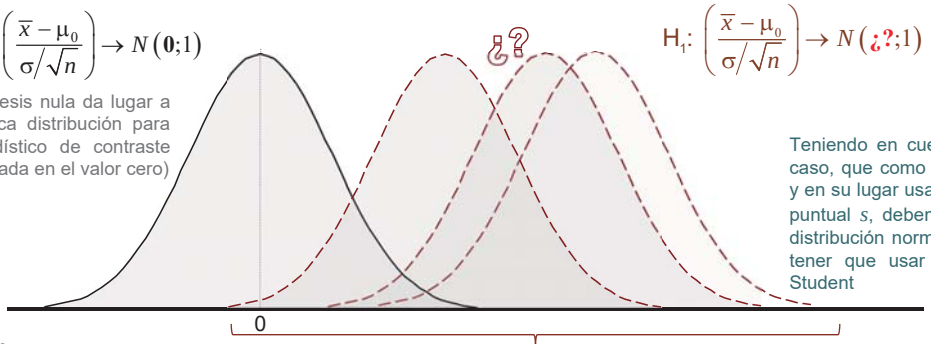
Resumen

La hipótesis alternativa no especifica una única distribución para el estadístico de contraste (¿en qué valor está centrada?)

$H_0: \left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right) \rightarrow N(0;1)$

$H_1: \left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right) \rightarrow N(¿?;1)$

La hipótesis nula da lugar a una única distribución para el estadístico de contraste (la centrada en el valor cero)



Teniendo en cuenta, en cualquier caso, que como σ es desconocida y en su lugar usamos su estimador puntual s, debemos '*corregir*' a la distribución normal, lo que supone tener que usar la distribución t-Student

ignoramos cuál es la distribución real de entre todas las posibles para H_1

En la práctica:

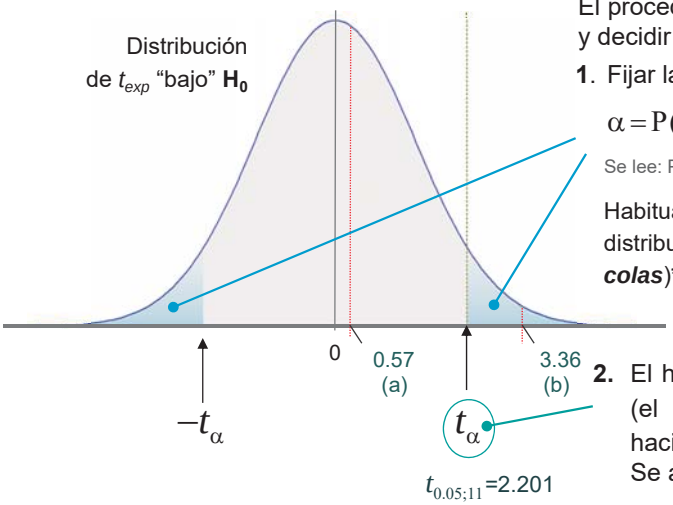
- La **hipótesis nula (H_0)** se construye de tal manera que especifica de forma precisa una distribución de probabilidad (en este caso, y casi siempre, centrada en el valor cero)
- La resolución del problema **comienza suponiendo cierta H_0**, es decir, que la muestra obtenida procede de la distribución especificada por dicha hipótesis
- La **distribución** asociada a **H_1 no se conoce de forma concreta**: ¿cuanto vale su parámetro de posición? (H_1 no establece ninguna distribución con la que se pueda trabajar). Por lo tanto, solo es posible manejar la distribución propuesta por H_0. Por eso, la identidad de H_0 no depende del interés del investigador. **H_0 es la hipótesis que proporciona unas condiciones conocidas con las que poder trabajar**
- La decisión a favor o en contra de H_0 se basará en determinar **si la muestra es probable** bajo tal distribución (se aceptará entonces H_0) o si por el contrario, constituye algo *raro* **o improbable** bajo la misma (se rechazará H_0 a favor de una H_1). Para ello, es necesario resumir la información muestral en una cantidad **el estadístico de contraste,** cuya distribución de probabilidad está especificada precisamente por H_0, si es cierta.
- Como σ es desconocido y usamos su estimador puntual, s en su lugar, en vez de la distribución normal hay que considerar la distribución *t-Student:*

$$t_{exp} = (\overline{x}-\mu_0)\Big/\left(s/\sqrt{n}\right) \rightarrow t\text{ - }Student$$

El error al rechazar H_0: Error de tipo I (α)

¿Cómo decidir?

Se debe asumir que no podemos tomar una decisión a favor en contra de H_0 de forma infalible: siempre va a haber cierto riesgo de cometer un error



El procedimiento es asumir un límite de error tolerable y decidir en consecuencia:

1. Fijar la probabilidad (α) de rechazar por error H_0

$$\alpha = P(\text{rechazar } H_0 | \text{ es cierta } H_0)$$ **= Error de tipo I**

Se lee: Probabilidad de rechazar H_0 ***cuando*** es cierta H_0

Habitualmente α=0.05 repartido en las dos colas de la distribución ($\alpha/2$=0.025 en cada una; ***test de dos colas***)*

2. El hecho de fijar α implica obtener un valor de t (el percentil $1-\alpha/2$) que deja un área de $\alpha/2$ hacia las colas de la distribución y $1-\alpha$ central. Se alude a él como t_α y se busca en las tablas.

3. El valor t_α (a menudo se le llama ***cantidad teórica***) actúa como criterio para aceptar o rechazar H_0 al comparar con él el valor de $t_{exp.}$ La decisión entonces consiste en:

$$\begin{cases} \text{Si } t_{exp} \le t_\alpha \rightarrow \text{Aceptar } H_0 \text{ (la muestra es probable bajo esta hipótesis)} \\ \text{Si } t_{exp} > t_\alpha \rightarrow \text{Rechazar } H_0 \text{ (la muestra es rara bajo esta hipótesis)} \end{cases}$$

* Se habla de test de una cola cuando el error de tipo I se concentra en una sola cola de la distribución. Se verá más adelante

Pedro Femia
pfemia@ugr.es

El error al rechazar H_0: Error de tipo I (α)

Fijado el error de tipo I (α) el conjunto de valores posibles para el estadístico de contraste que nos llevan a aceptar H_0 se denomina **región de aceptación del test** (RA). El resto de valores constituyen la **región crítica** (RC) o de **rechazo**, ya que si el estadístico toma uno de estos valores, la decisión será rechazar H_0

Es decir, fijar α se traduce en fijar estas dos regiones RA y RC. El valor t_α constituye la ***frontera*** entre ambas

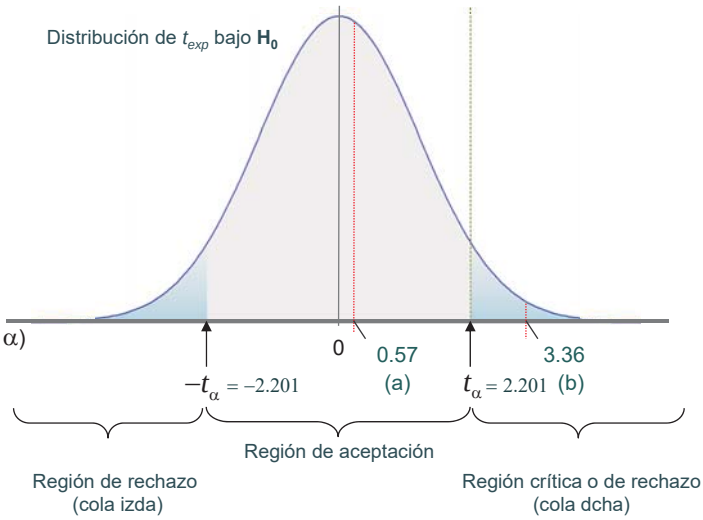
Normalmente, de los estadísticos de contraste interesa su magnitud (no su signo) por tanto, por simetría, toda la atención se puede centrar en lo que ocurre en la cola derecha de la distribución.

$t_{exp} \leq t_\alpha \rightarrow$ Aceptación de H_0 a nivel α

- El estadístico de contraste toma un valor que pertenece a la región de aceptación
- La información muestral es compatible con la H_0
- La hipótesis nula explica los datos observados
- La diferencia con μ_0 **no es significativa** (a nivel α)

$t_{exp} > t_\alpha \rightarrow$ Rechazo de H_0 a nivel α

- La H_0 no explica a los datos observados
- Los datos observados son improbables si es cierta la H_0
- La diferencia con μ_0 **es significativa** (a nivel α)



Significación estadística ≡ rechazo de H_0

El error al aceptar H_0: error de tipo II (β); y el acierto al rechazarla: la potencia (θ)

El error de tipo I (con probabilidad α) no es el único error que se puede cometer. ¿Qué ocurre si H_0 no es correcta pero el estadístico de contraste toma un valor en la región de aceptación, es decir, si $t_{exp} < t_\alpha$?

El **error de tipo II** o **error β** es aquel que se comete cuando se decide aceptar de forma incorrecta H_0

$$\beta = P(\text{aceptar } H_0 | \text{ no es cierta } H_0)$$

Se lee: Probabilidad de aceptar H_0 ***cuando*** no es cierta H_0

Observe que β es un área bajo la distribución dada por H_1 no por H_0

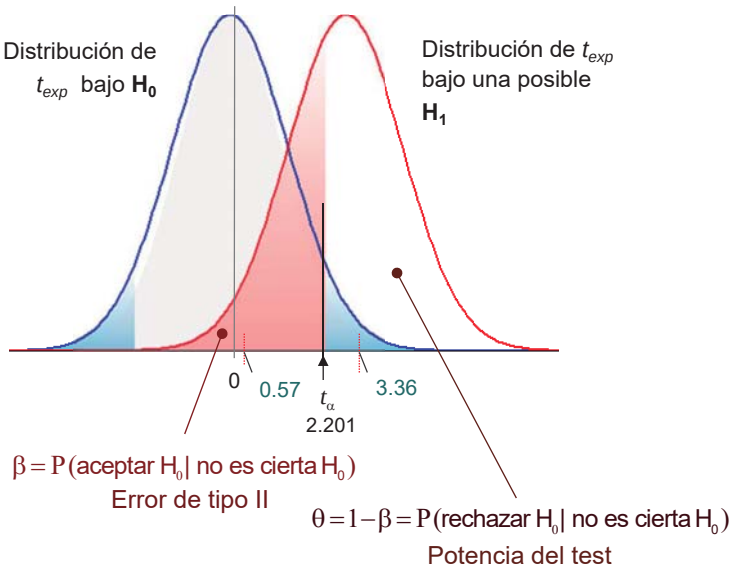
Como esa distribución es en principio desconocida, entonces ¡tampoco conocemos β!

El área bajo H_1 dada por

$$\theta = 1 - \beta$$

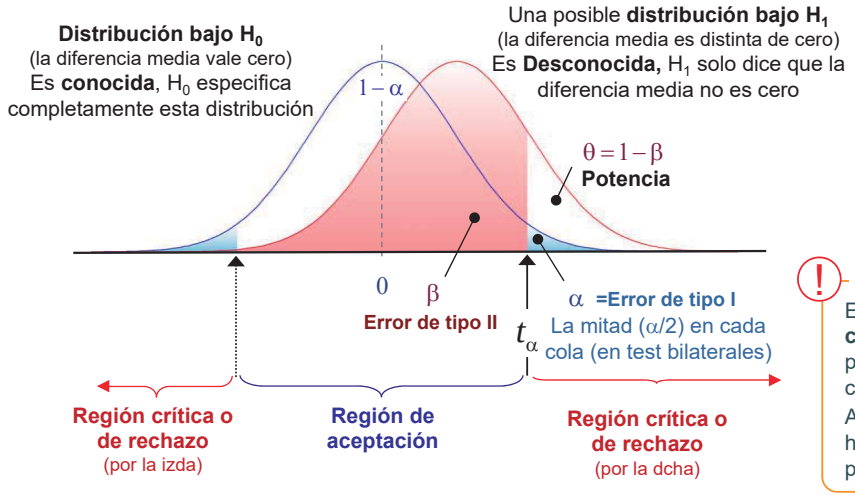
es de gran interés: representa la probabilidad de acertar al rechazar H_0 y se llama **potencia del test.**

Al igual que el error de tipo II, la potencia es un área bajo una distribución desconocida y por tanto tampoco sabemos su magnitud



Pedro Femia
pfemia@ugr.es

Resumen



Decisión por	Es cierta H_0	Es cierta H_1
H_0	1-α (*acierto*)	β (*error*)
H_1	α (*error*)	1−β=θ (*acierto*)

¡Cuidado!
En los test de hipótesis **no existe el concepto de confianza** (característica propia de los de los Intervalos de confianza)
Aquí 1-α no recibe ningún nombre, de hecho no es algo que interese en la práctica

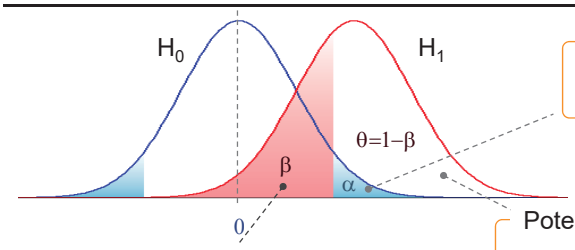
Conceptos fundamentales:

- **Estadístico de contraste** o **cantidad experimental:** t_{exp} = resumen de la información muestral que permite tomar la decisión a favor (por tomar un valor *probable*) o en contra (por tomar un valor *improbable*) de H_0
- **Región de aceptación** y **región crítica o de rechazo** = conjunto de valores del estadístico de contraste que nos llevan a aceptar y rechazar respectivamente H_0
- **Error de tipo I:** α = probabilidad de equivocarse al **rechazar** H_0
- **Error de tipo II:** β = probabilidad de equivocarse al **aceptar** H_0
- **Potencia** θ (=1-β) probabilidad de acertar al **rechazar** H_0

← Fijado por el investigador
Es un área bajo la distribución (conocida) dada por H_0

← Desconocidos
Se trata de áreas bajo la distribución (desconocida) compatible con H_1

De qué dependen los errores de tipo I y II (y, por extensión, la potencia)



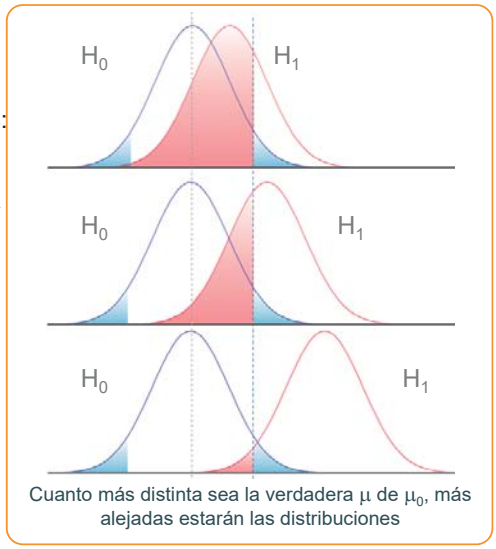
Error de tipo I
$\alpha = P(\text{rechazar } H_0 | \text{ es cierta } H_0)$
Fijado de antemano por el investigador (no depende de ni de *n* ni del fenómeno estudiado)

Potencia
$\theta = 1-\beta = P(\text{rechazar } H_0 | \text{ no es cierta } H_0)$

Error de tipo II
$\beta = P(\text{aceptar } H_0 | \text{ no es cierta } H_0)$

La potencia θ y, por lo tanto, el error β dependen de tres factores:

- Del **fenómeno** en sí, es decir, de la magnitud real de la diferencia entre μ y μ_0 (*tamaño del efecto* real)
- Del **nivel del error de tipo I** prefijado:
  si $\alpha \uparrow \Rightarrow \beta \downarrow$ y $\theta \uparrow$
- Del **tamaño de muestra** *n*:
  si $n \uparrow \Rightarrow \beta \downarrow$ y $\theta \uparrow$ ¿por qué?



Cuanto más distinta sea la verdadera μ de μ_0, más alejadas estarán las distribuciones

Pedro Femia
pfemia@ugr.es

En qué influye el tamaño de muestra

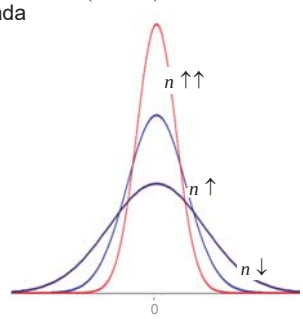
Como ya sabemos $(\bar{x}-\mu)\to N\left(0;\sigma/\sqrt{n}\right)$

tipificando, y por ser σ desconocida, teníamos

$t_{exp}=\dfrac{(\bar{x}-\mu)}{s/\sqrt{n}}\to t\text{-}Student$ y suponiendo cierta $H_0:\mu=\mu_0$

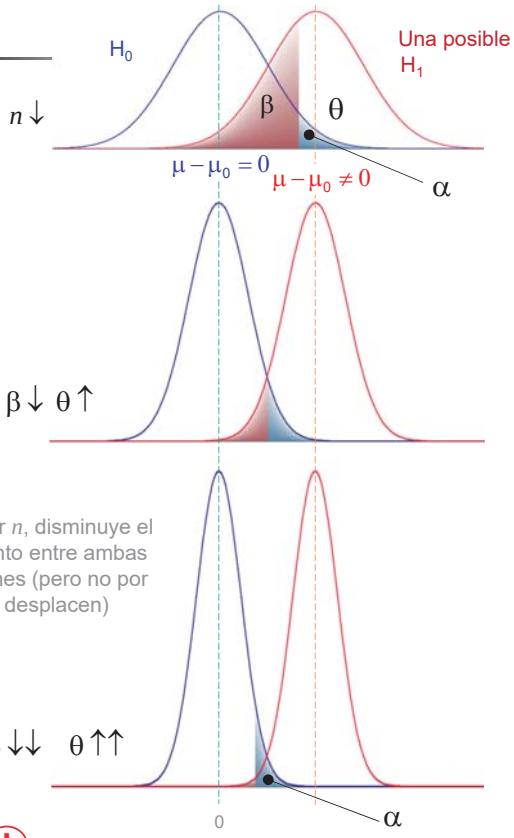
$t_{exp}=\dfrac{(\bar{x}-\mu_0)}{s/\sqrt{n}}\to t\text{-}Student$ de modo que la variabilidad de la distribución de $(\bar{x}-\mu_0)$ viene estimada por $s/\sqrt{n}$

Por lo tanto, al aumentar el tamaño muestral, la distribución de $(\bar{x}-\mu_0)$ se hace más estrecha y apuntada



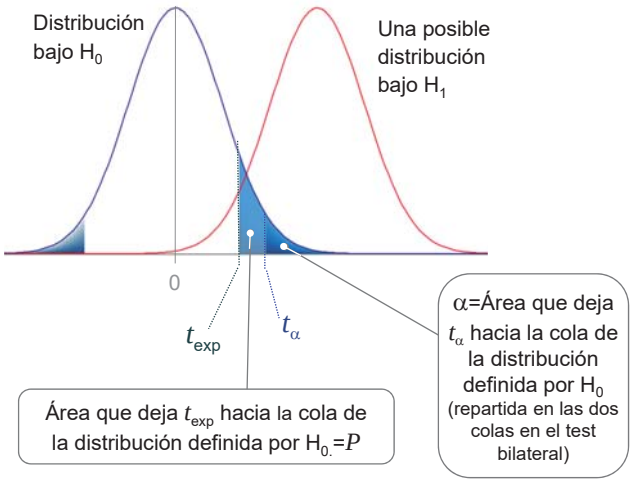
Esto se traduce en que para una misma diferencia $|\bar{x}-\mu_0|$ si n es mayor, el valor estandarizado t_{exp} será también mayor, permitiendo rechazar antes H_0.

si $n_1<n_2\Rightarrow\dfrac{|\bar{x}-\mu_0|}{s/\sqrt{n_1}}<\dfrac{|\bar{x}-\mu_0|}{s/\sqrt{n_2}}$



Recordar que ¡α NO cambia al variar n!
Se fijó de antemano y no depende de la muestra

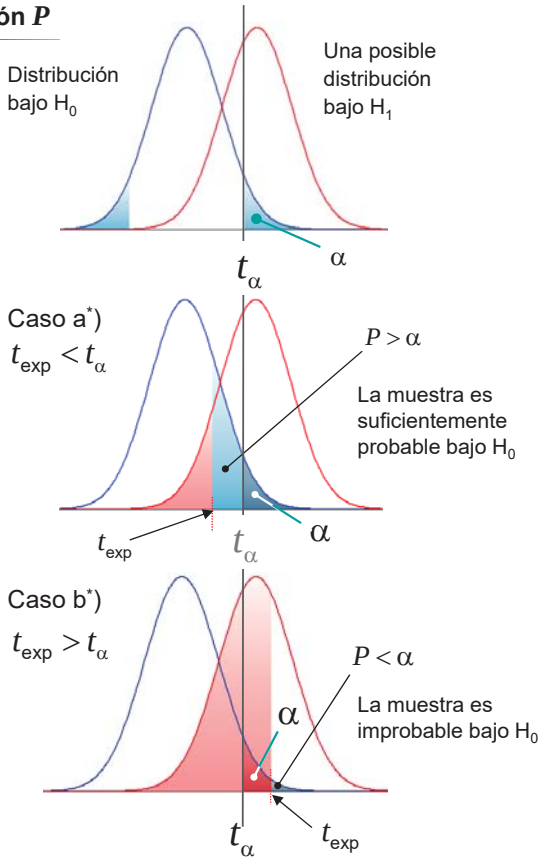
La mejor forma de decidir: el nivel mínimo de significación P



P es el *nivel mínimo de significación* y es un área bajo la distribución definida por H_0 que viene determinada por la información muestral

Comparar P con α es una alternativa mejor para decidir el resultado del test que comparar t_{exp} con t_α

- P depende de la muestra
- P responde a la misma definición que α, es un error de tipo I, solo que P lo da la muestra (α lo fija el investigador de antemano) → P viene a ser el "α_{exp}"

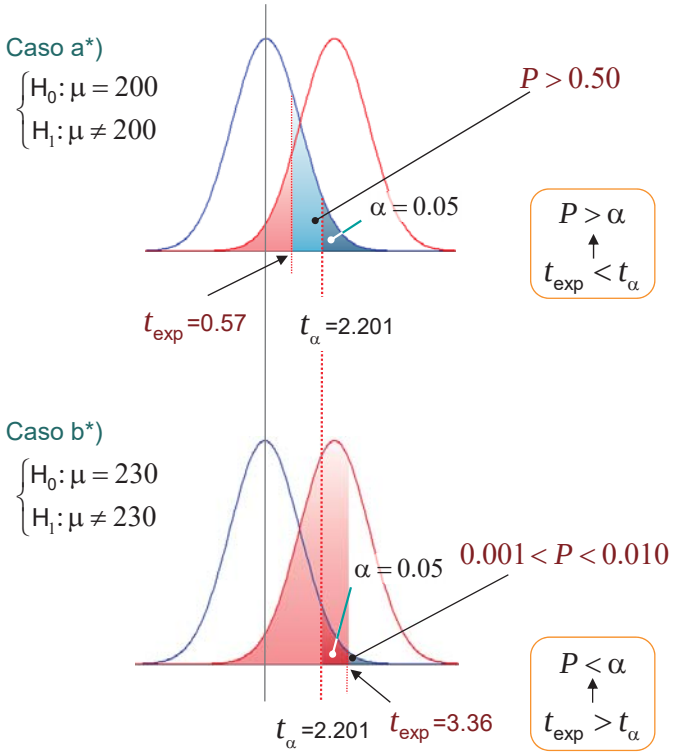


*Para simplificar, se ha representado el error α solamente en la cola dcha.

Pedro Femia
pfemia@ugr.es

Toma de decisiones. El nivel de significación P, la mejor forma de decidir

Volviendo al ejemplo del principio



*Para simplificar, se ha representado el error α solamente en la cola dcha.

Interpretación del caso (a)

Suponiendo cierta la hipótesis (nula) de que la media poblacional es de 200, la probabilidad de encontrar una muestra tan discrepante (o más) con dicha hipótesis como la actual es P>50%. Como esta probabilidad es alta, concluimos que nuestra muestra es algo que puede ocurrir con frecuencia si la H_0 es cierta, de manera que no podemos rechazarla

Interpretación del caso (b)

Suponiendo cierta la hipótesis (nula) de que la media poblacional es de 230, la probabilidad de encontrar una muestra tan discrepante (o más) con dicha hipótesis como la muestra actual es P<1%. Como esta probabilidad es muy baja, concluimos que si la hipótesis nula fuera cierta, entonces la observación de una muestra como la actual es un fenómeno raro (improbable). Si la muestra está bien tomada es improbable observar "cosas extrañas" por tanto la evidencia experimental apunta a que la H_0 no debe ser cierta y en consecuencia la rechazamos, a favor de una hipótesis alternativa que dice que la media poblacional no es 230.

Toma de decisiones

Resumen: Las dos estrategias para decidir

1. Fijar α (no depende de la muestra) → Esto permite **determinar t_α** (de las tablas de distribuciones de probabilidad)

2. Obtener t_{exp} (de la muestra)

1ª forma de decidir:

Si $t_{exp} \geq t_\alpha \rightarrow$ Rechazar H_0 para el nivel α prefijado

Si $t_{exp} < t_\alpha \rightarrow$ Aceptar H_0 para el nivel α prefijado

Esta forma de decidir debe evitarse en la medida de lo posible

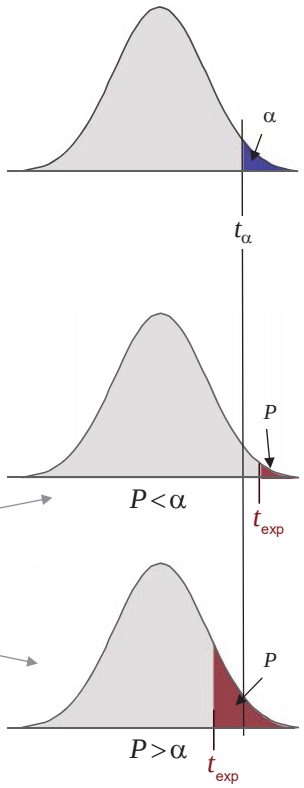
Obtener P (tablas)

2ª forma de decidir:

Si $P \leq \alpha \rightarrow$ Rechazar H_0 indicando el nivel de significación P

Si $P > \alpha \rightarrow$ Aceptar H_0 indicando el nivel de significación P

La literatura científica está repleta de niveles de significación P



Se ha representado el error α solo en una cola

Pedro Femia
pfemia@ugr.es

Nivel mínimo de significación P

Breve resumen:

De la información propuesta por H_0 y de la muestra se obtiene t_{exp} que es quien contiene toda la información respecto a la probabilidad de observar una muestra como la actual si H_0 es cierta.

El valor de t_{exp} es difícil de interpretar sin tablas de probabilidad (para empezar, depende de los g.l.) El nivel mínimo de significación P representa un modo de estandarizar esa información en forma de probabilidad (concretamente la probabilidad de un error)

Tanto α como P aluden al error de tipo I, la diferencia entre ellos es que

- α lo fija de antemano el investigador. No depende de la muestra observada, y es el nivel de error *máximo* tolerable para rechazar incorrectamente H_0
- P lo proporciona la muestra, viene a ser algo así como el "α experimental". Se puede interpretar como la probabilidad de rechazar de forma incorrecta H_0 a raíz de la información suministrada por esa muestra concreta

Cuando P toma un valor pequeño, por ejemplo P=0,001, se dice que la significación es alta. En este caso el error β asociado a la decisión tendrá un valor grande → no podemos aceptar H_0 (el error asociado a tal decisión es grande =β), mejor la rechazamos (el error asociado a tal decisión es pequeño =P)

Elección de α + valoración de P → Regla automática de decisión

Habitualmente se fija α=0.05, de modo que:

- Si $P > 0.15$ → La muestra es suficientemente probable bajo la H_0, por lo tanto no se puede rechazar dicha hipótesis
- Si $0.05 < P \leq 0.15$ → no se puede rechazar H_0 (para un α=0.05) pero hay indicios de significación. Si la muestra no es muy grande es posible que la potencia sea baja y convendría estudiar un aumento del tamaño muestral
- Si $P \leq 0.05$ → se rechaza H_0

! Cuidado con las reglas automáticas: ¿si P =0.049 rechazamos H_0 y si es P=0.051 la aceptamos?. No tiene mucho sentido. La decisión no debe recaer solo sobre P. Como veremos más adelante, los IC ayudan a tomar mejor la decisión.

Ejemplos de cómo obtener el *nivel mínimo de significación P*

Para centrar ideas **supongamos que la distribución de referencia es una *t-Student* con 10 g.l.** (tabla de la diapositiva siguiente)

En la fila de los gl implicados, se busca entre qué valores de t_α se encuentra t_{exp} de manera que P se encuentra entre los valores de α correspondientes:

1) $t_{exp}=2.40 \quad t_{0.05}=2.228 < t_{exp}=2.40 < t_{0.02}=2.764 \rightarrow 0.02 < P < 0.05 \rightarrow H_1$ (Significativo, $P < 0.050$)

2) $t_{exp}=1.21 \quad t_{0.30}=1.093 < t_{exp}=1.21 < t_{0.20}=1.372 \rightarrow 0.20 < P < 0.30 \rightarrow H_0$ (N.S. $P > 0.200$)

3) $t_{exp}=2.11 \quad t_{0.10}=1.812 < t_{exp}=2.11 < t_{0.05}=2.228 \rightarrow 0.05 < P < 0.10 \rightarrow H_0$ (Indicios de significación)
(Conviene estudiar n y la potencia)

4) $t_{exp}=6.33 \quad t_{0.001}=4.587 < t_{exp}=6.33 \rightarrow P < 0.001 \rightarrow H_1$ (Significativo, $P < 0.001$)

5) $t_{exp}=0.55 \quad t_{0.50}=0.700 > t_{exp}=0.55 \rightarrow P > 0.50 \rightarrow H_0$ (NS, $P > 0.500$)

Recordar que

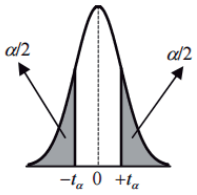
Si se **opta por H_1** se dice que el resultado es **estadísticamente significativo**
Si se **opta por H_0** se dice que el resultado es **no significativo**

En cualquier caso se acompaña tal afirmación con el **nivel *P*** obtenido (cuando P se obtiene de forma exacta, con un programa de ordenador, siempre se debe indicar con 3 decimales, o bien la expresión P<0.001 cuando sea procedente)

Pedro Femia
pfemia@ugr.es

Ejemplos de cómo obtener el *nivel mínimo de significación P*

Distribución *t*-Student



0.02 < P < 0.05

g.l. \ α	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850

t_{exp}=2.40

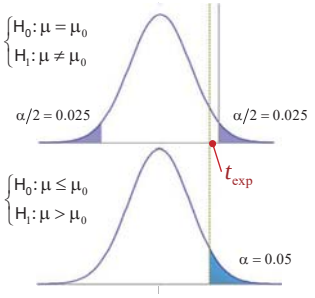
Test bilaterales y test unilaterales

El planteamiento desarrollado hasta ahora es el de un **test bilateral** o **a colas**, ya que el error se reparte entre las colas izquierda y derecha de la distribución.

Hipótesis alternativa bilateral $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$ La región crítica se reparte en las dos colas

Hipótesis alternativa unilateral $\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$ o bien $\begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$ La región crítica se concentra en una de las colas

- En H_1 aparece la desigualdad que debe ser comprobada sin duda
- En H_0 aparecen la igualdad y la desigualdad que no tiene interés demostrar



Un resultado, como el indicado con t_{exp}, puede ser significativo en un test unilateral pero no en uno bilateral; el test unilateral es *más potente*

Ejemplos

- ¿Es cierto que la duración de determinada prótesis de rodilla es de, al menos, 15 años? ⇨ $\begin{cases} H_0: \mu \geq 15 \quad (=\mu_0) \\ H_1: \mu < 15 \end{cases}$
- ¿Se ha reducido el número de cigarrillos diarios* consumidos tras una terapia de deshabituación? ⇨ $\begin{cases} H_0: \lambda_{Despues} \geq \lambda_{Antes} \\ H_1: \lambda_{Despues} < \lambda_{Antes} \end{cases}$

* X= número de cigarrillos consumidos al día sería una VA de tipo Poisson, λ es la media de dicha VA

- Al comparar *un grupo tratado* frente a un *grupo control* (← Ensayos clínicos)
  - El test bilateral contrasta tanto si el grupo tratado mejora como si empeora respecto al control (es decir, si es *distinto*)
  - El test unilateral contrasta solo si el grupo tratado mejora (o solo si empeora). A efectos prácticos, una de las desigualdades no se considera un hallazgo de interés (tiene la misma relevancia que la igualdad).

Resolución de un test unilateral

1. Formulación de las hipótesis. H_1 es la desigualdad que debe ser comprobada sin duda (la otra se incluye en la nula)
2. Comprobación de que la evidencia muestral es compatible con H_1. En caso de no serlo, se acepta la H_0 sin necesidad de hacer cálculos (por ejemplo, si el promedio muestral del número de cigarrillos consumidos al día es mayor después que antes de la terapia, $\bar{x}_{Despues} > \bar{x}_{Antes}$, la información muestral es incompatible con H_1 y no tiene sentido hacer cálculos)
3. En general, si la información muestral es compatible con H_1, los cálculos se hacen igual que si fuera un test bilateral, solo que el valor de *p* obtenido se divide por dos

Pedro Femia
pfemia@ugr.es

Conclusiones posibles

Cuando se rechaza H_0

- **La conclusión es fiable**, ya que la probabilidad de que ocurra un de tipo I, α, está controlada (prefijada) y es tan pequeña como se haya querido. Puede ser que la conclusión sea errónea, pero la (pequeña) probabilidad de esto ocurra (α) ha sido asumida de antemano.
- **No tiene sentido aumentar el tamaño de la muestra**, pues el error de tipo I (α) sigue siendo el mismo y el aumento de potencia que provoca el aumento del tamaño muestral conducirá a que lo más probable sea que nuevamente se rechace H_0

Cuando se acepta H_0

- Si no se ha estimado previamente el tamaño de muestra, la conclusión **no es fiable,** pues la probabilidad de cometer el error β no está controlada y puede haber sido grande. La **duda** razonable **es** si

  (1) realmente la muestra no contradice a H_0 (es decir, su aceptación es correcta), o bien si

  (2) lo que ocurre es que la muestra no aporta la suficiente información para ello (es decir, hay poca *potencia estadística*).

  Un **aumento del tamaño de muestra** puede ser conveniente, especialmente si el tamaño original era pequeño y el valor de p es relativamente bajo, aunque no lo suficiente como para declarar al test significativo (por ejemplo, entre el 5% y el 20%). En este caso, al aumentar n, el error β se hará más pequeño, la potencia θ aumentará y si H_0 no es cierta, será mas probable rechazarla. Hay mecanismos para estudiar la fiabilidad de la decisión a favor de H_0 basados en Intervalos de Confianza

- **Para estimar el tamaño de muestra**, es necesario establecer de antemano tanto el **nivel de α** como **la potencia deseada 1−β**. En consecuencia, al haber asumido el nivel tolerable de los dos errores posibles, tanto el rechazo como la aceptación de H_0 son fiables

Significación *estadística* y significación *sustantiva.* Fiabilidad de H_0

Tras un test de hipótesis, generalmente se debe matizar la decisión con un intervalo de confianza

Considerando el ejemplo inicial $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases} \equiv \begin{cases} H_0: \mu - \mu_0 = 0 \\ H_1: \mu - \mu_0 \neq 0 \end{cases} \rightarrow t_{exp} = \dfrac{|\overline{x} - \mu_0|}{s/\sqrt{n}}$

a) Rechazo de H_0

Como se concluye $\mu \neq \mu_0$, la pregunta procedente es ¿cuánto vale μ? $\rightarrow IC_{1-\alpha}(\mu) = \overline{x} \pm t_{\alpha,n-1}\, s/\sqrt{n}$

o ¿cuanto vale la diferencia observada respecto a μ_0? $\rightarrow IC_{1-\alpha}(\mu - \mu_0) = (\overline{x} - \mu_0) \pm t_{\alpha,n-1}\, s/\sqrt{n}$

Si la potencia es grande (n grande), el test rechazará H_0, pero se debe estudiar la *utilidad práctica* (clínica, o fisiológica, o biológica,… = ***sustantiva***) de la diferencia encontrada

b) Aceptación de H_0

Ahora es pertinente indicar a qué se está llamando "cero": $\mu - \mu_0 = 0 \rightarrow IC_{\alpha}(\mu - \mu_0) = (\overline{x} - \mu_0) \pm t_{\alpha,n-1}\, s/\sqrt{n}$

Este IC tendrá un límite negativo y otro positivo, y permite apreciar "a qué magnitud estamos llamando cero"

Por otra parte, si no se ha fijado la potencia de antemano y no se ha estimado el tamaño de muestra que permite obtenerla, el intervalo elaborado a un nivel de error 2β, es decir, con una confianza del (1−2β)% permite estudiar la fiabilidad de la decisión por H_0 (es decir, si la potencia era suficiente o no). Este intervalo sería:

$$IC_{1-2\beta}(\mu - \mu_0) = (\overline{x} - \mu_0) \pm t_{2\beta,n-1}\, s/\sqrt{n}$$

En el ejemplo:

$IC_{95\%}(\mu) = (150.35; 229.16)$

- En a) μ_0=200, se aceptó H_0. ¿Es tolerable asumir que es lo mismo 150 y 200?
- En b) μ_0=230, se rechazó H_0 Efectivamente no pertenece al IC (aunque por poco)

! El uso de IC para matizar la conclusión de un test de hipótesis se verá con detalle en el capítulo dedicado a los *test con dos muestras* (estudios comparativos)

Pedro Femia
pfemia@ugr.es

ESTADÍSTICA

Grado en Enfermería

Tema V

Tests con una muestra

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Estudio de la normalidad

- La distribución normal es un modelo de distribución de probabilidad para variables continuas
- Generalmente interesa más la **normalidad de la distribución de la media** de la variable que de la variable en si.

Recordemos que:

La media de una VA con distribución normal también tiene distribución normal (con dispersión menor cuanto mayor es el tamaño de la muestra n)

$$\text{Si } X \to \mathrm{N}(\mu;\sigma) \;\Rightarrow\; \bar{X} \to \mathrm{N}\left(\mu;\sigma/\sqrt{n}\right)$$

La media de una VA con distribución desconocida, tiene una distribución aproximadamente normal si el tamaño de muestra es $n \geq 60$. La dispersión disminuye y la calidad de la aproximación aumenta cuanto mayor sea el tamaño de la muestra n (Teorema del Límite Central)

$$\text{Si } X \to ?? \;\Rightarrow\; \bar{X} \underset{n\geq 60}{\overset{\text{aprox.}}{\to}} \mathrm{N}\left(\mu;\sigma/\sqrt{n}\right)$$

- En la práctica:
  - Si el tamaño de muestra es $n \geq 60$ ← La normalidad de la distribución de la media muestral viene avalada por el **teorema del límite central**, incluso si la variable es discreta (en tal caso se debe incluir una corrección por continuidad (cpc))
  - Si el tamaño de muestra es $n < 60$:
    - si la variable es continua: estudiar su normalidad estadística ← Análisis gráfico + Test de normalidad
    - si la variable es discreta: no deben utilizarse a priori métodos basados en la distribución normal

Estudio de la normalidad

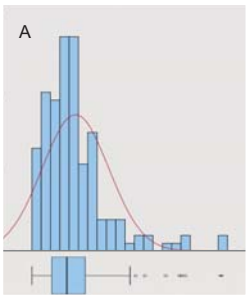
- Se trata de un problema que puede ser difícil de resolver. Especialmente si el tamaño de muestra es pequeño
- Se deben combinar los test de hipótesis con métodos gráficos

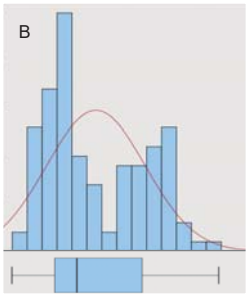
Análisis gráfico →
- Histograma de la variable con curva de referencia
- Diagramas de caja
- Otros métodos específicos (*diagramas probabilísticos P-P, Q-Q,…*)

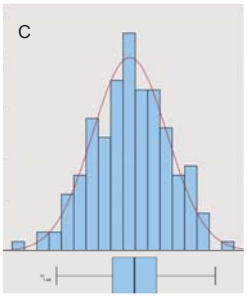
Pruebas (tests) de normalidad →
- Test de D'Agostino ← Se puede hacer a mano
- Test de Kolmogorov-Smirnov ← Aproximado. Solo para muestras grandes (SPSS)
- Test de Shapiro-Wilk ← Quizá la mejor opción (SPSS)

(… y hay más)

- Muchos métodos que requieren que la variable sea normal, a menudo funcionan bien si la desviación no es muy severa → Observar la simetría de la distribución







(A) Asimetría positiva
Muchos "tienen" poco y pocos "tienen" mucho

(B) Mezcla de 2 grupos heterogéneos

(C) ¿Normal?

Estudio de la normalidad

Test de normalidad de D'Agostino

1. Test de D'Agostino. Hipótesis a contrastar

$$\begin{cases} H_0: \text{La variable tiene distribución normal} \\ H_1: \text{La variable NO tiene distribución normal} \end{cases}$$

En todos los tests de normalidad, H_0 es siempre "La variable es normal"

2. Estadístico de contraste COMPLEMENTO

$$D_{exp} = \frac{\sum i\, x_{(i)} - (n+1)\left(\sum x_i\right)/2}{n\sqrt{n(n-1)\,s^2}}$$

(Confrontar con la tabla de D'Agostino)

El peligro de las muestras pequeñas

n pequeño ↓ *Test poco potente* ↓ *No se rechaza H_0* ↓ *¡La variable es normal!*

3. Ejemplo de cálculo COMPLEMENTO

	x_i	$x_{(i)}$	i	$i\,x_{(i)}$
1	2.1	1.3	1	1.3
2	8.6	2.1	2	4.2
3	6.1	2.6	3	7.8
4	3.5	3.4	4	13.6
5	6.2	3.5	5	17.5
6	3.4	4.9	6	29.4
7	5.5	5.5	7	38.5
8	9.7	6.1	8	48.8
9	1.3	6.2	9	55.8
10	2.6	8.5	10	85.0
11	8.5	8.6	11	94.6
12	4.9	9.7	12	116.4

H_0: x tiene distribución normal

$n = 12$

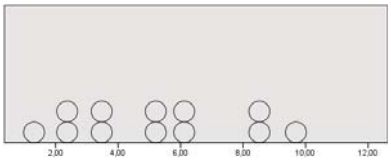
$\sum x_i = 62.37$

$\sum i\, x_{(i)} = 512.06$

$s^2 = 7.40$

$D_{exp} = 0.2844$

$0.10 < p < 0.20$ ← ¿conclusión?



Distribución observada de x
(Diagrama de puntos)

Pedro Femia
pfemia@ugr.es

Test con una proporción

COMPLEMENTO

Ejemplo:
Según la encuesta nacional de salud de 2012, la prevalencia del tabaquismo en los estudiantes de primer curso universitario es del 22%. Para estudiar si ha variado en 2013 se ha tomado una muestra de 300 estudiantes de los cuales: 48 son fumadores, 239 no lo son y el resto no contestan. ¿Puede afirmarse que la proporción de fumadores en este sector de la población ha cambiado respecto a 2012?

1. **Información muestral**:

   Tamaño de muestra original = 300; pero solo contestan 239+48=287 = n

$$\begin{cases} n = 287 \\ x = 48 \\ n\text{-}x = 239 \end{cases}$$

2. **Hipótesis**: se trata de comprobar si la prevalencia ha cambiado respecto al valor propuesto p_0=0.22

$$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases} \rightarrow \begin{cases} H_0: p = 0.22 \\ H_1: p \neq 0.22 \end{cases} \begin{array}{l} \leftarrow \text{La prevalencia no ha cambiado} \\ \leftarrow \text{La prevalencia si ha cambiado} \end{array}$$

   Fijamos el error de tipo I a un nivel α=0.05, de manera que declararemos significativo a todo resultado que de lugar a un nivel de significación $P \leq \alpha$=0.05

3. **Condiciones de validez**: debe cumplirse que

$$\begin{cases} np_0 > 5 \\ n(1-p_0) = nq_0 > 5 \end{cases} \rightarrow \begin{cases} np_0 = 287 \times 0.22 = 63.14 \\ nq_0 = 287 \times 0.78 = 223.86 \end{cases}$$

Se cumple: podemos aplicar el método
(de lo contrario no podemos seguir)

4. Evidencia muestral: el **estadístico de contraste**:

$$z_{exp} = \frac{|x - np_0| - 0.5}{\sqrt{np_0 q_0}} = \frac{|48 - 287 \times 0.22| - 0.5}{\sqrt{287 \times 0.22 \times 0.78}} = 2.086$$

Test con una proporción

COMPLEMENTO

5. **Nivel de significación** (se obtiene de la tabla de la distribución normal): $0.03 < P < 0.04$

6. **Interpretación**: el resultado es **significativo**, es decir que rechazamos la hipótesis nula de que la prevalencia en 2013 es la misma que la observada en 2012

   Lecturas posibles (y equivalentes) de este valor de P

   - La probabilidad de equivocarnos al rechazar la hipótesis nula con estos datos es menor al 4%. Como esta probabilidad es menor a α=5%, el nivel de error es tolerable y en consecuencia rechazamos H_0

   - Si H_0 es cierta, es decir si la prevalencia es del 22% en 2013, la probabilidad de observar una muestra como la actual, o más discrepante con dicha hipótesis, es del 4%. Consideramos que esa probabilidad es baja (<α) y en consecuencia rechazamos H_0

7. **Análisis de la significación**. Como rechazamos la H_0 no puede asumirse que la prevalencia es el valor propuesto, ¿Qué valor toma entonces dicho parámetro? ← IC para una proporción

					%		
Método	IC(-)	IC(+)	d	Validez	IC(-)	IC(+)	d
Wilson	0.12698	0.21665	0.04483	Si	12.70%	21.67%	4.48%
Wald	0.12233	0.21217	0.04492	Si	12.23%	21.22%	4.49%
Wald Ajustado	0.12848	0.21516	0.04334	Si	12.85%	21.52%	4.33%

Confianza: 95%

Atendiendo al método de Wilson, la prevalencia del tabaquismo es, con un 95% de probabilidad, un valor comprendido entre el 12.7% y el 21.7%. Podemos asumir que ha disminuido respecto al 22% del año anterior. Observese como p_0=0.22 no está en el IC para p

Pedro Femia
pfemia@ugr.es

Test con una proporción

COMPLEMENTO

¿Y si los datos muestrales hubieran sido? $\begin{cases} n = 200 \\ x = 33 \\ n\text{-}x = 167 \end{cases}$

$$\begin{cases} np_0 > 5 \\ n(1-p_0) = nq_0 > 5 \end{cases} \longrightarrow \begin{cases} np_0 = 200 \times 0.22 = 44 \\ nq_0 = 200 \times 0.78 = 156 \end{cases}$$ El método es válido

$$z_{exp} = \frac{|x - np_0| - 0.5}{\sqrt{np_0q_0}} = \frac{|33 - 200 \times 0.22| - 0.5}{\sqrt{200 \times 0.22 \times 0.78}} = 1.792 \longrightarrow 0.07 < P < 0.08 \longrightarrow$$

El resultado no es significativo. No podemos rechazar la hipótesis nula con estos datos, aunque hay **indicios de significación**.

Observemos ahora el 95%-IC para p

					%		
Método	IC(-)	IC(+)	d	Validez	IC(-)	IC(+)	d
Wilson	0.11782	0.22541	0.05380	Si	11.78%	22.54%	5.38%
Wald	0.11106	0.21894	0.05394	Si	11.11%	21.89%	5.39%
Wald Ajustado	0.11983	0.22330	0.05173	Si	11.98%	22.33%	5.17%

Confianza: 95%

Atendiendo al IC del método de Wilson:

- El valor p_0 pertenece al IC: $0.22 \in (0.1178;\ \ 0.2254)$
- El valor de este intervalo es el de apreciar que la prueba no encuentra diferencias entre un 22% y un 11.78% (el límite del IC mas discrepante). Como esto puede ser una discrepancia muy grande (11% es la mitad de 22%) observamos un síntoma de falta de potencia del test. La aceptación de H_0 no resulta fiable y sería recomendable aumentar el tamaño de muestra (para ganar potencia)

Test con una proporción

COMPLEMENTO

La **determinación del tamaño de muestra** requiere fijar de antemano:

- El nivel de error α asumible
- La potencia deseada (o de forma equivalente el nivel de error asumible β)
- La mínima diferencia con p_0 a detectar (δ)

Considerando el valor más cercano a 0.5 de $p_1 = p_0 \pm \delta$ $\longrightarrow$ $n \geq \left(\frac{z_\alpha \sqrt{p_0q_0} + z_{2\beta} \sqrt{p_1q_1}}{\delta} \right)^2$

- Por ejemplo, si interesa declarar significativa una diferencia del 6% con una potencia del 90% a un error α del 5%

$$\alpha = 0.05 \rightarrow z_a = 1.96$$
$$\theta = 1 - \beta = 0.90; \beta = 0.10 \rightarrow z_{2\beta} = z_{0.20} = 1.282$$
$$\delta = 0.06$$
$$p_0 \pm \delta = 0.22 \pm 0.06 = \begin{cases} 0.16 \\ 0.28 = p_1 \end{cases}$$ Por ser el más cercano a 0.5

$$n \geq \left(\frac{1.96\sqrt{0.22 \times (1-0.22)} + 1.282\sqrt{0.28 \times (1-0.28)}}{0.06} \right)^2 = 534.6 \rightarrow 535$$

Son necesarias más de 534 observaciones (=535)

Pedro Femia
pfemia@ugr.es

ESTADÍSTICA

Grado en Enfermería

Tema VI

Estudios comparativos con dos muestras

(1ª parte: comparación de dos medias)

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Pruebas a realizar

Comparación de medias

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

La hipótesis de **homogeneidad** de las medias ($\mu_1 = \mu_2$) equivale a la de **independencia** de la variable cuantitativa estudiada respecto a la variable dicotómica que divide a esta en dos grupos (con medias μ_1 y μ_2)

Homogeneidad entre **muestras** $\Rightarrow$ **independencia** entre **variables**

Implica considerar:

- El tipo de **muestreo**: muestras **independientes** o muestras **relacionadas**
- La **normalidad**: prueba de normalidad de la variable estudiada o bien posibilidad de aproximación a dicha distribución
- En caso de asumir normalidad: **homogeneidad de las varianzas** (test accesorio)

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Comparación de proporciones

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$$

La hipótesis de **homogeneidad** de las proporciones ($p_1 = p_2$) equivale a la de **independencia** de la variable binomial estudiada respecto a la variable dicotómica que divide a esta en dos grupos (con proporciones p_1 y p_2) (Se verá en el capítulo que viene)

Implica considerar:

- El tipo de **muestreo**: muestras **independientes** o muestras **relacionadas**

Muestras Independientes y muestras relacionadas

Muestras *independientes*

Las observaciones de una de las muestras no condiciona, en nada, a ninguna de las observaciones de la otra. No hay ninguna relación entre las observaciones de cada muestra

→ Ej. Estudio de un indicador fisiológico en relación al género, masculino o femenino

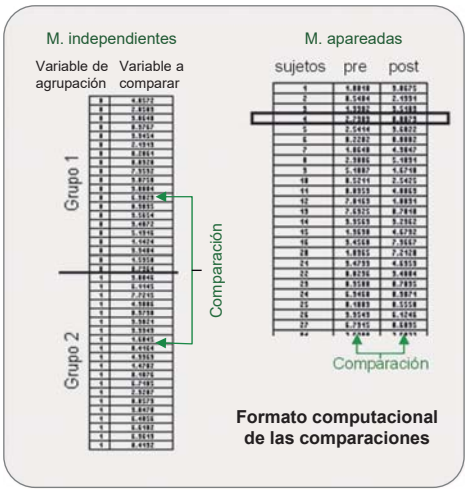
Muestras *relacionadas o apareadas*

Cada observación de una muestra tiene un vínculo con una observación de la otra muestra (la información viene en parejas de datos)

Tipos y ejemplos

COMPLEMENTO

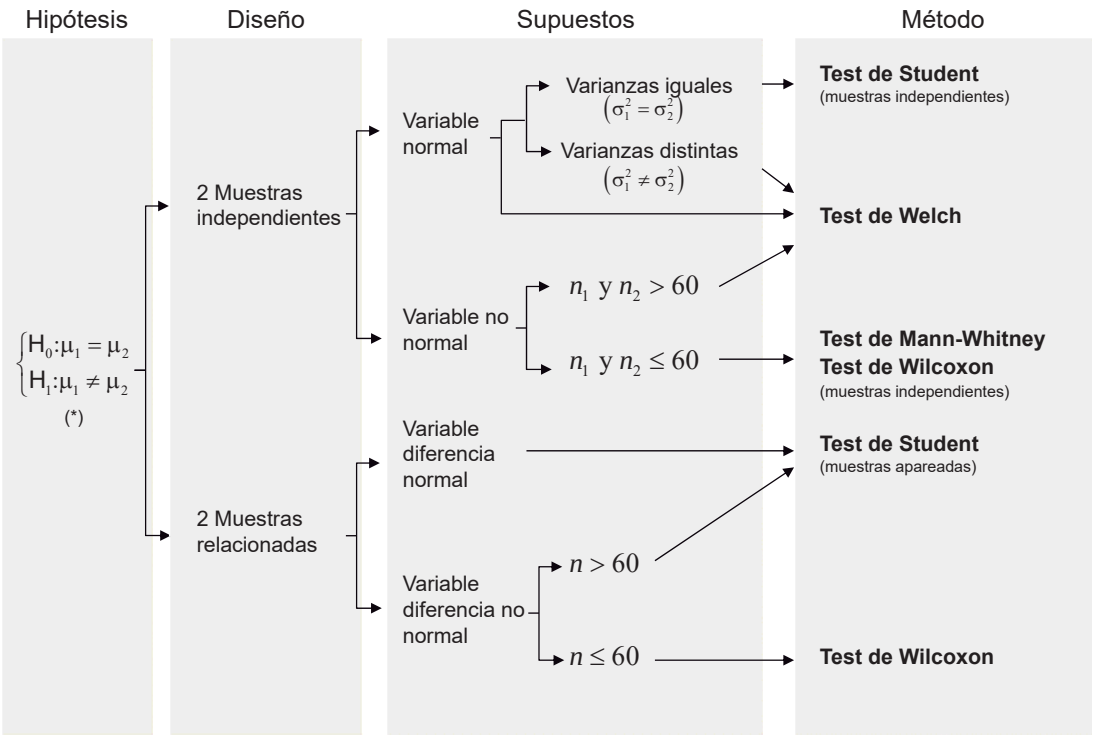
- ***Autoapareamiento***: el mismo sujeto aporta los dos datos
  → Ej: estudios pretest/posttest, el mismo individuo aporta un dato *antes* y otro *después* de una intervención
- ***Apareamiento natural***: no se trata del mismo sujeto, pero hay un vínculo natural entre los datos de la pareja
  → Ej: En experimentación en laboratorio, animales de la misma camada, hermanos gemelos,…
- ***Apareamiento artificial***: establecido por el investigador en términos de variables de interés (cuando no es viable el autoapareamiento)
  → Ej: Se persigue que cada observación de una muestra tenga un homólogo en la otra muestra



Interés de las muestras relacionadas

- Cada sujeto sirve como su propio **control** (autoapareamiento)
- Permite hacer **comparaciones homogéneas**
- En general la **potencia del test es mayor** que con un diseño independiente equivalente

Comparación de dos medias: esquema general



* O bien las alternativas unilaterales: $H_1:\mu_1>\mu_2$; o $H_1:\mu_1<\mu_2$

Pedro Femia
pfemia@ugr.es

Test preliminar a la comparación de medias: Prueba de homogeneidad de las varianzas

1. Hipótesis a contrastar

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 & \rightarrow \sigma_1^2/\sigma_2^2 = 1 \\ H_1: \sigma_1^2 \neq \sigma_2^2 & \rightarrow \sigma_1^2/\sigma_2^2 \neq 1 \end{cases}$$

*Para probar si dos **varianzas** son iguales, se **dividen**, a ver si el cociente vale 1*

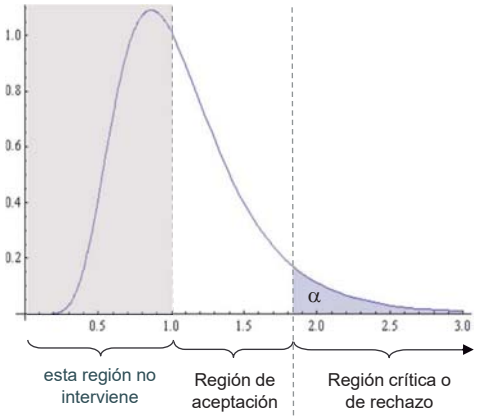
2. Evidencia empírica: el estadístico de contraste

$$F_{exp} = \frac{s_1^2}{s_2^2} \quad \left(\text{para } s_1^2 \geq s_2^2\right)$$

Si H_0 es cierta, F_{exp} debe ser "próximo" a 1

¿cuánto de "próximo"? → pertenecer a la región de aceptación

Distribución F de Snedecor



$$\left(s_1^2 \geq s_2^2\right) \Rightarrow F_{exp} \geq 1$$

F_α

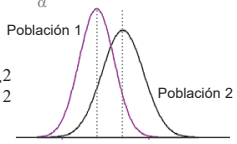
3. Decisión (para α fijado de antemano)*

Comparar F_{exp} vs $F_{0.10;(n_1-1);(n_2-1)}$

Si

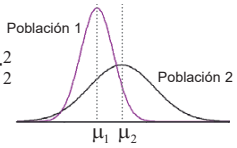
- $F_{exp} \leq F_{0.10;(n_1-1);(n_2-1)}$ →
  - Se puede asumir que $F_{exp} = s_1^2/s_2^2 \approx 1$
  - El estadístico experimental toma un valor en la región de aceptación

  → $\sigma_1^2 = \sigma_2^2$



- $F_{exp} > F_{0.10;(n_1-1);(n_2-1)}$ →
  - No se puede asumir que $F_{exp} = s_1^2/s_2^2 \approx 1$
  - El estadístico experimental toma un valor en la región crítica o de rechazo

  → $\sigma_1^2 \neq \sigma_2^2$



* Este test es el único que resolveremos sin determinar el nivel de significación *p*, y lo hacemos así por simplicidad, debido al formato de las tablas. En el fondo la comparación que interesa es la de las medias. Este test solo nos dice qué método se debe aplicar en la comparación de medias.

Prueba de homogeneidad de medias

1. Hipótesis a contrastar

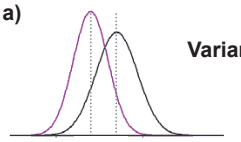
$$\begin{cases} H_0: \mu_1 = \mu_2 & \rightarrow \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 \neq \mu_2 & \rightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$$

*Para probar si dos **medias** son iguales, se **restan**, a ver si la diferencia vale 0*

2. Evidencia empírica: el estadístico de contraste

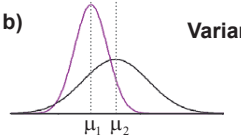
- Si H_0 es cierta, $|\overline{x}_1 - \overline{x}_2|$ debe ser una magnitud pequeña: $|\overline{x}_1 - \overline{x}_2| \rightarrow 0$
- ¿Cuánto de pequeña? → dependerá de la variabilidad y del tamaño de muestra
- El Estadístico de contraste es una estandarización de esa diferencia: $t_{exp} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{*}}$ con *f* grados del libertad
- Punto de vista clásico:
  - El modo de estandarizar (quién es $\sqrt{*}$)
  - Los grados de libertad f a considerar

  } dependen de cómo sean las varianzas poblacionales

a) 

Varianzas homogéneas

$\sigma_1^2 = \sigma_2^2$ → **Test de Student** { Es uno de los test más clásicos (y más utilizados) de la Estadística

b) 

Varianzas NO homogéneas

$\sigma_1^2 \neq \sigma_2^2$ → **Test de Welch** { Algunos autores recomiendan usar siempre este, especialmente si los tamaños de muestra son muy distintos

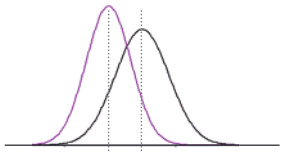
Pedro Femia
pfemia@ugr.es

1. Hipótesis a contrastar

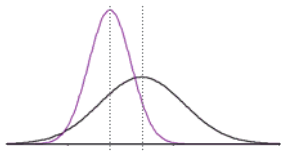
$$\begin{cases} H_0: \mu_1 = \mu_2 & \rightarrow \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 \neq \mu_2 & \rightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$$

2. Evidencia empírica: el estadístico de contraste

• Modelo supuesto:



$(\sigma_1^2 = \sigma_2^2)$



$(\sigma_1^2 \neq \sigma_2^2)$

• Método: **Test de Student** **Test de Welch**

- Estadístico de contraste:

$$t_{exp} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{s_p^2 (n_1 + n_2)/(n_1\, n_2)}} \qquad t_{exp} = \frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{A + B}}$$

con:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \qquad A = \frac{s_1^2}{n_1};\ B = \frac{s_2^2}{n_2}$$

- Grados de libertad:

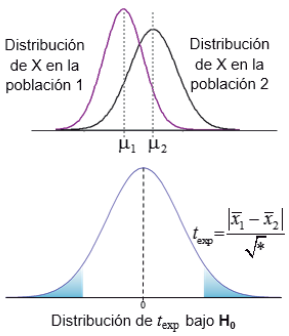
$$f = (n_1 + n_2 - 2)\ \text{g.l.} \qquad f = \frac{(A+B)^2}{A^2/(n_1 - 1) + B^2/(n_2 - 1)}\ \text{g.l.}$$

en cualquier caso, si H_0 es cierta será $t_{exp} \rightarrow 0$

Muestras independientes // Variable normal

Prueba de homogeneidad de medias

No confundir:

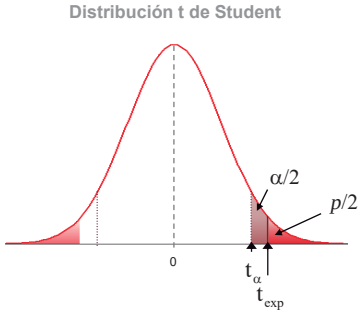


3. Decisión: obtención e interpretación del nivel de significación *p*

• El *estadístico de contraste* t_{exp} tiene una distribución *t*-Student con f grados de libertad.

• En base a ella, se trata de determinar si realmente $t_{exp} \rightarrow 0$. El procedimiento a seguir será obtener el nivel de significación *p* y compararlo con el valor de α fijado de antemano.
Dos resultados posibles:

[1]



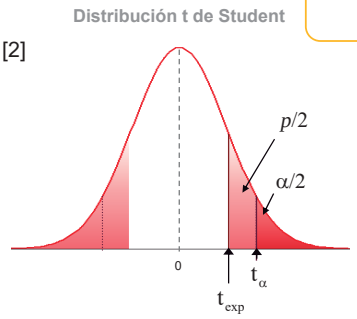
$p \leq \alpha$
Rechazar H_0

La probabilidad de cometer un error al asumir que μ_1 es diferente de μ_2 con la muestra actual, es de *p* (%) (menor que el error α tolerable, luego SI se puede asumir)

↓

Test **significativo**

[2]



$p > \alpha$
No se puede rechazar H_0

La probabilidad de cometer un error al asumir que μ_1 es diferente de μ_2 con la muestra actual, es de *p* (%) (mayor que el error α tolerable, luego NO se puede asumir)

↓

Test **no significativo**

Pedro Femia
pfemia@ugr.es

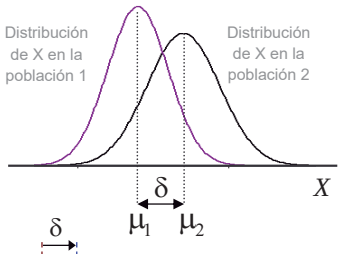
4. Análisis de la significación: IC para la diferencia de medias

- Tras realizar el test debe darse el **intervalo de confianza (IC) para la diferencia** $(\mu_1-\mu_2)$
- Formato general del IC tras un t-test:

$IC_\alpha =$ (Numerador del test) $\pm\ t_{\alpha;f}$ (Denominador del test) Es decir, si $t_{exp}=\dfrac{|\bar{x}_1-\bar{x}_2|}{\sqrt{*}} \longrightarrow (\bar{x}_1-\bar{x}_2)\pm t_{\alpha;f}\sqrt{*}$

con $\sqrt{*}$ y f los correspondientes al test de Student o de Welch que se haya utilizado y (1-α) la confianza del intervalo

Interés y particularidades del IC para $(\mu_1-\mu_2)$ tras el test

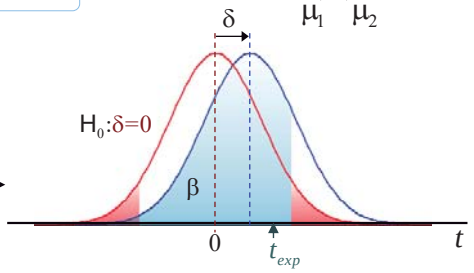


a) Si se ha rechazado H_0 y se ha decidido $(\mu_1-\mu_2)\neq 0$

- El error asociado a tal afirmación es conocido: p
- Interesa estimar la magnitud de la diferencia, el **tamaño del efecto δ** →

b) Si no se ha rechazado H_0 y se ha decidido $(\mu_1-\mu_2)=0$

- El error asociado a tal afirmación no es conocido (**β**)
- Interesa **estudiar la fiabilidad de la decisión por H_0** →

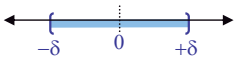


Resultado no significativo ante un **efecto δ**
¿es fiable la decisión **δ=0**?

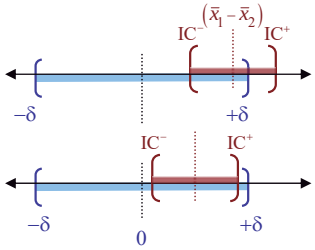
4. Análisis de la significación: IC para la diferencia de medias

a) Si el test SI ha sido significativo: tamaño del efecto

- El intervalo $(\bar{x}_1-\bar{x}_2)\pm t_{\alpha;f}\sqrt{*}$ representa el tamaño del efecto
- ¿cuál es la magnitud de la diferencia que se considera relevante? → **δ**
  - Denotamos así a la mínima **diferencia relevante**



  - La establece el investigador
- El Intervalo de confianza se hace al nivel de error α preestablecido



Significación estadística + Significación física

La magnitud de la diferencia detectada puede ser mayor a δ

Significación estadística + No significación física

La magnitud de la diferencia detectada no es mayor a δ

Medidas de efecto estandarizado

COMPLEMENTO

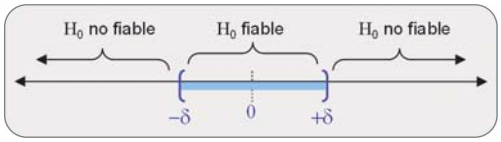
Omega cuadrado (Thomas & Nelson; 2001)	$\omega^2=(t^2_{exp}-1)/(t^2_{exp}+n_1+n_2-1)$	% de variabilidad explicada por la variable de agrupación
Efecto de Cohen (1988)	$d=(\bar{x}_1-\bar{x}_2)/s_c$	con s_c la d.t. agrupada del test de Student, o la relativa al grupo de referencia (d ~0.2 efecto no relevante; ~ 0.5 efecto moderado; >0.8 efecto importante)
% de cambio	$pc=(\bar{x}_2-\bar{x}_1)/\bar{x}_1\times 100\%$	con $\bar{x}_1$ la media del grupo de referencia

Pedro Femia
pfemia@ugr.es

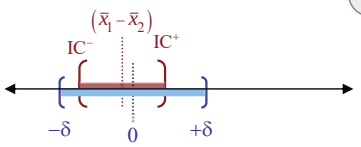
4. Análisis de la significación: IC para la diferencia de medias

b) Si el test NO ha sido significativo: Fiabilidad de la decisión por H_0

- El estudio de la fiabilidad de H_0 requiere establecer:
  - La diferencia que se considere significativa δ (como antes)
  - El error de tipo II (β) máximo tolerable o equivalentemente La potencia (1-β) con que interesa detectar a δ
- Para vincularlo a la potencia del test (1-β), el IC se hace a nivel de error 2β
- ¿Qué valores toma el intervalo $(\bar{x}_1-\bar{x}_2)\pm t_{2\beta;f}\sqrt{*}$ respecto a δ?
  ¿a qué le llamamos "no diferencia"?

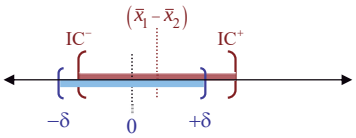


Decisión por H_0 fiable



$(\bar{x}_1-\bar{x}_2)\pm t_{2\beta;f}\sqrt{*}$
no toma valores mas allá de $\pm\delta$

Decisión por H_0 no fiable



$(\bar{x}_1-\bar{x}_2)\pm t_{2\beta;f}\sqrt{*}$
toma valores mas allá de la diferencia admisible

Complemento

5. Tamaño de muestra

Propósito

Se trata de determinar el tamaño de muestra (o muestras) necesario para que un test al error α, de significativo el (1-β)x100% de las veces en que las medias difieran al menos $\delta=|\mu_1-\mu_2|$

Requisitos:

Muestras piloto que permitan inferir la variabilidad
- Estimación: $s_1^2=\hat{\sigma}_1^2$ y $s_2^2=\hat{\sigma}_2^2$
- Saber si es admisible H_0: $\sigma_1^2=\sigma_2^2$

Cantidades a fijar (por el investigador)
- Magnitud del error de tipo I: α
- Magnitud del error de tipo II (o potencia deseada): β (ó $1-\beta$)
- Diferencia mínima a detectar: $\delta=|\mu_1-\mu_2|$

Situación:

Varianzas iguales

$$n=\left(\frac{t_\alpha+t_{2\beta}}{\delta}\right)2s_p^2;\quad n_1=n_2=n$$
$$(n_1+n_2-2)\ g.l.$$

← Es preferible que las muestras sean de igual tamaño

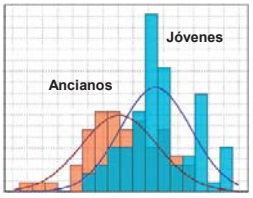
Varianzas distintas

$$n_1=\left(\frac{t_\alpha+t_{2\beta}}{\delta}\right)(r+1)s_1^2;\quad n_2=r\times n_1$$
$$r=\hat{\sigma}_2/\hat{\sigma}_1$$
$$g.l.=(1+r)^2\Big/\left[\frac{1}{n_1'-1}+\frac{r^2}{n_2'-1}\right]$$

← La muestra con mayor varianza debe ser la mayor
A mayor variabilidad, mayor es la información necesaria para poder caracterizar a la población

Pedro Femia
pfemia@ugr.es

Ejemplo 1: ¿Cambia el nivel de proteína α-klotho sérica con la edad?



$\mu_1 \equiv$ Promedio del nivel de proteina plasmática (pg/ml) en jovenes (~25 años)

$\mu_2 \equiv$ Promedio del nivel de proteina plasmática (pg/ml) en ancianos (~80 años)

Hipótesis $\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \rightarrow$ Los dos grupos son **homogéneos** respecto al nivel de proteína, es decir, el nivel de proteína no depende de la edad

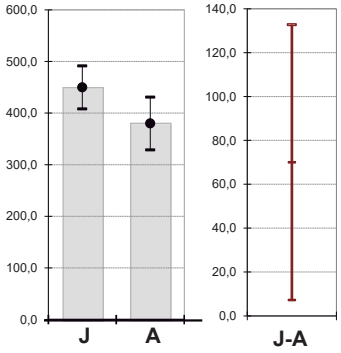
Muestras: $\begin{cases} 1(J): \ n_1 = 20, \bar{x}_1 = 450, s_1 = 89 \\ 2(A): \ n_2 = 16, \bar{x}_2 = 380, s_2 = 96 \end{cases}$

Validez del método: suponemos que el nivel de proteina es una VAC con distribución normal

Comparación de varianzas:

$\begin{cases} H_0 \equiv \sigma_1^2 = \sigma_2^2 \\ H_1 \equiv \sigma_1^2 \neq \sigma_2^2 \end{cases}$ $F_{exp} = \left(\frac{96}{89}\right)^2 = 1.16 \ < \ F_{0.10;(15,\,19)} = 1.86$

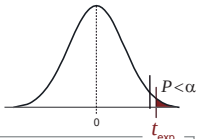
$F_{exp} < F_{0.10;(15,\,19)} \Rightarrow$ Varianzas homogéneas



Comparación de medias (student):

$$s^2 = \frac{(19)(89^2)+(15)(96^2)}{19+15} = 8492.32$$

$$t_{exp} = \frac{|450-380|}{\sqrt{(8492.32)\left(\frac{1}{20}+\frac{1}{16}\right)}} = \frac{70}{30.91} = 2.265 \ \ (34\ g.l.) \Rightarrow$$



$0.02 < P < 0.05$

La diferencia es **estadísticamente significativa**

IC: estimación del efecto de la edad sobre el nivel de proteina

$$IC_{95\%}(\mu_1 - \mu_2) = 70 \pm 2.042 \times 30.91 = 70 \pm 62.8 = (7.18;\ 132.81)$$

La diferencia es un valor que, con un 95% de confianza, se espera que esté comprendido entre 7.18 y 132.81 pg/ml

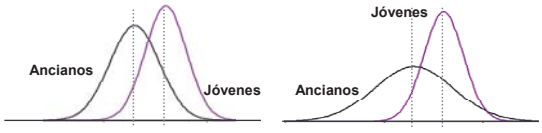
Ejemplo 2: en el contexto del Ejemplo 1 ¿y si los datos hubieran sido estos?

Muestras: $\begin{cases} 1: \ n_1 = 20, \bar{x}_1 = 450, s_1 = 59.0 \\ 2: \ n_2 = 16, \bar{x}_2 = 380, s_2 = 96.0 \end{cases}$

Comparación de varianzas:

$\begin{cases} H_0 \equiv \sigma_1^2 = \sigma_2^2 \\ H_1 \equiv \sigma_1^2 \neq \sigma_2^2 \end{cases}$ $F_{exp} = \left(\frac{96}{59}\right)^2 = 2.65; \quad F_{0.10;(15;19)} \approx 1.86$

$F_{exp} > F_{0.10;(15,\,19)} \Rightarrow$ Varianzas distintas



Situación del ejemplo 1 | Situación actual

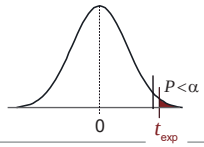
Comparación de medias: Test de Welch

$$A = \frac{s_1^2}{n_1} = \frac{59^2}{20} = 174.05 \qquad B = \frac{s_2^2}{n_2} = \frac{96^2}{16} = 576.00$$

$$t_{exp} = \frac{|450-380|}{\sqrt{174.05+576.00}} = \frac{70}{27.387} = 2.56$$

$$gl = \frac{(174.05+576.00)^2}{\frac{174.05^2}{20-1}+\frac{576.00^2}{16-1}} = 23.72 \rightarrow 23$$

¡Antes, en el test de Student, eran 34 *gl*!



$0.01 < P < 0.02$

IC: estimación del efecto de la edad sobre el nivel de proteina

$$IC_{95\%}(\mu_1 - \mu_2) = 70 \pm 2.069 \times 27.387 = 70 \pm 56.65 = (13.34;\ 126.65)$$

La diferencia es un valor que, con un 95% de confianza, se espera que esté comprendido entre 13.34 y 126.65 unidades. El aumento de la variabilidad y el hecho de que no sea homogénea en las dos muestras genera un intervalo menos preciso que en el ejemplo anterior. Se han perdido 11 gl.

Pedro Femia
pfemia@ugr.es

Ejemplo 3: ¿Depende la concentración de glucocorticoides en orina (mg/24h) del sexo del individuo?

$\mu_1 \equiv$ Promedio del nivel de glucocorticoides en orina en varones

$\mu_2 \equiv$ Promedio del nivel de glucocorticoides en orina en mujeres

Hipótesis $\begin{cases} H_0 \equiv \mu_1 = \mu_2 & \text{(la concentración de glucocorticoides es la misma en varones y mujeres)} \\ H_1 \equiv \mu_1 \neq \mu_2 & \text{(la concentración de glucocorticoides depende del sexo)} \end{cases}$

Muestras: $\begin{cases} 1: \ n_1 = 15, \bar{x}_1 = 5.5, s_1 = 1.7 \text{ mg/24h} \\ 2: \ n_2 = 11, \bar{x}_2 = 6.1, s_2 = 1.5 \text{ mg/24h} \end{cases}$

* Suponemos NORMALIDAD de la variable:
X = Nivel de glucocorticoides $\rightarrow N(\mu,\sigma)$

Varianzas: $\begin{cases} H_0 \equiv \sigma_1^2 = \sigma_2^2 \\ H_1 \equiv \sigma_1^2 \neq \sigma_2^2 \end{cases}$ $F_{exp} = \left(\frac{1.7}{1.5}\right)^2 = 1.28 < F_{0.10;(14,\,10)} \approx 2.28$

Asumimos que las varianzas son homogéneas $\rightarrow$ Test de Student

$s^2 = \frac{(14)(1.7^2)+(10)(1.5^2)}{15+11-2} = 2.623$ $\quad t_{exp} = \frac{|5.5-6.1|}{\sqrt{(2.623)\left(\frac{1}{15}+\frac{1}{11}\right)}} = 0.93 \ (24\ g.l.)$

$0.30 < P < 0.40$

No significativo

¿Es fiable el resultado si se desea detectar una diferencia de δ=0.5 unidades con una potencia del 90%?

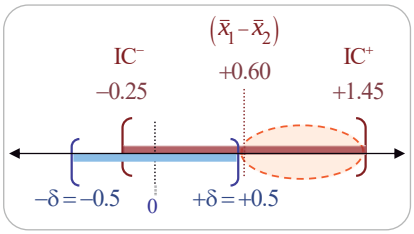
Información de partida:
$\alpha = 5\%$
$1-\beta = 90\% \Rightarrow \beta = 10\%$
$\delta = 0.5$

$IC_{80\%}(\mu_1-\mu_2) = (6.1-5.5) \pm 1.318\sqrt{2.623(1/15+1/11)} = (-0.25;\ +1.45)$

Cómo $\delta=0.5 \in IC_{80\%}(\mu_1-\mu_2) = (-0.25;\ +1.45)$

la decisión por H_0 NO es fiable

* Si interesara δ=2, entonce si que es fiable:
$\delta=2 \notin IC_{80\%}(\mu_1-\mu_2)$



Tamaño de muestra necesario para detectar δ=2 con un 80% de potencia a un error del 5%

$n \geq \left(\frac{2.064+1.318}{0.5}\right)^2 \times 2 \times 2.623$

$\rightarrow n_1 = n_2 \geq 241$

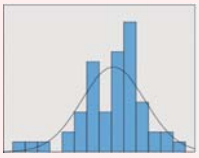
COMPLEMENTO

Prueba de homogeneidad de medias

1. Test de Student. Hipótesis a contrastar

$\begin{cases} H_0: \mu_{antes} = \mu_{después} & \rightarrow \mu_{antes} - \mu_{después} = \mu_{Diferencial} = 0 \\ H_1: \mu_{antes} \neq \mu_{después} & \rightarrow \mu_{antes} - \mu_{después} = \mu_{Diferencial} \neq 0 \end{cases}$

- Es la variable diferencia d la que debe tener distribución normal



- La información relevante es:
  - n ← Tamaño de muestra
  - $\bar{d}$ ← Media
  - s_d^2 ← Varianza

2. Evidencia empírica: el estadístico de contraste

- Con las muestras apareadas, lo que se estudia es la variable diferencia

Individuo:	1	2	⋯	n
Antes: x_a	x_{a_1}	x_{a_2}	⋯	x_{a_n}
Después: x_d	x_{d_1}	x_{d_2}	⋯	x_{d_n}
Diferencia:	$d_1 = x_{a_1} - x_{d_1}$	$d_2 = x_{a_2} - x_{d_2}$	⋯	$d_n = x_{a_n} - x_{d_n}$

- Si H_0 es cierta, $\bar{d} = \left(\bar{x}_{antes} - \bar{x}_{después}\right)$ debe de ser una magnitud pequeña: $\bar{d} \rightarrow 0$

a) Estadístico de contraste: $t_{exp} = \left|\bar{d}\right| \Big/ \sqrt{s_d^2/n}$

b) Intervalo de confianza: $\bar{d} \pm t_{\alpha;n-1}\sqrt{s_d^2/n}$

c) Tamaño de muestra: $n = \left(\frac{t_\alpha + t_{2\beta}}{\delta}\right) s_d^2$

El proceso a seguir es como antes, si

$p \leq \alpha$ Rechazar H_0 ⟶ Estimar el efecto con el IC

$p > \alpha$ No se puede rechazar H_0 ⟶ Estudiar la fiabilidad de la decisión con el IC

* En todos los casos considerar la distribución t de Student con $(n-1)\ g.l.$

Pedro Femia
pfemia@ugr.es

Ejemplo 4

Se estudia si la práctica continuada de ejercicio físico permite reducir el nivel de colesterol en individuos con hipercolesterolemia familiar. Participaron en el estudio n=10 sujetos a los que se les determinó el nivel de colesterol antes y después de 6 meses de desarrollar un programa de práctica de AF regular.

	Nivel de colesterol									
Sujeto	1	2	3	4	5	6	7	8	9	10
Antes	220	225	212	198	214	236	214	211	235	210
Después	190	201	213	199	205	201	189	196	185	174
Diferencia	30	24	-1	-1	9	35	25	15	50	36

Hipótesis: $\begin{cases} H_0: \mu_{antes} = \mu_{después} \rightarrow \mu_{antes} - \mu_{después} = 0 \\ H_1: \mu_{antes} \neq \mu_{después} \rightarrow \mu_{antes} - \mu_{después} \neq 0 \end{cases}$

* Suponemos normalidad de la variable diferencia: $(X_{antes} - X_{despues}) \rightarrow N(\mu,\sigma)$

Muestra: $n = 10;\ \bar{x}_{dif} = 22.2;\ s_{dif} = 16.67$

Test: $t_{exp} = |\bar{d}| \Big/ \sqrt{s_d^2/n} = |22.2| \Big/ \left(16.67/\sqrt{10}\right) = 4.21$ P= 0.002

Hay diferencias significativas

Efecto: $IC_{95\%}(\mu_1 - \mu_2) = 22.2 \pm 11.92 = (10.27;\ 34.13)$

Con un 95% de confianza se puede esperar que la reducción del nivel de colesterol sea un valor comprendido entre 10 y 34 unidades (redondeando)

Inferencias sobre la diferencia de las medias por aproximación a la Normal

- **Validez:** Si los tamaños de muestra son **mayores a 60** (o a 30 si la distribución es simétrica)
- **Test:**

	Variable continua	Variable discreta (COMPLEMENTO)
Muestras independientes	$t_{exp} = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$t_{exp} = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert - c}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
Muestras apareadas	$t_{exp} = \dfrac{\lvert\bar{d}\rvert}{\sqrt{s_d^2/n}}$	$t_{exp} = \dfrac{\lvert\bar{d}\rvert - c}{\sqrt{s_d^2/n}}$

- P se obtiene de la distribución *t-Student*
- La cantidad c es una corrección para variables discretas
  - en independientes $c = 1/2 \max\{n_1; n_2\}$
  - en apareadas $c = 1/2n$
- **Intervalo:**

Variable continua: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha;f}\sqrt{*}$

Variable discreta: $(\bar{x}_1 - \bar{x}_2) \pm (t_{\alpha;f}\sqrt{*} + c)$

$\sqrt{*}$ Es el denominador de la t_{exp} correspondiente

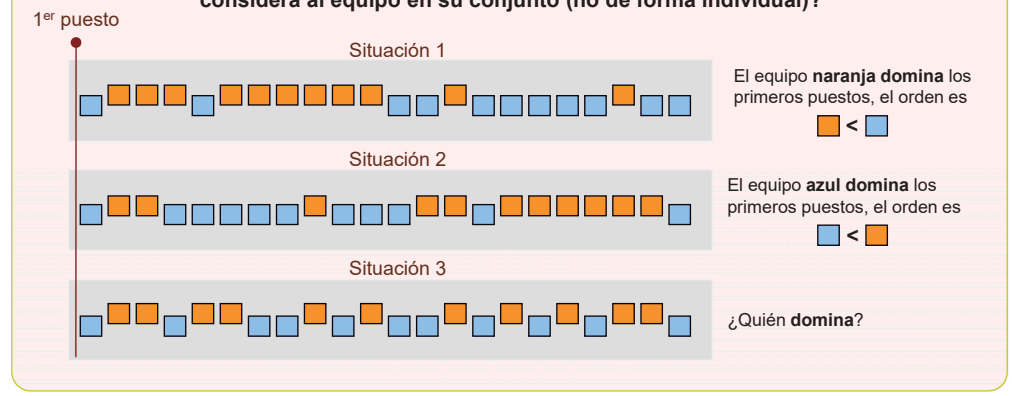
$t_{\alpha;f}$ en la distribución *t-Student* con los gl (f) adecuados

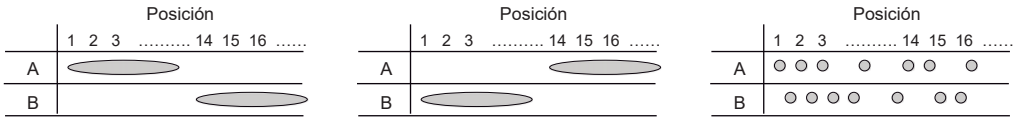
Pedro Femia
pfemia@ugr.es

Pruebas no paramétricas

En qué consiste un test de homogeneidad *no paramétrico*

Supongamos una carrera entre el equipo azul y el naranja ¿Cuál gana en cada situación si se considera al equipo en su conjunto (no de forma individual)?





A domina las posiciones bajas y B las posiciones altas
A < B

A domina las posiciones altas y B las posiciones bajas
A > B

Ni A ni B dominan posiciones concretas de la escala
A = B

¿Cómo se cuantifica esto?
¿qué criterio permite decidir si una de las muestras domina una parte de la escala ordenada?

Muestras independientes // Variable No normal

Test de Wilcoxon para muestras independientes

1. Hipótesis a contrastar

$$\begin{cases} H_0: \text{Las dos poblaciones son homogéneas} \\ H_1: \text{Una de las poblaciones tiende a dar valores mas altos o mas bajos que la otra} \end{cases}$$

2. Información muestral

Lo vemos con un ejemplo:

Se estudia si en el personal de enfermería adscrito a un centro de atención primaria A, percibe el mismo nivel de sobrecarga que el personal adscrito a un centro de atención B. Se utilizó una escala con puntuaciones enteras de 1 a 40 puntos (mayor puntuación indica mayor sobrecarga). Los datos obtenidos se indican a continuación

Centro A	12, 14, 11, 30, 10	$(n_1 = 5)$
Centro B	16, 11, 14, 21, 18, 34, 22, 7, 12, 12	$(n_2 = 10)$

3. Procedimiento

1) Ordenar las observaciones de menor a mayor (sabiendo cual es la muestra de procedencia)
2) Asignar ***rangos*** = nº de orden si no hay empates o el nº de orden promedio del grupo empatado
3) La suma de rangos de la muestra de menor tamaño es el estadístico de contraste: R_{exp}
4) Obtener el nivel de significación p

Datos	A		10	11		12			14						30		
ordenados	B	7			11			12	12		14	16	18	21	22		34	
Nº orden		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	*Totales*
Rangos	A		2	3.5		6			8.5						14		$34 = R_1$
r_i	B	1			3.5		6	6		8.5	10	11	12	13		15	$86 = R_2$

$R_{exp} = 34 \quad \rightarrow p > 0.10$ No hay evidencia de que en un centro se de mayor sobrecarga que en el otro

En caso de haberse dado diferencias significativas, la conclusión habría sido B>A, ya que se observa que $\bar{R}_2 > \bar{R}_1$
Siendo $\bar{R}_i = R_i / n_i$ los rangos promedio:

$$\bar{R}_2 = 86/10 = 8.6 \quad > \quad \bar{R}_1 = 34/5 = 6.8$$

Pedro Femia
pfemia@ugr.es

Test de Wilcoxon para muestras apareadas

1. Hipótesis a contrastar

$\begin{cases} H_0: \text{Las medidas antes (pre) y después (post) de la intervención son equivalentes} \\ H_1: \text{Las observaciones post-tratamiento toman valores que tienden a ser mayores o menores que las pre-tratamiento} \end{cases}$

2. Información muestral

Lo vemos con un ejemplo:

Se estudia si la frecuencia cardiaca (FC) cambia como respuesta a un tratamiento antihipertensivo. A continuación figuran los valores de FC (en ppm) correspondientes a n=10 pacientes tomados antes y después de dicho entrenamiento.

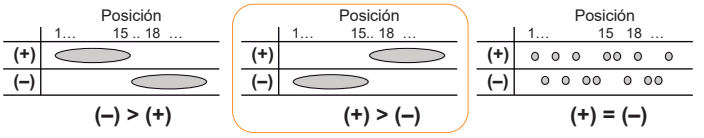
Paciente:	1	2	3	4	5	6	7	8	9	10
FC Antes:	140	165	160	160	175	190	170	175	155	160
FC Después:	145	150	150	160	170	175	160	165	145	170

3. Procedimiento

1) Obtener la variable diferencia (como en el test de Student). Si una diferencia es nula, se quita (y se corrige n).
2) Ordenar las diferencias por su magnitud pero clasificándolas por su signo
3) Aplicar el test de Wilcoxon para muestras independientes a las "muestras" de diferencias positivas y negativas
4) Interpretación. A partir de cómo se obtuvo la diferencia:

Paciente:	1	2	3	4	5	6	7	8	9	10
Antes – Desp.	−5	+15	+10	0	+5	+15	+10	+10	+10	−10

Ordenación	(+)			5		10	10	10	10	15	15	
$\lvert d_i \rvert$	(−)	5			10							
Num. orden		1	2	3	4	5	6	7	8	9	*Total*
Rangos	(+)		1.5		5	5	5	5	8.5	8.5	38.5
r_i	(−)	1.5		5							6.5



$R_{exp} = 6.5 \rightarrow 0.05 < p < 0.10$

Hay indicios de significación

Como es $R^+ > R^-$ el cambio sugerido es que se reduce la FC (ya que se ha hecho Antes–Después)

Comparativa entre las pruebas paramétricas y las no paramétricas

La clave es **"qué información se tiene y qué método la aprovecha mejor"**

1. Supuestos sobre los datos

Los supuestos mas básicos se deben mantener en ambos ⟶ la muestra debe ser aleatoria, las observaciones deben ser independientes

***Métodos paramétricos (MP)*:** Dependen que los supuestos sobre el **modelo** sean correctos y por tanto deben de comprobarse.

Métodos no paramétricos (MNP): No se especifica ningún **modelo** relativo a la distribución de la población cuyas observaciones componen la muestra. Son aplicables siempre.

2. Nivel de medida

MP: requieren mayor sofisticación de la **medida**. Por ejemplo, el modelo normal es para variables continuas.

MNP: No requieren mediciones tan 'fuertes'. En general, son **aplicables siempre** (sobre datos en escala ordinal, e incluso nominal)

3. Potencia del test

Los **MP** en general son **más potentes** que los MNP. Si se satisfacen los supuestos paramétricos, una prueba paramétrica requiere 95 datos por cada 100 que necesita una paramétrica para poder rechazar una hipótesis nula falsa.

4. Capacidad de análisis

MP permiten **mayor profundidad en el análisis** y resultados que suelen ser mas intuitivos (por ejemplo, diferencia de efecto, IC, etc)

MNP. Los métodos no paramétricos no ofrecen resultados tan intuitivos. No se puede cuantificar el tamaño del efecto, solo decir que una población da valores (de una escala) mayores que la otra.

Pedro Femia
pfemia@ugr.es

COMPLEMENTO

El peligro de comparar muchas variables a la vez

1. Planteamiento del problema

- El propósito es comparar dos grupos entre sí, pero a través de k variables (no una como hasta ahora)
- Una comparación individual se declara significativa cuando se obtenga $p<\alpha$ pero esto supone que por cada 100 comparaciones realizadas en $\alpha \times 100\%$ de ellas se dirá que hay diferencias por error
- Cuando se comparan k variables de forma simultánea, el error asociado a la afirmación global no será α sino mucho mayor. Concretamente

$$\alpha_{global} = 1-(1-\alpha)^k$$

2. Solución de Bonferroni

- Declarar cada comparación individual como significativa cuando sea $p < \alpha/k$

k	$1-(1-\alpha)^k$	α/k
2	0.0975	0.0250
3	0.1426	0.0167
4	0.1855	0.0125
5	0.2262	0.0100
6	0.2649	0.0083
7	0.3017	0.0071
8	0.3366	0.0063
9	0.3698	0.0056
10	0.4013	0.0050

3. Ejemplo

Una batería de 6 test de coordinación motora, TD-1 a TD-6, cada uno de los cuales mide una habilidad motora, se aplica a un total de 15 niños afectados de parálisis cerebral antes y después de un programa rehabilitación. Se trata de saber en cuál de las habilidades ha habido un cambio

H_0: En ninguno de los test de destreza hay cambio

Variable	t_{exp}	$g.l.$	p	k	6	3	2
				α/k	0.0083	0.0167	0.0250
TD-1	4.11	13	0.0012		SI	-	-
TD-2	2.91	11	0.0142		NS	SI	-
TD-3	6.86	12	0.0000		SI	-	-
TD-4	2.02	9	0.0741		NS	NS	NS
TD-5	4.18	12	0.0013		SI	-	-
TD-6	0.48	14	0.6386		NS	NS	NS

A nivel global del 5%, hay cambio en TD-1, TD-2, TD-3 y TD-5

- Bonferroni es un método muy severo (corrige demasiado)
- Si al aplicar el método de Bonferroni, se eliminan las variables que han resultado significativas y se repite con las que quedan (cambiando k en consecuencia) el método se llama de **Newman-Keuls**

Pedro Femia
pfemia@ugr.es

ESTADÍSTICA

Grado en Enfermería

Tema VI

Estudios comparativos con dos muestras

(2ª parte: comparación de dos proporciones)

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Estudios comparativos con dos muestras

Pruebas a realizar

Comparación de medias $\begin{cases} H_0 : \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$

Implicaba considerar:

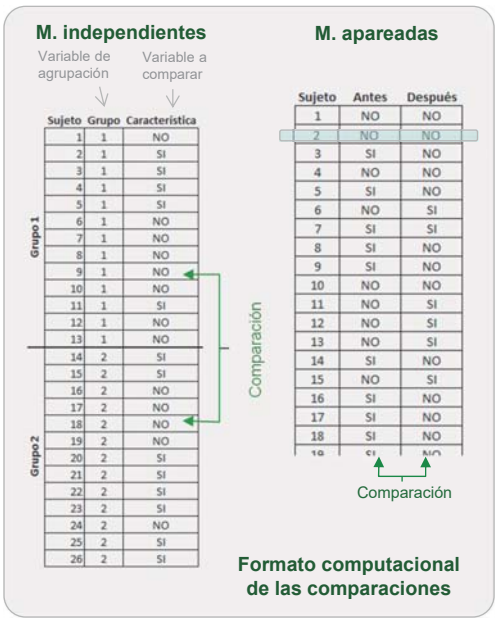
- El tipo de muestreo: muestras **independientes** o muestras **relacionadas**
- La **normalidad**
- La **homogeneidad de las varianzas** (si se verifica la normalidad)

Comparación de proporciones $\begin{cases} H_0 : p_1 = p_2 \rightarrow p_1 - p_2 = 0 \\ H_1 : p_1 \neq p_2 \rightarrow p_1 - p_2 \neq 0 \end{cases}$

La hipótesis de homogeneidad de las proporciones (p_1 = p_2) equivale a la de independencia de la variable binomial estudiada respecto a la variable dicotómica que divide a esta en dos grupos (con proporciones p_1 y p_2)

Implica considerar:

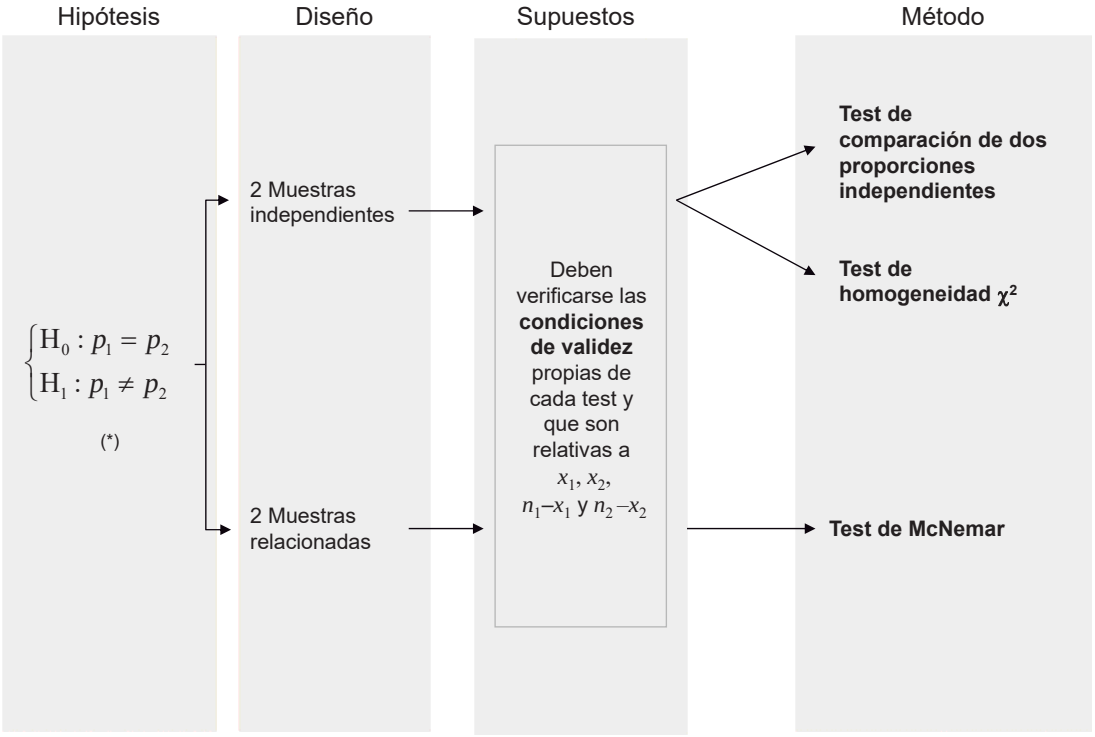
- El tipo de muestreo: muestras **independientes** o muestras **relacionadas**
- La inferencia con proporciones siempre requiere del cumplimiento de unas **condiciones de validez**



Formato computacional de las comparaciones

Estudios comparativos con dos muestras

Comparación de dos proporciones binomiales: esquema general



* O bien las alternativas unilaterales: $H_1: p_1 > p_2$; o $H_1: p_1 < p_2$

Estudios comparativos con dos muestras

Presentación de los datos

Observaciones

MUESTRAS INDEPENDIENTES

Característica

	Presente	Ausente	Total
Muestra I	x_1	$n_1 - x_1 = y_1$	n_1
Muestra II	x_2	$n_2 - x_2 = y_2$	n_2
Total	$a_1 = x_1 + x_2$	$a_2 = y_1 + y_2$	$N = n_1 + n_2$

MUESTRAS APAREADAS

Característica A \ Característica B	Presente	Ausente	
Presente	n_{11}	n_{12}	
Ausente	n_{21}	n_{22}	
			n

Porcentajes

Característica

	Presente	Ausente	Total
Muestra I	$\hat{p}_1 = x_1/n_1$	$\hat{q}_1 = 1 - p_1$	1
Muestra II	$\hat{p}_2 = x_2/n_2$	$\hat{q}_2 = 1 - p_2$	1

Característica A \ Característica B	Presente	Ausente	
Presente	$\hat{p}_{11} = n_{11}/n$	$\hat{p}_{12} = n_{12}/n$	$p_1 = \frac{n_{11} + n_{12}}{n}$
Ausente	$\hat{p}_{21} = n_{21}/n$	$\hat{p}_{22} = n_{22}/n$	
	$p_2 = \frac{n_{11} + n_{21}}{n}$		1

Ejemplo

Riesgo cardiovascular

	Alto	Normal	Total
Fumadores	58	25	**83**
No fumadores	19	28	**47**
Total	77	53	130

Respuesta a dos tratamientos, A y B, para paliar el dolor crónico en pacientes con fibromialgia

Tratamiento B

Tratamiento A \	Si	No	
Si	38	15	53
No	31	82	113
	69	97	**166**

Pedro Femia
pfemio@ugr.es

Se estudia la asociación entre el riesgo cardiovascular (medido en términos de indicadores como la calcificación coronaria, el índice aterogénico, etc) y el consumo de tabaco. Para ello se consideró una muestra de 105 fumadores de los cuales 45 resultaron ser individuos de alto riesgo cardiovascular. En otra muestra (control) de 119 individuos no fumadores se observaron a 15 individuos con alto riesgo cardiovascular. ¿Puede decirse que la presencia de riesgo cardiovascular alto es diferente en fumadores y no fumadores?

1. Identificación del problema:

   Se trata comparar la **proporción** de individuos con alto riesgo en la población de fumadores (=p_1) con la **proporción** de individuos de alto riesgo en la población de no fumadores (=p_2). Se han tomado dos **muestras independientes** correspondientes a cada una de ellas, por lo tanto se trata de realizar un **test de homogeneidad de dos proporciones independientes**

2. Información muestral:

   Fumadores: $n_1 = 105;\ x_1 = 45;\ n_1 - x_1 = 60$

   No fumadores: $n_2 = 119;\ x_2 = 15;\ n_2 - x_2 = 119$ $\longrightarrow$

   Riesgo cardiovascular

	Alto	Normal	Total
Muestra I: Fumadores	45	60	105
Muestra II: No fumadores	15	104	119
Total	60	164	224

   La estimación puntual de las proporciones de interés es:

   $\hat{p}_1 = 45/105 = 0.429$
   $\hat{p}_2 = 15/119 = 0.126$

   Puntualmente, el riesgo elevado se observa en el 42.9% de los fumadores y en el 12.6% de los no fumadores.

3. Hipótesis

   $\begin{cases} H_0: p_1 = p_2 & \text{La proporción de RC elevado es la misma en las dos poblaciones (el riesgo es independiente del tabaquismo)} \\ H_1: p_1 \neq p_2 & \text{En una de las poblaciones se dan más casos de individuos con alto RC (riesgo asociado del tabaquismo)} \end{cases}$

   Declararemos significativo al test si $P \leq \alpha = 0.05$

Estudios comparativos con dos muestras

Test de comparación de dos proporciones independientes

4. **Condiciones de validez** del test de comparación de dos proporciones independientes

   Riesgo cardiovascular

	Alto	Normal	Total
Muestra I: Fumadores	45	60	105
Muestra II: No fumadores	15	104	119
Total	60	164	224

   $$E = \frac{\min(60;164)\times\min(105;119)}{224} = \frac{60\times105}{224} = 28.1 \geq 7.7$$

   Cómo 28.1>7.7, el método que sigue a continuación es aplicable

   * Si hubiera sido N>500, debería verificarse E ≥14,9

   $\hat{p} = 60/224 = 0.268$
   $\hat{q} = 1 - \hat{p} = 0.732$

5. **Estadístico de contraste** (método incondicionado)

   $$z_{exp} = \frac{|\hat{p}_1 - \hat{p}_2| - c}{\sqrt{\hat{p}\hat{q}\dfrac{N}{n_1 n_2}}} \text{ con } c = 1/n_1 n_2 \text{ por ser } n_1 \neq n_2 \longrightarrow z_{exp} = \frac{|0.429 - 0.126| - 0.00016}{\sqrt{0.268\times0.732\dfrac{224}{105\times119}}} = 5.10$$

   * Si hubiera sido n_1=n_2, la *cpc* sería $c = 2/n_1 n_2$

   $\rightarrow$ P<0.001 (en la tabla de la distribución normal estándar)

   Suponiendo cierta la hipótesis (nula) de que la proporción de casos con alto riesgo cardiovascular es la misma en fumadores y no fumadores, la probabilidad de encontrar unos datos tan discrepantes o más con dicha hipótesis como los actuales es menor al 0.001 (menor al 1 por mil). Por tanto rechazamos dicha hipótesis asumiendo que las proporciones comparadas son diferentes

6. **Magnitud de la diferencia** (tamaño del *efecto*)

   Considerando el método de Agresti-Caffo:

   $$IC(p_1 - p_2) = \left(\frac{x_1 + h}{n_1 + 2h} - \frac{x_2 + h}{n_2 + 2h}\right) \pm z_\alpha \sqrt{\frac{(x_1 + h)(y_1 + h)}{(n_1 + 2h)^3} + \frac{(x_2 + h)(y_2 + h)}{(n_2 + 2h)^3}}$$

Pedro Femia
pfemia@ugr.es

6. Magnitud de la diferencia (continuación)

Riesgo cardiovascular

	Alto	Normal	Total
Muestra I: Fumadores	45	60	105
Muestra II: No fumadores	15	104	119
Total	60	164	224

Al 95% de confianza $\begin{cases} z_\alpha = 1.96 \\ h = z_\alpha^2 / 4 = 1.96^2 / 4 = 0.9604 \end{cases}$

$$95\% - IC\left(p_1 - p_2\right) = 0.2978 \pm 0.1145 = \left(0.183, 0.412\right)$$

Con un 95% de confianza, la proporción de casos con alto riesgo cardiovascular es un valor comprendido entre 18.3% y 41.2% veces mayor en fumadores que en no fumadores.

Observaciones:

- Vemos que el porcentaje de alto RC es mayor en fumadores: $\begin{cases} \hat{p}_1 = 42.9\% \\ \hat{p}_2 = 12.6\% \end{cases}$
- Como el test ha resultado significativo (P<0.001), el intervalo de confianza no contiene el valor propuesto por la hipótesis nula: $p_1 - p_2 = 0$.
- Como el resultado es **significativo**, no es necesario plantear un aumento del tamaño de muestra para detectar la diferencia entre las proporciones estudiadas. No obstante, si que se podría hacer tal consideración si lo que se desea es aumentar la precisión del intervalo obtenido para $\left(p_1 - p_2\right)$.
- El **test** también se podría haber resuelto mediante la **prueba χ^2** del capítulo siguiente a este.

Supongamos que los datos del ejemplo anterior hubieran sido estos →

Riesgo cardiovascular

	Alto	Normal	Total
Fumadores	11	25	36
No fumadores	8	40	48
Total	19	65	84

Test: (el planteamiento y las hipótesis son como antes)

- Validez: $E = \frac{19 \times 36}{284} = 8.1 \geq 7.7$ el método es válido
- Estadístico de contraste: $z_{\text{exp}} = 1.493$
- Significación: P = 0.135 > 0.05 (=α)

Suponiendo cierta la hipótesis nula de homogeneidad de las proporciones, la probabilidad de observar una discrepancia igual o mayor que la actual con dicha hipótesis es del 13.5%. Por tanto, para un error de tipo I fijado de antemano al 5%, no se puede rechazar la hipótesis nula de homogeneidad (no podemos decir que haya diferencia entre los porcentajes estudiados).

Como el nivel de significación no es inferior al 5% pero tampoco tiene una magnitud grande (es menor al 15%) todo apunta a que el test pueda estar resultando poco potente (decisión por H_0 no fiable).

IC(p_1−p_2) (método de Agresti-Caffo)

- Como se no se ha rechazado H_0 el interés de este intervalo es **estudiar la fiabilidad** de la decisión de aceptar la hipótesis nula.
- De forma rigurosa deberíamos fijar la diferencia a detectar δ y la potencia deseada para el test 1-β y obtener el IC al 1-2β de confianza. No lo hacemos y construimos el IC al 95% de confianza:

$$95\% - IC\left(p_1 - p_2\right) = 0.2923 \pm 0.1087 = \left(-0.0598, 0.3317\right)$$

- EL IC contiene al valor p_1-p_2=0 coherentemente con la aceptación de H_0 (aunque ¡por poco!)
- Observemos el límite más discrepante de 0: el 33.17%. El test considera que una diferencia del 33% ¡no es diferencia!

Pedro Femia
pfemia@ugr.es

Riesgo cardiovascular

	Alto	Normal	Total
Fumadores	11	25	36
No fumadores	8	40	48
Total	19	65	84

$$95\% - IC\left(p_1 - p_2\right) = 0.1359 \pm 0.1957 = \left(-0.0598, 0.3317\right)$$

Tamaño de muestra: sin considerar información previa (muestra piloto)

Para detectar una diferencia (un *efecto*) $|p_1 - p_2| = \delta$

con una potencia θ

a un error α

$$n_1 = n_2 \geq \frac{1}{2}\left(\frac{z_\alpha + z_{2\beta}\sqrt{1-\delta^2}}{\delta}\right)^2$$

Para declarar significativa una diferencia del 10% (δ=0.1) con una potencia del 80% (β=0.20) a un error del 5% (α=0.05) son necesarios 392 casos en cada muestra:

$$n_1 = n_2 \geq \frac{1}{2}\left(\frac{1.96 + 0.842\sqrt{1-0.1^2}}{0.1}\right)^2 = \frac{1}{2}\left(\frac{2.7978}{0.1}\right)^2 = 391.4 \rightarrow 392$$

$$\leftarrow \begin{cases} z_{0.05} = 1.96 \\ 2\beta = 0.4 \rightarrow z_{0.40} = 0.842 \\ \delta = 0.1 \end{cases}$$

Ejemplo: Se estudia la eficacia de dos terapias alternativas, A (analgésica) y B (terapia física) para paliar el dolor crónico en pacientes con fibromialgia. En el estudio han participado 166 mujeres que fueron sometidas durante dos periodos de 6 meses a cada uno de los métodos. El orden en que se aplicaron las terapias fue aleatorizado. Del conjunto de pacientes considerado, 38 respondieron a ambos métodos, 82 no respondieron a ninguno y 15 de los que respondieron al método A no respondieron al B ¿Puede considerarse una terapia más eficaz que la otra?

1. **Identificación del problema:**

   Cada paciente aporta dos observaciones: su respuesta a la terapia A y su respuesta a la terapia B, se trata por tanto de **muestras apareadas.** Como los parámetros de interés son las proporciones de casos que responden a cada terapia, el problema consiste en **contrastar la homogeneidad de dos proporciones de muestras apareadas**

2. **Información muestral:**

   Podríamos pensar en dos tipos de tabla:

Terapia

Respuesta	A	B
Si	53	69
No	113	97

No es la tabla que interesa, se pierde la información apareada: no es posible saber cuál es la respuesta a la terapia B por parte de un paciente que ha respondido positiva o negativamente a la A. Además, obsérvese que con este planteamento, la suma de frecuencias es 332 (el doble que el número de pacientes)

Terapia B

	Respuesta	Si	No	
Terapia A	Si	38	15	53
	No	31	82	113
		69	97	166

Este formato **anidado** (la respuesta esta anidada en cada modalidad de la terapia) si que permite conservar la información apareada y es por tanto el formato apropiado.

Pedro Femia
pfemia@ugr.es

3. Planteamiento de la inferencia. Hipótesis

$p_1 =$ Proporción de pacientes que responden al tto. A

$$\rightarrow \hat{p}_1 = \frac{n_{11}+n_{12}}{n} = \frac{n_{11}}{n} + \frac{n_{12}}{n} = \hat{p}_{11} + \hat{p}_{12}$$

$p_2 =$ Proporción de pacientes que responden al tto. B

$$\rightarrow \hat{p}_2 = \frac{n_{11}+n_{21}}{n} = \frac{n_{11}}{n} + \frac{n_{21}}{n} = \hat{p}_{11} + \hat{p}_{21}$$

Comparar p_1 con p_2 supone comparar p_{12} con p_{21}.

Terapia B

Terapia A	Si	No	
Si	$n_{11}=38$	$n_{12}=15$	53
No	$n_{21}=31$	$n_{22}=82$	113
	69	97	$n=166$

$$\hat{p}_1 = \frac{38+15}{166} = \frac{53}{166} = 0.3193$$

$$\hat{p}_2 = \frac{38+31}{166} = \frac{69}{166} = 0.4157$$

Los casos que superan ambas terapias (o los que no superan ninguna) no aportan información diferencial

$$\begin{cases} H_0: p_1 = p_2 \\ H_1: p_1 \neq p_2 \end{cases} \rightarrow \begin{cases} H_0: p_{12} = p_{21} \\ H_1: p_{12} \neq p_{21} \end{cases}$$

Por tanto, se trata de comparar si la proporción de casos que SI supera la terapia A y NO la B es igual a la proporción de casos que NO supera la A y SI la B: **Test de McNemar**

$$\hat{p}_{12} = \frac{15}{166} = 0.0904;\ \hat{p}_{21} = \frac{31}{166} = 0.1867$$

Declararemos el test significativo si $P \leq \alpha$=0.05

4. Condición de validez

Debe ser $n_{12}+n_{21} > 10$. Como $(n_{12}=15)+(n_{21}=31) = 46 > 10$ el método es aplicable.

Estudios comparativos con dos muestras

Test de comparación de dos proporciones Apareadas

5. Estadístico de contraste y nivel de significación

$$z_{exp} = \frac{|n_{12}-n_{21}|-0.5}{\sqrt{n_{12}+n_{21}}} = \frac{|15-31|-0.5}{\sqrt{15+31}} = 2.212 \longrightarrow 0.02 < P < 0.03$$

en la normal estándar ($P = 0.027$)

Terapia B

Terapia A / Respuesta	Si	No	
Si	38	15	53
No	31	82	113
	69	97	166

No puede aceptar la homogeneidad de las dos terapias. A una se responde más que a la otra (la probabilidad de error implícita en esta afirmación es menor al 3%):

$\hat{p}_1 = 31.93\%$

$\hat{p}_2 = 41.57\%$ ⟶ A la terapia B se responde en mayor proporción que a la A

6. Estimación de la diferencia de porcentajes (efecto diferencial de la terapia B respecto a la A)

$$IC_{1-\alpha}(p_1-p_2) = \frac{n_{12}-n_{21}}{n+2} \pm z_\alpha \sqrt{(n_{12}+n_{21}+1) - \frac{(n_{12}-n_{21})^2}{n+2}}$$

(método de Agresti-Min)

Para un nivel de confianza del 95% (z_α=1.96)

$$IC_{95\%}(p_1-p_2) = -0.0952 \pm 0.0787 = (-0.1739; -0.0166)$$

Obsérvese que como se ha rechazado la hipótesis de que los dos porcentajes sean homogéneos (iguales), la diferencia="cero" no está contenida en el intervalo

La terapia B tiene mayor respuesta positiva que la A, puntualmente la diferencia es de un 9.5% de pacientes que responden a la B más que a la A. Con un 95% de confianza, esta diferencia es un valor superior al 1.66% e inferior al 17.39%

Pedro Femia
pfemia@ugr.es

BIOESTADÍSTICA

Grado en Enfermería

Tema VII

Análisis de datos cualitativos

Análisis de tablas de contingencia – Test χ^2

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Análisis de datos cualitativos: el test χ^2

Introducción

- Se van a estudiar dos caracteres (variables) cualitativos que pueden presentarse bajo diferentes modalidades
- La forma de abordar tal estudio es mediante el **recuento** (V.A. discreta) de las veces que se presenta cada una de las modalidades en cada carácter
- Los *recuentos* (frecuencias absolutas) se deben normalizar en forma de ***proporciones*** (frecuencias relativas)
- El problema principal consiste en comprobar si las ***proporciones observadas*** (información empírica) difiere de forma significativa de aquella ***proporciones esperadas*** bajo el supuesto de una determinada teoría o de una hipótesis

El test χ^2

- Debido a Karl Pearson, es uno de los test mas clásicos de la Estadística
- Varias aplicaciones:

I - Ajuste a distribuciones

No lo vamos a estudiar aquí

  - Por ejemplo el ajuste de los datos empíricos a la distribución de frecuencias fenotípicas propuesta por la Teoría de Mendel
  - También hay un test de normalidad basado en la prueba χ^2 (obsoleto)

II - Análisis de ***tablas de contingencia***.

Nos centramos en este tipo de análisis

2 tipos de problemas:

- Test de **homogeneidad** entre varias muestras
- Test de **independencia** de dos variables cualitativas

- Son problemas diferentes
- En la práctica se resuelven igual

Homogeneidad (entre muestras)
⇓
Independencia (entre variables)

Pedro Femia
pfemia@ugr.es

Formato de los datos

Tablas de contingencia. Notación y terminología

- Una **tabla de contingencia** es una tabla de frecuencias en donde se dispone la información (cruzada) de al menos dos características cualitativas. Se habla también de **tablas de clasificación cruzada**.

Ejemplo 1

Se consideran **4 tratamientos** que curan una misma enfermedad. Se aplica cada uno de ellos a otras tantas **muestras independientes** de 150, 120, 130 y 160 enfermos respectivamente. Tras un periodo de seguimiento se anota el **estado de cada paciente** clasificándolo en *peor, igual* o *mejor* tal y como se presenta a la derecha.

	Estado del paciente tras el tratamiento			
Tratamiento	*Peor*	*Igual*	*Mejor*	*Total*
1	7	28	115	150
2	15	20	85	120
3	10	30	90	130
4	5	40	115	160

Ejemplo 2

En un estudio sobre tumores cerebrales se desea averiguar si existe asociación entre la **localización** del tumor y su **naturaleza**. A tal efecto se tomaron al azar **una muestra** de 141 pacientes afectados de tumor cerebral y se les clasificó como se indica en la tabla adjunta

	Naturaleza del tumor		
Localización del tumor	Benigno	Maligno	Otros
Lóbulo frontal	23	9	6
Lóbulo temporal	21	4	3
Otras áreas	34	24	17

141

Formato de los datos

Tablas de contingencia. Notación y terminología

Ejemplo 1

- 2 características:
  - Por filas: el tipo de tratamiento (4 categorías)
  - Por columnas: el estado del paciente tras el tto. (3 categorías)
- Se dice que es una tabla de contingencia 4×3
- En el interior de la tabla aparecen las frecuencias (absolutas) observadas de cada estado para cada tratamiento
- En este caso hay cuatro muestras independientes, tomadas en función de las modalidades del tratamiento: Se trata de un **test de homogeneidad** entre 4 muestras independientes

	Estado del paciente tras el tratamiento			
Tratamiento	*Peor*	*Igual*	*Mejor*	*Total*
1	7	28	115	150
2	15	20	85	120
3	10	30	90	130
4	5	40	115	160

Ejemplo 2

- 2 características:
  - Por filas: localización del tumor (3 categorías)
  - Por columnas: naturaleza del tumor (3 categorías)
- Se dice que es una tabla de contingencia 3×3
- En el interior de la tabla aparecen las frecuencias (absolutas) observadas de cada una de las modalidades de cada variable
- En este caso hay una sola muestra (no se puede hablar de homogeneidad): Se trata de un **test de independencia** entre dos variables cualitativas

	Naturaleza del tumor		
Localización del tumor	Benigno	Maligno	Otros
Lóbulo frontal	23	9	6
Lóbulo temporal	21	4	3
Otras áreas	34	24	17

141

Pedro Femia
pfemia@ugr.es

Tablas de contingencia. Notación y terminología

Tabla de contingencia $r \times c$ (= r filas y c columnas)

Variable B

Variable A	B_1	B_2	…	B_c	*Totales*
A_1	O_{11}	O_{12}	…	O_{1c}	F_1
A_2	O_{21}	O_{22}	…	O_{2c}	F_2
⋮	⋮	⋮	⋱	⋮	⋮
A_r	O_{r1}	O_{r2}	…	O_{rc}	F_r
Totales	C_1	C_2	…	C_c	T

O_{ij}	Frecuencia observada en la fila i y en la columna j
C_j	Total de la columna j
F_i	Total de la fila i
T	Total global

Porcentajes observados derivados de la tabla anterior → 3 tipos de tablas de porcentajes

Porcentajes por filas

	B_1	…	B_c	*Totales*
A_1	$p_{11} = O_{11} / F_1$	…	$p_{1C} = O_{1c} / F_1$	1
⋮	⋮	⋱	⋮	⋮
A_R	$p_{R1} = O_{r1} / F_r$	…	$p_{rc} = O_{rc} / F_r$	1
Totales	C_1 / T	…	C_c / T	1

Porcentajes por columnas

	B_1	…	B_C	*Totales*
A_1	$p_{11} = O_{11} / C_1$	…	$p_{1C} = O_{1C} / C_C$	F_1 / T
⋮	⋮	⋱	⋮	⋮
A_R	$p_{R1} = O_{R1} / C_1$	…	$p_{RC} = O_{RC} / C_C$	F_R / T
Totales	1	…	1	1

Porcentajes totales

	B_1	…	B_C	*Totales*
A_1	$p_{11} = O_{11} / T$	…	$p_{1C} = O_{1C} / T$	F_1 / T
⋮	⋮	⋱	⋮	⋮
A_R	$p_{R1} = O_{R1} / T$	…	$p_{RC} = O_{RC} / T$	F_R / T
Totales	C_1 / T	…	C_C / T	1

Ejemplo (Tabla de contingencia 2×2)

¿Esta asociado el peso del niño con el hábito de fumar de la madre?

Frecuencias observadas

Madre fumadora

Peso del niño al nacer	*Si*	*No*	*Total*
Bajo	20	26	46
Normal	60	294	354
Total :	80	320	400

Proporciones observadas

Madre fumadora

Peso del niño al nacer	*Si*	*No*	*Total*
Bajo	0.435	0.565	1
Normal	0.169	0.831	1
Total :	0.200	0.800	1

Del total de niños con el peso bajo al nacer, el 43.5% son hijos de madres fumadoras

Madre fumadora

Peso del niño al nacer	*Si*	*No*	*Total*
Bajo	0.250	0.081	0.115
Normal	0.750	0.919	0.885
Total :	1	1	1

Del total de madres fumadoras, el 25% han tenido un niño con el peso demasiado bajo

Madre fumadora

Peso del niño al nacer	*Si*	*No*	*Total*
Bajo	0.050	0.065	0.115
Normal	0.150	0.735	0.885
Total :	0.200	0.800	1

Del total de madres, el 5% eran fumadoras y su hijo ha tenido el peso demasiado bajo

Pedro Femia
pfemia@ugr.es

Test de homogeneidad con tablas de contingencia (distintas a 2×2)

Valores observados e Hipótesis

Estado del paciente tras el tratamiento

	Peor	*Igual*	*Mejor*	*Total*
Tratamiento 1	7	28	115	150
Tratamiento 2	15	20	85	120
Tratamiento 3	10	30	90	130
Tratamiento 4	5	40	115	160

Hipótesis

H_0: Los cuatro tratamientos son equivalentes (homogéneos)

H_1: Al menos un tratamiento es diferente a los demás

¿Qué supone H_0?

Porcentajes por filas

(según se han dispuesto las muestras)

	Peor	*Igual*	*Mejor*	*Total*
1	p_{11}	p_{12}	p_{13}	1
2	p_{21}	p_{22}	p_{23}	1
3	p_{31}	p_{22}	p_{33}	1
4	p_{41}	p_{32}	p_{43}	1

$H_0: p_{11} = p_{21} = p_{31} = p_{41}$

$p_{12} = p_{22} = p_{32} = p_{42}$

La proporción de los que empeoran es la misma en todos los ttos.

La proporción de los que quedan igual es la misma en todos los ttos.

La proporción de los que mejoran es también la misma. Va obligada, para cada fila i:

$$p_{i3} = 1 - (p_{i1} + p_{i2})$$

¿Qué forma tendá H_1?

H_1: al menos un $p_{ij} \neq p_{kj}$ para $i \neq k$

Test de homogeneidad con tablas de contingencia (distintas a 2×2)

¿Contra qué se comparan las frecuencias observadas?

Estado

Tratamiento	*Peor*	*Igual*	*Mejor*	*Totales*
1	$7 = O_{11}$	$28 = O_{12}$	$115 = O_{13}$	$150 = F_1$
2	$15 = O_{21}$	$20 = O_{22}$	$85 = O_{23}$	$120 = F_2$
3	$10 = O_{31}$	$30 = O_{32}$	$90 = O_{33}$	$130 = F_3$
4	$5 = O_{41}$	$40 = O_{42}$	$115 = O_{43}$	$160 = F_4$
Totales	$37 = C_1$	$118 = C_2$	$405 = C_3$	$560 = T$

Frecuencias esperadas (E_{ij})

(Suponiendo cierta H_0 y manteniendo fijos los totales observados)

$$E_{ij} = \frac{F_i C_j}{T}$$

$$\Downarrow$$

$$\left(E_{11} = \frac{F_1 C_1}{T} = \frac{(150)(37)}{560} = 9.91; \ldots \right)$$

Estado

Tratamiento	*Peor*	*Igual*	*Mejor*	*Totales*
1	$E_{11} = 9.91$	$E_{12} = 31.61$	$E_{13} = 108.48$	$F_1 = 150$
2	$E_{21} = 7.93$	$E_{22} = 25.28$	$E_{23} = 86.79$	$F_2 = 120$
3	$E_{31} = 8.59$	$E_{32} = 27.39$	$E_{33} = 94.02$	$F_3 = 130$
4	$E_{41} = 10.57$	$E_{42} = 33.72$	$E_{43} = 115.71$	$F_4 = 160$
Totales	$C_1 = 37$	$C_2 = 118$	$C_3 = 405$	$T = 560$

Pedro Femia
pfemia@ugr.es

Test de homogeneidad con tablas de contingencia (distintas a 2×2)

Condiciones de validez (la prueba χ^2 no es aplicable siempre)

- No puede haber frecuencias esperadas demasiado pequeñas.

> El test solo **es válido si** se verifica:
> - Ninguna E_{ij} es inferior a 1
> - No mas del 20% de las E_{ij} son inferiores o iguales a 5

- **En el ejemplo:**

Frecuencias esperadas bajo H_0

		Estado		
	E_{ij}	*Peor*	*Igual*	*Mejor*
Trat.	1	9.91	31.61	108.48
	2	7.93	25.28	86.79
	3	8.59	27.39	94.02
	4	10.57	33.72	115.71

$E_{ij} > 5$ para todos los $i, j \Rightarrow$ el método es válido

Se admitirían 2 valores $1 < E_{ij} \le 5$, pero no 3 $(12 \times 20\% = 2.4)$

Test de homogeneidad con tablas de contingencia (distintas a 2×2)

Obtención de la cantidad experimental. El estadístico χ^2

$$\chi^2_{exp} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j \frac{O_{ij}^2}{E_{ij}} - T$$

$$\begin{cases} \chi^2_{exp} \downarrow \Rightarrow O_{ij} \approx E_{ij} \Rightarrow P \uparrow \quad (H_0) \\ \chi^2_{exp} \uparrow \Rightarrow \uparrow \text{ discrepancia } O_{ij} \text{ vs } E_{ij} \Rightarrow P \downarrow \quad (H_1) \end{cases}$$

En el ejemplo: $\chi^2_{exp} = \frac{7^2}{9.91} + \frac{28^2}{31.61} + \cdots + \frac{115^2}{115.71} - 560 = 13.87$

Obtención del nivel de significación

En la distribución χ^2 con los grados de libertad:

$$g.l. = (n^o \text{ de filas} - 1) \times (n^o \text{ de columnas} - 1)$$

En el ejemplo $P < 0.05$ por lo tanto no se puede asumir que todos los tratamientos sean iguales (homogéneos). Al menos hay uno diferente.

$0.025 < P < 0.050$

g.l. \ α	0,200	0,100	0,050	0,025	0,010	0,005	0,001
1	1,642	2,706	3,842	5,024	6,637	7,905	10,808
2	3,219	4,604	5,995	7,379	9,220	10,589	13,690
3	4,642	6,252	7,817	9,356	11,325	12,819	16,291
4	5,989	7,782	9,492	11,150	13,280	14,824	18,431
5	7,291	9,237	11,073	12,838	15,087	16,762	20,750
6	8,559	10,646	12,596	14,459	16,810	18,549	22,676
7	9,804	12,020	14,070	16,020	18,470	20,270	24,526

$\chi^2_{exp}(6\,gl) = 13.87$

Análisis de las causas de significación

El test χ^2 es un test ***omnibus***, es decir, permite detectar si algún tratamiento es diferente al resto, pero no dice ni cuantos ni cuales son distintos. En caso de ser significativo, las causas de la significación se pueden analizar mediante diferentes métodos:

- Análisis de la tabla de **porcentajes** →método fácil, pero relativamente subjetivo
- Análisis de los ***residuos estandarizados*** → Los da SPSS; método objetivo y muy intuitivo
- **Partición de tablas** (reducción a tablas 2×2) → Método potente, permite obtener mucha información, es objetivo pero laborioso)

Pedro Femia
pfemia@ugr.es

Test de homogeneidad con tablas de contingencia (distintas a 2×2)

Análisis de las causas de significación. Tabla de porcentajes

Frecuencias observadas		Resultado				
		Peor	Igual	Mejor	Total	
Tratamiento	Tratamiento 1	7	28	115	150	n_1
	Tratamiento 2	15	20	85	120	n_2
	Tratamiento 3	10	30	90	130	n_3
	Tratamiento 4	5	40	115	160	n_4
Total		37	118	405	560	

Proporciones observadas

Proporción de *Peor* con el
Tratamiento 1: 7/150=0.047 (4.7%)
Tratamiento 2: 15/120=0.125 (12.5%)
Tratamiento 3: 10/130=0.077 (7.7%)
Tratamiento 4: 5/160=0.031 (3.1%)

Proporción de *Igual* con el
Tratamiento 1: 28/150=0.187 (18.7%)
Tratamiento 2: 20/120=0.167 (16.7%)
Tratamiento 3: 30/130=0.231 (23.1%)
Tratamiento 4: 40/160=0.250 (25.0%)

Proporción de *Mejor* con el
Tratamiento 1: 115/150=0.767 (76.7%)
Tratamiento 2: 85/120=0.708 (70.8%)
Tratamiento 3: 90/130=0.692 (69.2%)
Tratamiento 4: 115/160=0.719 (71.9%)

$H_0 \Rightarrow$ Estos porcentajes son iguales (Los cuatro tratamientos *empeoran* por igual) y estos también (y dejan *igual*, por igual) …. y estos también (y *mejoran* por igual)

Como $P<0.05$ no podemos aceptar H_0, es decir, no se pueden considerar equivalentes los tratamientos, entonces,

¿qué tratamiento(s) es (o son) distinto(s) al resto?

… antes de seguir, a la vista de la tabla que sigue a continuación, de porcentajes por filas, **¿qué tratamiento usaría usted?**

% de Tratamiento

		Resultado			
		Peor	Igual	Mejor	Total
Tratamiento	Tratamiento 1	4.7%	18.7%	76.7%	100.0%
	Tratamiento 2	12.5%	16.7%	70.8%	100.0%
	Tratamiento 3	7.7%	23.1%	69.2%	100.0%
	Tratamiento 4	3.1%	25.0%	71.9%	100.0%
Total		6.6%	21.1%	72.3%	100.0%

Test de homogeneidad con tablas de contingencia (distintas a 2×2)

Análisis de las causas de significación. Tabla de porcentajes (continuación)

La tabla **de porcentajes por filas** permite detectar qué porcentajes son los mas dispares (y en principio allí radica la significación del test)

% de Tratamiento

		Resultado			
		Peor	Igual	Mejor	Total
Tratamiento	Tratamiento 1	4.7%	18.7%	76.7%	100.0%
	Tratamiento 2	12.5%	16.7%	70.8%	100.0%
	Tratamiento 3	7.7%	23.1%	69.2%	100.0%
	Tratamiento 4	3.1%	25.0%	71.9%	100.0%
Total		6.6%	21.1%	72.3%	100.0%

La discrepancia mas relevante se muestra en el % de '**peor**' correspondiente a los tratamientos 2 (es el que mas empeora) y 4 (el que menos empeora)

Las proporciones observadas de '**igual**' parecen definir dos grupos: los tratamientos 1 y 2 son similares y los 3 y 4 también

La % de **mejorías** no es muy dispar, ronda en todos los casos el 70%

Entonces, ¿qué tratamiento usaría usted?

No se puede negar que el método es un tanto subjetivo, pero debe contemplarse como una primera aproximación. El estudio en profundidad de las causas de significación en las tablas de contingencia se hace mediante los otros métodos indicados

Pedro Femia
pfemia@ugr.es

Test de independencia con tablas de contingencia (distintas a 2×2)

Naturaleza del tumor

Localización del tumor	Benigno	Maligno	Otros	
Lóbulo frontal	23	9	6	
Lóbulo temporal	21	4	3	
Otras áreas	34	24	17	
				141

Ahora solo hay **una muestra** (no cuatro como antes). No se puede hablar de homogeneidad.

Antes se trataba de la ***homogeneidad*** *entre muestras*, ahora de la ***independencia*** *entre variables*

Hipótesis:

H_0: La naturaleza del tumor es **independiente** de su localización

H_1: La naturaleza del tumor **esta asociada** a su localización

Cantidades esperadas
(suponiendo cierta H_0)

El fundamento para obtenerlas es diferente que en el test de homogeneidad, pero al final la *receta* es la misma, así que no cambia nada: $E_{ij}=F_i\times C_j/T$

Naturaleza

Localización	Benigno	Maligno	Otros	*Totales*
Lóbulo frontal	$23\,(=O_{11})$ $(E_{11}=)21.02$	$9\,(=O_{12})$ $(E_{12}=)9.97$	6 7.01	$38=F_1$
Lóbulo temporal	21 15.49	4 7.35	3 5.16	$28=F_2$
Otras áreas	34 41.49	24 19.68	$17\,(=O_{33})$ $(E_{33}=)13.83$	$75=F_3$
Totales	$78=C_1$	$37=C_2$	$26=C_3$	$141=T$

Test de independencia con tablas de contingencia (distintas a 2×2)

Validez del método

$E_{ij}>5$ para todos los $i, j \Rightarrow$ el método es válido

Se admitiría un valor $1<E_{ij}\leq 5$ pero no dos ($9\times 20\% = 1.8$)

Igual que en el test de homogeneidad

Estadístico de contraste:

$$\chi^2_{\exp}=\sum_i\sum_j\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\sum_i\sum_j\frac{O_{ij}^2}{E_{ij}}-T=$$

$$=\frac{23^2}{21.02}+\frac{9^2}{9.77}+\cdots+\frac{17^2}{13.83}-141=7.64$$

Igual que en el test de homogeneidad

Nivel de significación:

En la tabla χ^2 con $(3\text{-}1)\times(3\text{-}1)=4$ *g.l.* $\Rightarrow$ $\boxed{0.10<P<0.20}$

Aunque no se puede rechazar la hipótesis nula, hay indicios de significación. Un aumento de tamaño de muestra sería recomendable para ganar potencia y dar fiabilidad al resultado

En caso de haber obtenido significación, los porcentajes se pueden hacer **por filas** o **por columnas** (como mas intuitivo resulte)

Pedro Femia
pfemia@ugr.es

En la práctica (análisis de tablas de contingencia "grandes")

En el análisis de tablas de contingencia deberían ser consideradas las siguientes etapas:

1. Identificar si es un test de homogeneidad o un test de independencia (el número de muestras presentes en el estudio lo debe de dejar claro (1 muestra→ indep; 2 o más → homogeneidad)

2. Formular las hipótesis (**homogeneidad** o **independencia** en H_0)

3. Si la tabla es 2x2 → **NO seguir** este esquema, acudir al apartado dedicado a tablas de este tipo

   Si la tabla es "más grande" de 2x2 (continuar)

4. Obtener las cantidades esperadas (suponiendo cierta la H_0) $E_{ij} = \frac{F_i C_j}{T}$

5. Comprobar las condiciones de validez (su cumplimiento se debe indicar de forma explícita)

6. Si el método es aplicable, obtener el estadístico de contraste $\chi^2_{\exp} = \sum_i \sum_j \frac{O_{ij}^2}{E_{ij}} - T$

7. Obtener el nivel de significación P e interpretarlo (como siempre):

   Si $P \leq \alpha \longrightarrow$ Significativo (la muestra es improbable bajo H_0)

   Si $P > \alpha \longrightarrow$ No Significativo (no se puede rechazar H_0)

   Habitualmente α al 5% y si $0.05<P<0.20$ la aceptación de H_0 no es fiable y se debe plantear un aumento del tamaño muestral (regla automática tradicional)

8. Si el test ha resultado significativo, obtener una **tabla de porcentajes** que sea adecuada para estudiar las causas de dicha significación (comentar detalladamente el resultado)

EL CASO PARTICULAR DE LAS TABLAS DE CONTINGENCIA 2×2

Introducción

- Las tablas 2x2 surgen cuando las dos cualidades son de **tipo dicotómico** (o binario)
- Permiten resolver problemas de gran interés en el ámbito biosanitario (en principio, toda variable cuantitativa es *dicotomizable*). Consideraremos en particular la relación entre la exposición a un determinado **factor de riesgo** (FR) y la aparición de una **enfermedad** (E) (en términos generales, se aludirá a las variables ***factor*** y ***respuesta*** respectivamente)
- Permiten obtener "buenas" **medidas de la intensidad de la asociación** entre las variables implicadas
- Aunque el planteamiento general es el mismo que para las tablas "mas grandes", las tablas 2×2 requieren un tratamiento especial**: la fórmula general del estadístico de contraste χ^2 ¡NO ES APLICABLE!**

Disposición de la información y notación asumidas en las tablas 2x2:

- Una tabla 2x2 puede presentarse de 8 formas distintas (cambiando filas por columnas, filas entre sí o columnas entre sí). → Es necesario adoptar un **formato** para facilitar la formulación implicada y la interpretación del análisis.

Se va a considerar el siguiente **formato estándar**

- **Variable respuesta** (la enfermedad) por filas. Modalidad de mayor interés (SI enfermo) en la primera fila
- **Variable Factor** (el factor de riesgo) por columnas. Modalidad que induce a tener la enfermedad en la primera columna

Frecuencias observadas

Factor de Riesgo / ***Enfermedad***	***Expuesto*** FR	***No expuesto*** $\overline{FR}$	*Total*
Presente E	O_{11}	O_{12}	F_1
Ausente $\overline{E}$	O_{21}	O_{22}	F_2
Total :	C_1	C_2	T

Pedro Femia
pfemia@ugr.es

Análisis de tablas de contingencia 2x2

Ejemplo:

Hábito de Fumar de la madre

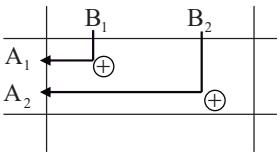
Peso del niño al nacer		*Si* (FR)	*No* $(\overline{FR})$	*Total*
	Bajo (E)	20	26	46
	Normal $(\overline{E})$	60	294	354
	Total :	80	320	400

Observaciones:

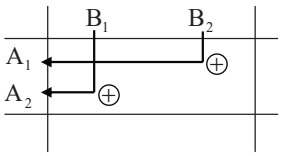
- La **enfermedad** (= *respuesta*) esta por **filas**, y la primera fila es la categoría que interesa estudiar (*peso=Bajo*)
- El **factor de riesgo** (= *factor*) se dispone por **columnas**, y la primera columna es la categoría que, en caso de asociación significativa, induce a tener la enfermedad (*Fumar=Si*)

Tipos de asociación: dos variables dicotómicas se pueden ordenar de forma arbitraria, así pues, en este punto conviene entender el tipo de asociación como una característica alusiva a la forma en que se ha construido la tabla 2x2

Asociación positiva
(la primera columna induce a estar en la primera fila)



Asociación negativa
(la primera columna induce a estar en la segunda fila)



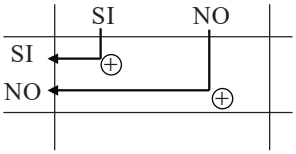
- Es mas fácil e intuitiva la interpretación de los resultados cuando la tabla se construye de forma que, en caso de haber asociación significativa, esta sea de tipo **positivo**

Análisis de tablas de contingencia 2x2

El formato de tabla más conveniente

Exposición al Factor de riesgo

Presencia de la enfermedad



Si hay asociación, mejor que sea positiva **(la primera columna =FR⁺ induce a estar en la primera fila =E⁺)**

A veces no es obvio qué categoría constituye el factor de riesgo (por ejemplo, si la variable es el sexo). Si se hace la tabla de **porcentajes por columnas**, se ve enseguida qué categoría está asociada de forma positiva con la enfermedad en caso de que haya significación

Si	*No*
$O_{11}=20$	$O_{12}=26$
$O_{21}=60$	$O_{22}=294$
$80=C_1$	$320=C_2$

→ Peso del niño al nacer

Madre fumadora

	Si	*No*	*Total*
Bajo	0.250	0.081	0.115
Normal	0.750	0.919	0.885
Total :	1	1	1

Del total de madres fumadoras, el 25% han tenido un niño con el peso demasiado bajo

Del total de madres no fumadoras, el 8% han tenido un niño con el peso demasiado bajo

Conclusión: 25% > 8%, es decir que el problema se da más en las madres fumadoras, es decir, si las variables están asociadas, fumar=si es realmente la modalidad de riesgo para que el niño tenga el peso bajo al nacer

Pedro Femia
pfemia@ugr.es

Análisis de tablas de contingencia 2x2

Tipos de muestreo o tipos de estudio epidemiológicos

El **tipo de muestreo** (o de **estudio**) no es un aspecto exclusivo de las tablas 2×2, pero para proceder a su análisis es preciso tenerlo en cuenta

La tabla considerada puede responder a tres tipos de muestreo
Obsérvese que la tabla es la misma. Cambia quién o quiénes son las muestras.

Fuma / *Peso*	*Si* (FR)	*No* $(\overline{FR})$	*Total*
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal $(\overline{E})$	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

M

Tipo I - Estudio transversal. Se toman T (=400) individuos al azar y se clasifican en base a la presencia o ausencia de la enfermedad y su exposición o no al factor de riesgo.

Hay **una sola muestra**
De ella se observa cuántos enfermos y cuantos expuestos hay

Fuma / *Peso*	*Si* (FR)	*No* $(\overline{FR})$	*Total*
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal $(\overline{E})$	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

M_1 M_2

Tipo II - a) Estudio prospectivo, longitudinal o de cohortes. Se toman C_1 (=80) y C_2 (=320) individuos al azar (muestras M_1 y M_2) y se clasifican en base a la presencia o no de la enfermedad.

Hay **dos muestras independientes**:
Una de **expuestos** y otra de **no expuestos**
De cada muestra se observan los enfermos y no enfermos

Fuma / *Peso*	*Si* (FR)	*No* $(\overline{FR})$	*Total*
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal $(\overline{E})$	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

M_1 M_2

Tipo II - b) Estudio Retrospectivo o de Caso-Control. Se toma F_1 (=46) y F_2 (=354) individuos al azar (muestras M_1 y M_2) y se clasifican en base a la exposición al factor de riesgo.

Hay **dos muestras independientes**
Una de **enfermos** y otra de **no enfermos**
De cada muestra se observan los expuestos y no expuestos

Análisis de tablas de contingencia 2x2

Test χ^2 en tablas 2x2

Factor de Riesgo / *Enfermedad*	*Expuesto* FR	*No expuesto* $\overline{FR}$	*Total*
Presente E	O_{11}	O_{12}	F_1
Ausente $\overline{E}$	O_{21}	O_{22}	F_2
Total :	C_1	C_2	T

- **Hipótesis:**
  - H_0: **Homogeneidad** entre grupos o **Independencia** entre variables, según el tipo de muestreo
  - H_1: Heterogeneidad o Asociación

- **Estadístico de contraste:**

$$\chi^2_{exp}=\frac{\left(\left|O_{11}O_{22}-O_{12}O_{21}\right|-c\right)^2}{F_1\,F_2\,C_1\,C_2}T$$

ATENCIÓN: **No** es válida la expresión general del estadístico χ^2 usada con tablas "grandes"

Estudio	Muestra(s)	Condición de validez*	Corrección** c
Transversal	T	$E\geq 6.2$ $(E\geq 3.9$ si $T\leq 500)$	$c=0.5$
Prospectivo	C_1 y C_2	$E\geq 14.9$ $(E\geq 7.7$ si $T\leq 500)$	$c=\begin{cases}1 & \text{si } C_1\neq C_2\\ 2 & \text{si } C_1=C_2\end{cases}$
Retrospectivo	F_1 y F_2		$c=\begin{cases}1 & \text{si } F_1\neq F_2\\ 2 & \text{si } F_1=F_2\end{cases}$

* Con $E=\dfrac{\text{Mín}\left(F_1;F_2\right)\text{Mín}\left(C_1;C_2\right)}{T}$ (que es la frecuencia mínima esperada)

La condición clásica es que fuera $E\geq 5$ independientemente del tipo de estudio. Es incorrecto

** Obsérvese que si el estudio es transversal c=0.5, en caso contrario es c=1 si los tamaños de muestra son distintos o c=2 si son iguales

- **Grados de libertad:** En tablas 2x2, siempre $g.l.$=1

Pedro Femia
pfemia@ugr.es

Test χ^2 en tablas 2x2

Resolución del ejemplo

Si el estudio es de tipo:

La frecuencia mínima esperada es $E = \dfrac{\text{Mín}(46;354)\,\text{Mín}(80;320)}{400} = 9.2$

- Transversal

Fuma / Peso	Si (FR)	No ($\overline{FR}$)	Total
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal ($\overline{E}$)	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

$\begin{cases} \mathbf{H_0}: \text{El peso del niño es independiente del hábito de fumar de la madre} \\ \mathbf{H_1}: \text{Hay asociación entre ambos caracteres} \end{cases}$

Validez: $T<500; E>3.9$

$$\chi^2_{exp} = \frac{\left(\left|(20)(294)-(26)(60)\right|-0.5\right)^2}{(46)(354)(80)(320)}400 = \mathbf{17.903} \qquad P<0.001$$

- Prospectivo

Fuma / Peso	Si (FR)	No ($\overline{FR}$)	Total
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal ($\overline{E}$)	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

$\begin{cases} \mathbf{H_0}: \text{La proporción de niños con peso bajo es la misma en madres fumadoras que en no fumadoras} \\ \mathbf{H_1}: \text{La proporción no es la misma (por tanto hay asociación entre ambos caracteres)} \end{cases}$

Validez: $T<500; E>7.7$

$$\chi^2_{exp} = \frac{\left(\left|(20)(294)-(26)(60)\right|-1\right)^2}{(46)(354)(80)(320)}400 = \mathbf{17.899} \qquad P<0.001$$

- Retrospectivo

Fuma / Peso	Si (FR)	No ($\overline{FR}$)	Total
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal ($\overline{E}$)	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

$\begin{cases} \mathbf{H_0} \equiv \text{La proporción de madres fumadoras con hijos de bajo peso es igual a la proporción de madres fumadoras en el grupo de niños de peso normal} \\ \mathbf{H_1} \equiv \text{La proporción no es la misma y por tanto hay asociación entre ambos caracteres} \end{cases}$

Validez: $T<500; E>7.7$

$$\chi^2_{exp} = \frac{\left(\left|(20)(294)-(26)(60)\right|-1\right)^2}{(46)(354)(80)(320)}400 = \mathbf{17.899} \qquad P<0.001$$

Medidas de asociación en tablas de contingencia 2x2

¿Es muy peligroso que la madre fume para que el niño tenga el peso bajo al nacer?

- Si el test es significativo, P no indica la intensidad de la **asociación** entre las variables (no indica lo peligroso que es que la madre fume para que el recién nacido tenga el peso demasiado bajo)

  P solo indica la magnitud de la evidencia muestral a favor (si es alto) o en contra (si es bajo) de H_0

- Una **medida de asociación** es un indicador de la fuerza o intensidad de la asociación entre las variables

- Las mejores medidas para cuantificar la asociación entre variables categóricas son las definidas sobre las tablas 2×2

- Toda medida de asociación va a tener un valor que es indicador de la falta de asociación o independencia

- Las **diferentes medidas** se pueden definir:
  - **Como cocientes** de las frecuencias o de las proporciones observadas

    En este caso, si la medida vale 1 indica que no hay asociación (en el test χ^2 será P>0.05)

    Son medidas de este tipo: El **riesgo relativo** y la **razón de producto cruzado** (*odds ratio*)
  - **Como diferencias** de las frecuencias o de las proporciones observadas

    En este caso, si la medida vale 0 indica que no hay asociación (en el test χ^2 será P>0.05)

    Son medidas de este tipo: el **riesgo absoluto** (diferencia de Berkson) y el **riesgo atribuible**

Pedro Femia
pfemia@ugr.es

Riesgo absoluto (diferencia de Berkson)

Definición: $d = P(E \mid FR) - P(E \mid \overline{FR}) = p_1 - p_2$

Indica cuánto es mayor (o menor) la probabilidad de tener la enfemedad cuando se está expuesto al factor de riesgo (p_1) que cuando no se está (p_2). Es la diferencia entre dos proporciones independientes ya vista en el capítulo anterior

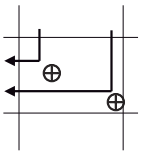
	FR	$\overline{FR}$	Total
E	O_{11}	O_{12}	*
*	*	*	*
Total	C_1	C_2	*

$\hat{p}_1 = \frac{O_{11}}{C_1}$ $\hat{p}_2 = \frac{O_{12}}{C_2}$

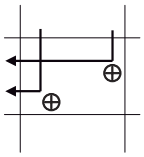
Estimador puntual (clásico):

$$\hat{d} = \hat{p}_1 - \hat{p}_2 = \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2} = \frac{O_{11}O_{22} - O_{12}O_{21}}{C_1 C_2}$$

Asociación positiva $d>0$



Asociación negativa $d<0$



Valores posibles: $-1 \le d \le +1$

$\begin{cases} -1 \le d < 0 & Asoc.(-) \longrightarrow P<0.05 \text{ en el test } \chi^2 \\ d = 0 & Independencia \longrightarrow P>0.05 \text{ en el test } \chi^2 \\ 0 < d \le 1 & Asoc.(+) \longrightarrow P<0.05 \text{ en el test } \chi^2 \end{cases}$

En el ejemplo:

Fumar / Peso	Si	No	Total
Bajo	$O_{11}=20$	$O_{12}=26$	$F_1=46$
Normal	$O_{21}=60$	$O_{22}=294$	$F_2=354$
Total	$C_1=80$	$C_2=320$	$T=400$

$$\hat{d} = \frac{20}{80} - \frac{26}{320} = 0.169$$

Puntualmente, la probabilidad de que el recién nacido tenga un peso bajo en madres fumadoras supera a la correspondiente a madres no fumadoras un 16.9% (estimador puntual clásico)

Tipo de estudios en que se puede estimar
- Transversales
- Prospectivos

En los estudios **retrospectivos** no se puede estimar (ello es debido a que este tipo de estudios no permiten estimar la prevalencia de la enfermedad P(E)

Riesgo absoluto (diferencia de Berkson)

Intervalo de confianza por el método de Agresti-Caffo

Frecuencias originales

Fuma / Peso	Si (FR)	No ($\overline{FR}$)	Total
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal ($\bar{E}$)	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

$h = z_\alpha^2/4$

$\alpha = 0.05$ ⇓

$h = 1.96^2/4 = 0.9216$

Frecuencias + h

Fuma / Peso	Si (FR)	No ($\overline{FR}$)	Total
Bajo (E)	$O_{11}=20.96$	$O_{12}=26.96$	$46.92=F_1$
Normal ($\bar{E}$)	$O_{21}=60.96$	$O_{22}=294.96$	$355.92=F_2$
Total	$81.92=C_1$	$320.92=C_2$	$403.84=T$

$$\hat{d} = \hat{p}_1 - \hat{p}_2 = \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2} = \frac{O_{11}O_{22} - O_{12}O_{21}}{C_1 C_2} = \frac{20.96\times294.96 - 26.96\times60.96}{81.92\times320.92} = 0.1721$$

Estimación puntual por el método de Agresti-Caffo

$$SE(\hat{d}) = \sqrt{\frac{O_{11}O_{21}}{C_1^3} + \frac{O_{12}O_{22}}{C_2^3}} = \sqrt{\frac{20.96\times60.96}{81.92^3} + \frac{26.96\times294.96}{320.92^3}} = 0.0506$$

$$IC_\alpha(d) = \hat{d} \pm z_\alpha SE(\hat{d}) \longrightarrow IC_{0.05}(d) = 0.1721 \pm 1.96\times0.0506 = 0.1721 \pm 0.0992$$

$$IC_{0.05}(d) = (0.0729,\ 0.2713)$$

Con un 95% de confianza, la probabilidad de que el niño tenga el peso bajo al nacer aumenta en un valor que debe de estar comprendido entre el 7.3% y el 27.1% cuando la madre es fumadora respecto a cuando no lo es.

En coherencia con la significación obtenida en el test χ^2 ($P<\alpha=0.05$): $0 \notin IC_{0.05}(d)$

Pedro Femia
pfemia@ugr.es

Riesgo relativo

Definición: $R = \dfrac{P(E \mid FR)}{P(E \mid \overline{FR})} = \dfrac{p_1}{p_2}$

Indica cuántas veces es mayor (o menor) la probabilidad de tener la enfemedad cuando se está expuesto al factor de riesgo (p_1) que cuando no se está (p_2)

	FR	$\overline{FR}$	Total
E	O_{11}	O_{12}	*
*	*	*	*
Total	C_1	C_2	*

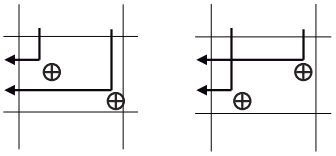
$\hat{p}_1 = \dfrac{O_{11}}{C_1}$ $\hat{p}_2 = \dfrac{O_{12}}{C_2}$

Estimador puntual: $\hat{R} = \dfrac{\hat{p}_1}{\hat{p}_2} = \dfrac{O_{11}/C_1}{O_{12}/C_2} = \dfrac{O_{11}C_2}{O_{12}C_1}$

Valores posibles: $0 \le R < \infty$

$$\begin{cases} 0 \le R < 1 & Asoc.(-) \longrightarrow P<0.05 \text{ en el test } \chi^2 \\ R = 1 & Independencia \longrightarrow P>0.05 \text{ en el test } \chi^2 \\ 1 < R < \infty & Asoc.(+) \longrightarrow P<0.05 \text{ en el test } \chi^2 \end{cases}$$

Asociación positiva $R>1$ **Asociación negativa** $R<1$



En el ejemplo:

Fumar / Peso	Si	No	Total
Bajo	$O_{11}=20$	$O_{12}=26$	$F_1=46$
Normal	$O_{21}=60$	$O_{22}=294$	$F_2=354$
Total	$C_1=80$	$C_2=320$	$T=400$

$\hat{R} = \dfrac{(20)(320)}{(26)(80)} = 3.07$

La probabilidad de que el recién nacido tenga un peso bajo es 3.07 veces mayor (¡el triple!) si la madre es fumadora respecto a si no lo es

Tipo de estudios en que se puede calcular

- Transversales
- Prospectivos

En los estudios **retrospectivos** solo se puede **aproximar** si la **enfermedad** es relativamente **rara** (esto es, si su prevalencia es P(E)<0.10, esto incluye a un buen número de enfermedades) en cuyo caso tanto el estimador puntual como el IC se *aproximan* a través de la razón de producto cruzado (la siguiente medida que veremos a continuación)

Riesgo Relativo

Intervalo de confianza

Frecuencias originales

Fuma / Peso	Si (FR)	No $(\overline{FR})$	Total
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal $(\bar{E})$	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

$h=0.5$ *clásico*

Frecuencias + *h*

Fuma / Peso	Si (FR)	No $(\overline{FR})$	Total
Bajo (E)	$O_{11}=20.5$	$O_{12}=26.5$	$46.5=F_1$
Normal $(\bar{E})$	$O_{21}=60.5$	$O_{22}=294.5$	$355.5=F_2$
Total	$81.0=C_1$	$321.0=C_2$	$402.0=T$

$$\hat{R} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{O_{11}/C_1}{O_{12}/C_2} = \frac{O_{11}C_2}{O_{12}C_1} = \frac{20.5\times321.0}{26.5\times81.0} = 3.0657$$

$$IC_\alpha(R) = \hat{R}\times\exp\left\{\pm z_\alpha\sqrt{\frac{1}{O_{11}}+\frac{1}{O_{12}}-\frac{1}{C_1}-\frac{1}{C_2}}\right\} \longrightarrow IC_{0.05}(R) = 3.0657\times\exp\left\{\pm1.96\sqrt{\frac{1}{20.5}+\frac{1}{26.5}-\frac{1}{81.0}-\frac{1}{321.0}}\right\}$$

$$IC_{0.05}(R) = 3.0657\times\exp\{\pm1.96\times0.2666\} = 3.0657\times\exp\{\pm0.5225\} = 3.0657\times\{0.5931 \quad 1.6862\} = (1.8182, \quad 5.1692)$$

$$IC_{0.05}(R) = (1.8182, \quad 5.1692)$$

Con un 95% de confianza, la probabilidad de que el niño tenga el peso bajo al nacer es un valor comprendido entre 1.82 y 5.17 veces mayor si la madre fuma que si no lo hace.

En coherencia con la significación obtenida en el test χ^2 ($P<\alpha=0.05$): $1 \notin IC_{0.05}(R)$

Pedro Femia
pfemia@ugr.es

Medidas de asociación en tablas de contingencia 2x2

Razón del producto cruzado (*odds ratio* o razón de ventajas)

Fundamento: Superioridad de la enfermedad en cada categoría del FR:

$\hat{S}_1 = O_{11}/O_{21}$ Superioridad de la enfermedad en los expuestos

$\hat{S}_2 = O_{12}/O_{22}$ Superioridad de la enfermedad en los no expuestos

$$\hat{O} = \widehat{OR} = \frac{\hat{S}_1}{\hat{S}_2} = \frac{O_{11}O_{22}}{O_{12}O_{21}}$$

	FR	$\overline{FR}$	Total
E	O_{11}	O_{12}	*
$\bar{E}$	O_{21}	O_{22}	*
Total	*	*	*

$\hat{S}_1 = \frac{O_{11}}{O_{21}}$ $\hat{S}_2 = \frac{O_{12}}{O_{22}}$

Si alguna $O_{ij} = 0 \Rightarrow \hat{O}' = \dfrac{(O_{11}+0.5)(O_{22}+0.5)}{(O_{12}+0.5)(O_{21}+0.5)}$

Válido siempre, es el estimador puntual que se utiliza para construir el IC

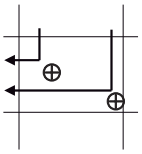
Valores posibles: $0 \le O < \infty$

$0 \le O < 1$ $Asoc.(-)$ ⟶ P <0.05 en el test χ^2

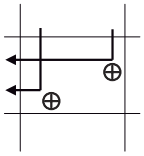
$O = 1$ $Independencia$ ⟶ P >0.05 en el test χ^2

$1 < O < \infty$ $Asoc.(+)$ ⟶ P <0.05 en el test χ^2

Asociación positiva O >1



Asociación negativa O <1



En el ejemplo:

Fumar / Peso	Si	No	Total
Bajo	$O_{11}=20$	$O_{12}=26$	$F_1=46$
Normal	$O_{21}=60$	$O_{22}=294$	$F_2=354$
Total	$C_1=80$	$C_2=320$	$T=400$

$$\widehat{OR} = \frac{(20)(294)}{(26)(60)} = 3.769$$

La fracción de niños con peso bajo al nacer frente a los de peso normal es 3.77 veces mayor si la madre es fumadora que si no lo es.

Observemos que: $\hat{S}_1 = 20/60 = 0.333$, $\hat{S}_2 = 26/294 = 0.088$; $\widehat{OR} = 0.333/0.088 = 3.769$

Tipo de estudios en que se puede calcular En todos:

- Transversales
- Prospectivos
- Retrospectivos

La razón de producto cruzado es la medida de asociación por excelencia en epidemiología:

- **Ventajas**: (1) Es válida en cualquier tipo de estudio; (2) forma parte de modelos más complejos (regresión logística)
- **Desventaja**: su interpretación es menos intuitiva que la del riesgo. A veces se la interpreta (inadecuadamente) como si fuera el riesgo relativo

Medidas de asociación en tablas de contingencia 2x2

Razón del producto cruzado (*Odds ratio*)

Intervalo de confianza

Frecuencias originales

Fuma / Peso	Si (FR)	No $(\overline{FR})$	Total
Bajo (E)	$O_{11}=20$	$O_{12}=26$	$46=F_1$
Normal $(\bar{E})$	$O_{21}=60$	$O_{22}=294$	$354=F_2$
Total	$80=C_1$	$320=C_2$	$400=T$

$h = 0.5$ ⟶ *clásico*

Frecuencias + *h*

Fuma / Peso	Si (FR)	No $(\overline{FR})$	Total
Bajo (E)	$O_{11}=20.5$	$O_{12}=26.5$	$46.5=F_1$
Normal $(\bar{E})$	$O_{21}=60.5$	$O_{22}=294.5$	$355.5=F_2$
Total	$81.0=C_1$	$321.0=C_2$	$402.0=T$

$$\widehat{OR} = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)} = \frac{\hat{S}_1}{\hat{S}_2} = \frac{O_{11}/O_{21}}{O_{12}/O_{22}} = \frac{O_{11}O_{22}}{O_{12}O_{21}} = \frac{20.5\times 294.5}{26.5\times 60.5} = 3.7656$$

$$IC_\alpha(R) = \widehat{OR}\times\exp\left\{\pm z_\alpha\sqrt{\frac{1}{O_{11}}+\frac{1}{O_{12}}+\frac{1}{O_{21}}+\frac{1}{O_{22}}}\right\} \longrightarrow IC_{0.05}(R) = 3.7656\times\exp\left\{\pm 1.96\sqrt{\frac{1}{20.5}+\frac{1}{26.5}+\frac{1}{60.5}+\frac{1}{294.5}}\right\}$$

$$IC_{0.05}(OR) = 3.7656\times\exp\{\pm 1.96\times 0.3262\} = 3.7656\times\exp\{\pm 0.6394\} = 3.7656\times\{0.5276 \quad 1.8954\} = (1.9867,\ 7.4812)$$

$$\boxed{IC_{0.05}(OR) = (1.9867,\ 7.4812)}$$

Con un 95% de confianza, la fracción de niños con el peso bajo al nacer frente a la de niños con el peso normal, es un valor comprendido entre 1.98 y 7.48 veces mayor si la madre fuma que si no lo hace.

En coherencia con la significación obtenida en el test χ^2 ($P<\alpha=0.05$): $1 \notin IC_{0.05}(OR)$

Pedro Femia
pfemia@ugr.es

Medidas de asociación en tablas de contingencia 2x2

Comentarios adicionales sobre los tipos de estudio

Factor de Riesgo / *Enfermedad*	*Expuesto* FR	*No expuesto* $\overline{FR}$	*Total*
Presente E	O_{11}	O_{12}	F_1
Ausente $\overline{E}$	O_{21}	O_{22}	F_2
Total :	C_1	C_2	T

- Los **estudios transversales**
  - Menor dificultad/coste en su ejecución. Más rápidos de realizar que los prospectivos.
  - No permiten establecer relaciones causales
  - Son los únicos que permiten estimar la **prevalencia de la enfermedad** $\hat{P}(E)=F_1/T$
  - Admiten cualquier medida de asociación

- Los **estudios prospectivos**
  - Son más difíciles/costosos de realizar (requieren seguimiento; expuestos al ***desgaste muestral***)
  - Permiten estimar la incidencia (casos nuevos de enfermedad/unidad de tiempo), pero no la prevalencia
  - Permiten analizar el efecto de factores (de riesgo) extraños

- Los **estudios retrospectivos**
  - Son rápidos de realizar, ya que no requieren seguimiento
  - Son los más adecuados para el estudio de enfermedades raras (de prevalencia baja)
  - No permiten estimar parámetros de interés epidemiológico como la **prevalencia** o la **incidencia** de la enfermedad.
  - No admiten la estimación de riesgos, aunque si que es posible **aproximar el riesgo relativo** cuando la enfermedad es poco frecuente (con prevalencia menor al 10%, esto ocurre con muchas enfermedades)

En los estudios retrospectivos sobre enfermedades raras, el ***riesgo relativo*** se puede **aproximar** a través de la **razón de producto cruzado.** Se acepta la aproximación cuando la prevalencia es $P(E)<10\%$ (esta información se debe dar de forma adicional y se suele obtener de *estudios previos de prevalencia*)

$$\text{Si } P(E)<0.1 \rightarrow \hat{R}\approx\hat{O}$$

Cuidado: Solo es lícita esta aproximación/interpretación en la situación indicada. Es un error frecuente interpretar siempre la razón de producto cruzado O como si fuera el riesgo relativo (dado que la interpretación de O es menos intuitiva que la de R).

En el análisis de tablas 2×2 deberían ser consideradas las siguientes etapas:

1. Representar la tabla en el formato estándar e identificar el **tipo de estudio** realizado e indicarlo de forma explícita. Si es un estudio prospectivo o retrospectivo, el test será de **homogeneidad**; si es transversal el test será de **independencia**. Recuérdese que el número de muestras presentes en el estudio lo debe de dejar claro (1 muestra→ test de independencia (estudio transversal); 2 o más → test de homogeneidad (estudio prospectivo o retrospectivo)
2. Formular las **hipótesis** (***homogeneidad*** o ***independencia*** siempre en $\mathbf{H_0}$)
3. Con las tablas 2x2:
   - **NO** se puede aplicar la fórmula χ^2 general. Debe usarse aquella que viene corregida con una *corrección por continuidad*
   - **NO** es necesario calcular las frecuencias esperadas
4. Comprobar las **condiciones de validez** (su cumplimiento se debe indicar de forma explícita)
5. Si el método es aplicable, obtener el **estadístico de contraste** χ^2_{exp} adecuado para el tipo de estudio
6. Obtener el **nivel de significación** P e interpretarlo (con la regla de siempre):

   Si $\begin{cases} P\le\alpha \longrightarrow \text{Test significativo (la muestra es improbable bajo } H_0) \\ P>\alpha \longrightarrow \text{Test no significativo (no se puede rechazar } H_0) \end{cases}$
7. Obtener una o más **medidas de asociación**, que sean adecuadas dado el tipo de estudio, y su IC al 95%,
   - Si el test ha resultado significativo: permiten medir la fuerza de la asociación entre las variables. Debe ocurrir, en general*:
   $$0\notin IC_{1-\alpha}(d),\quad 1\notin IC_{1-\alpha}(R),\quad 1\notin IC_{1-\alpha}(OR)$$
   - Si el test no ha sido significativo el IC proporciona una idea de la fiabilidad de la aceptación de H_0. Debe ocurrir, en general*:
   $$0\in IC_{1-\alpha}(d),\quad 1\in IC_{1-\alpha}(R),\quad 1\in IC_{1-\alpha}(OR)$$
   La precisión del intervalo permite hacer una apreciación de la idoneidad del tamaño de muestra utilizado

* A veces esto no ocurre, el test puede ser no significativo y, sin embargo, el IC para la medida de asociación no contiene al valor que indica independencia (pero por poco). Esto es debido a que el test y los IC se obtienen por métodos distintos. En general estas situaciones no se dan, pero si lo hacen constituyen un indicador claro de falta de potencia (muestras pequeñas). En estos casos, los IC suelen ser muy anchos (poco precisos), lo que corrobora aún más esta idea.

Pedro Femia
pfemia@ugr.es

BIOESTADÍSTICA

Grado en Enfermería

Tema VIII

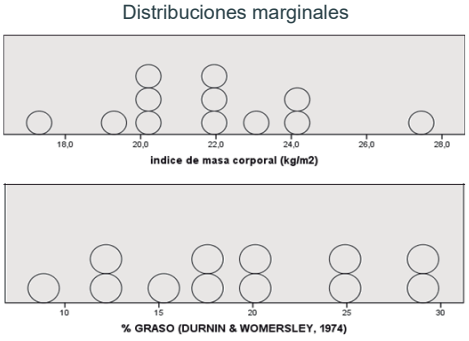
Regresión y correlación

Pedro Femia Marzo
Unidad de Bioestadística – Facultad de Medicina
Universidad de Granada

Regresión y correlación
Motivación

Problema: A continuación figuran las medidas del índice de masa corporal (IMC) y el porcentaje de peso graso (determinado en laboratorio) de una muestra de 12 adolescentes (8 hombres y 5 mujeres) con una edad comprendida entre 15 y 19 años.

#	IMC	% graso
1	20,36	17,36
2	20,34	12,72
3	20,05	11,65
4	23,07	17,81
5	23,94	19,34
6	27,45	29,70
7	22,17	24,19
8	24,37	15,26
9	22,18	28,41
10	17,30	8,87
11	21,72	25,62
12	19,29	20,91

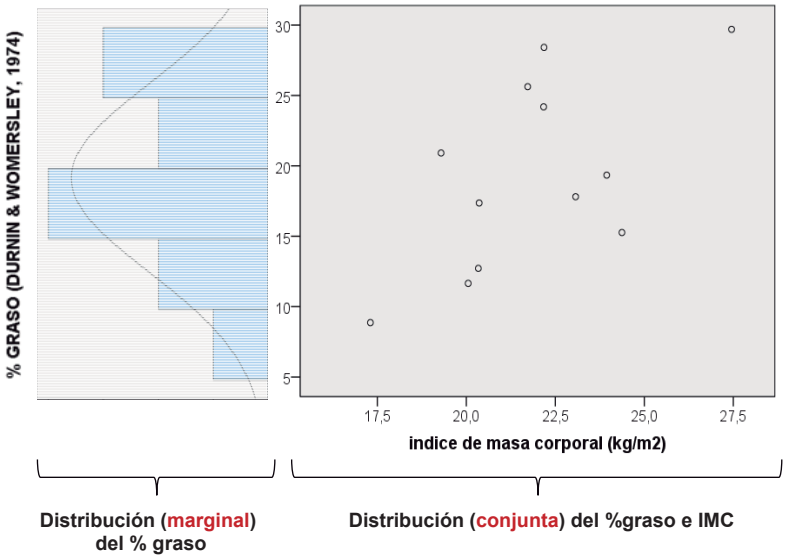


- ¿**Existe relación** entre el IMC y el porcentaje de masa grasa?
- De ser así, ¿es ***intensa*** la relación? ¿existe algún ***modelo*** que permita caracterizar tal relación?
- Sabiendo el IMC de una persona ¿es posible hacer algún tipo de **pronóstico** sobre su % graso?*
- En tal caso ¿es **fiable** tal pronóstico?

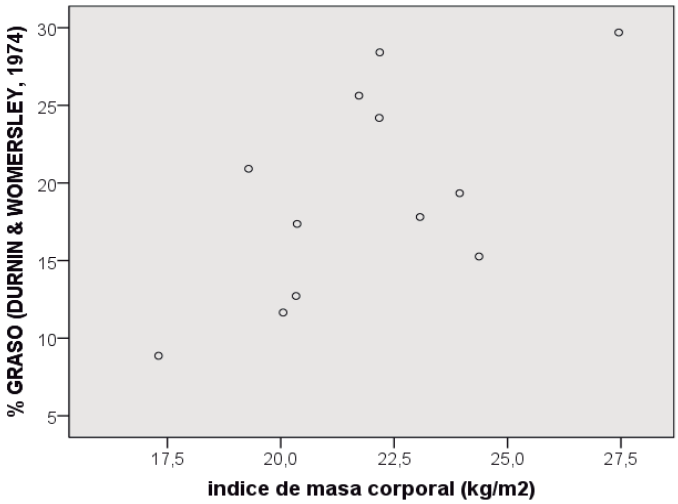
* El IMC=peso (kg) /(talla (m)) ^{2} es una medida fácil de calcular, el % de masa grasa no tanto, requiere la instrumentación adecuada.

Regresión y correlación
El primer paso: Diagrama de dispersión o nube de puntos

¿ Aporta alguna información sobre la distribución de una de las variables el hecho de conocer qué valores toma la otra?



Regresión y correlación
El primer paso: Diagrama de dispersión o nube de puntos



- Se dice que hay ***correlación*** si hay algún tipo de ***asociación***. Es decir, cuando el cambio en una variable se acompaña por un cambio "sistemático" en la otra. Mediante un ***coeficiente de correlación*** se cuantifica la ***fuerza de la asociación***.

- Por ***regresión*** se alude a un ***modelo*** (= expresión matemática) que permite caracterizar *cómo* cambia una de las variables cuando cambia la otra. Un modelo de regresión permite no solo describir la forma del cambio, también pronosticar el valor esperado de una de las variables a partir del valor que toma la otra.

El diagrama de dispersión permite hacer una primera valoración sobre ambas cuestiones.

Pedro Femia
pfemia@ugr.es

Modelos de regresión
Rol de las variables y tipos de muestreo

Rol de las variables

- Una variable actúa como variable ***explicativa, regresora*** o ***independiente*** (habitualmente se indica como x)
- La otra variable es la variable ***explicada*** o ***dependiente*** (habitualmente se representa por y)

Tipo de Muestreo

Cada individuo aporta una pareja de datos (muestras apareadas) solo que, en general, se trata de variables diferentes (aunque no necesariamente). Hay dos formas posibles de obtener los datos:

- **Muestreo de tipo I** (de una muestra aleatoria de individuos se observan ambas variables) ← **observación** (es el caso del ejemplo que estamos considerando). Las dos variables pueden desempeñar el rol de variable explicativa.
- **Muestreo de tipo II** (fijados de antemano los valores de x, se observan los de y) ← **experimentación** (el rol de las variables no es intercambiable)

Asociación no implica causalidad

Que dos variables estén asociadas no supone necesariamente que una sea la causa de la otra.

Ejemplo, consideremos niños con edad ≤ 12 meses

Relación causal unidireccional

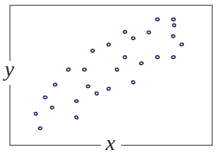
Edad ⟶ peso

x y

Influencia (causal) mutua

Actividad hipofisaria ⇄ Tiroxina plasmática

x y



Asociación espuria

Estatura ⟶ Destreza motriz

x y

Estatura ✕ Destreza motriz

y y

Edad

x

El muestreo de tipo I solo permite hablar de asociación (en el sentido de *covariación*), no de causalidad. Para establecer una relación causal se debe comenzar por diseñar el muestreo de tipo II

Modelos de regresión
El modelo de regresión como *modelo estadístico*

Que y se relacione con x se expresa, de la forma general: $y = f(x)$ (esto es un modelo)

Para establecer, o especificar, el ***modelo*** deberemos concretar la forma que toma $f(x)$

En cualquier caso, $y = f(x)$ debe poder representar una ***relación no determinista*** → para los mismos valores de x no se observan necesariamente los mismos valores de y

En el ámbito científico se presentan básicamente dos tipos de modelos:

Modelos mecanicistas

- Se establecen a partir del **conocimiento intimo** del proceso (ej. Ecuación de Michaelis-Menten)
- Son **deterministas** → para un mismo valor de x se obtiene siempre la misma y
- Experimentalmente nunca se observarán los datos exactamente como los predice el modelo mecanicista **(A)**, siempre habrá error de medida y otros factores *no asignables* que hacen que las observaciones presenten cierta variabilidad en torno al valor pronosticado por dicho modelo **(B)**

Modelos estadísticos

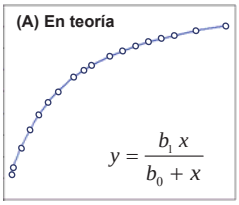
Se pueden expresar

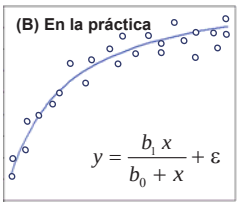
$y =$ **componente determinista** + **componente aleatorio**

(forma de la dependencia de y respecto a x) (influencia de causas no asignables)

- El **componente determinista** puede corresponder a un ***modelo mecanicista*** (sugerido por el conocimiento del mecanismo que relaciona las variables) o por un ***modelo empírico*** (sugerido por los propios datos, la forma que toma la *nube de puntos*)
- El **componente aleatorio** es el responsable de la **variabilidad** observada en los datos respecto al componente determinista. En general (nosotros lo haremos así), se asume que se trata de una variable aleatoria con **distribución normal** y **media cero** (los errores por exceso se compensan con los errores por defecto)

Efecto del componente aleatorio



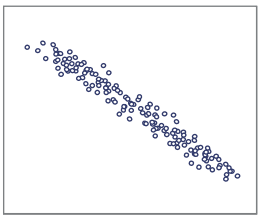


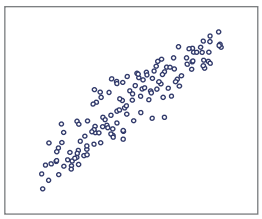
	x	y
#	IMC	% graso
1	20,36	17,36
2	20,34	12,72
3	20,05	11,65
4	23,07	17,81
5	23,94	19,34
6	27,45	29,70
7	22,17	24,19

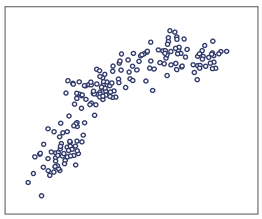
Para valores similares de x se observan diferentes valores de y

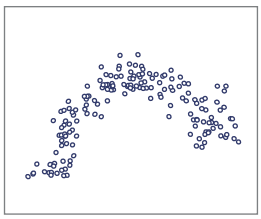
Pedro Femia
pfemia@ugr.es

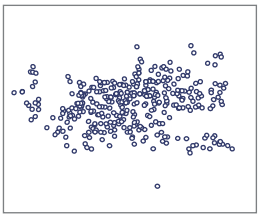
Modelos de regresión
Diagnóstico del diagrama de dispersión

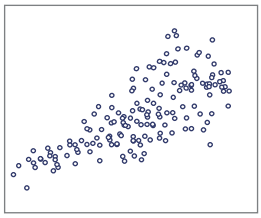

Relación lineal inversa

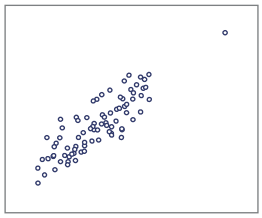

Relación lineal directa

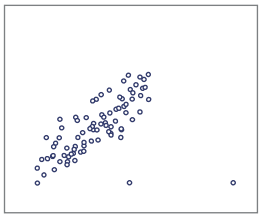

Relación no lineal (monotónica)


Relación no lineal (no monotónica)

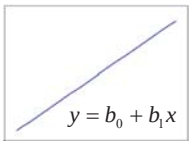

Ausencia de relación


Heterocedasticidad

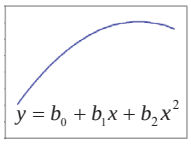

Observación influyente (pero que mantiene el modelo lineal)


Observaciones atípicas (no mantienen el modelo lineal)

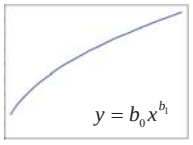
Algunos modelos (con base determinista o sugeridos por los datos) de uso frecuente



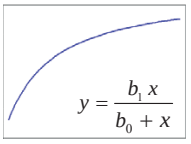
Lineal



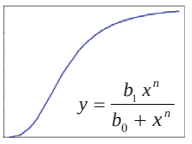
Polinómico



Potencial



Hiperbólico (Michaelis-Menten)

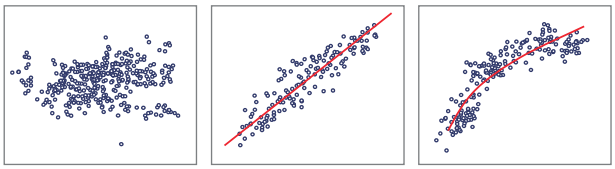


Sigmoidal (Hill)

Modelos de regresión
Diagnóstico del diagrama de dispersión

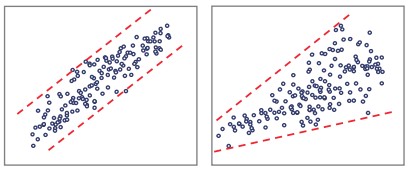
Qué se debe observar en un diagrama de dispersión

1. Si **hay relación** o no entre las variables

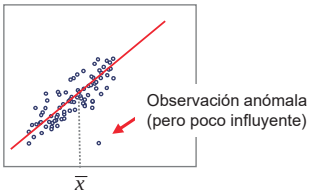


2. En caso de haber relación, si es ***lineal*** o ***no lineal***

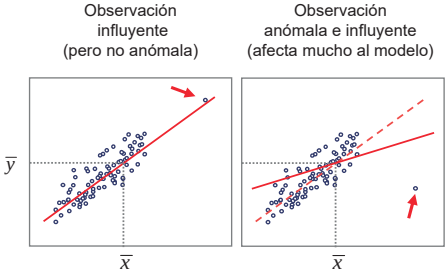
3. En caso de haber relación lineal → al aumentar x aumenta o disminuye linealmente la media de y pero, ¿cambia también la variabilidad de y al cambiar x (***heterocedasticidad****) ?
   La heterocedasticidad es un fenómeno frecuente y "perjudica" la *calidad* del modelo (las estimaciones pierden precisión). A menudo se puede corregir considerando el logaritmo de la respuesta, o del predictor y de la respuesta


homocedasticidad heterocedasticidad

4. ¿Hay puntos ***anómalos*** respecto al modelo que viene sugerido por la mayoría de los datos?



5. ¿Hay observaciones ***influyentes***?
   El nivel de influencia de una observación viene dado por su distancia a la media de la variable explicativa, cuanto más alejada esté de $\overline{x}$, más influyente resulta sobre el modelo.



* - *cedasticidad* procede del griego *skedasis* (σκεδασις) que quiere decir *dispersión*

Pedro Femia
pfemia@ugr.es

El Modelo de Regresión Lineal

Desde el punto de vista algebraico Ecuación general de la línea recta:

$y = a + b\,x$ 2 parámetros
- ***b*** = pendiente (= inclinación de la recta)
- ***a*** = ordenada en el origen (valor de y cuando x=0)

Punto de vista estadístico $y = \beta_0 + \beta_1\,x + \varepsilon$ **Modelo poblacional**

La combinación lineal de los dos parámetros descritos constituyen el ***componente determinista*** del modelo, que caracteriza la forma en que y depende de x. Además, el modelo presenta un ***componente aleatorio*** que es el responsable de la variabilidad observada en y para un mismo valor de x, ambos componentes son también ***aditivos***

$$y = (\beta_0 + \beta_1\,x) + \varepsilon \qquad \varepsilon \to N(0, \sigma_\varepsilon) \qquad y \mid X = x \to N(\beta_0 + \beta_1 x, \sigma_\varepsilon)$$

Respuesta = componente determinista + componente aleatorio

Valor medio (esperado) de y condicionado a que el valor del predictor sea x

Por hipótesis, se espera que el ***componente aleatorio*** ε tenga distribución normal con un valor promedio de cero (los *excesos* se compensan con los *defectos*).

Ajustar el modelo estadístico consiste en ***estimar*** los parámetros que constituyen el componente determinista para caracterizar cómo afecta x a y

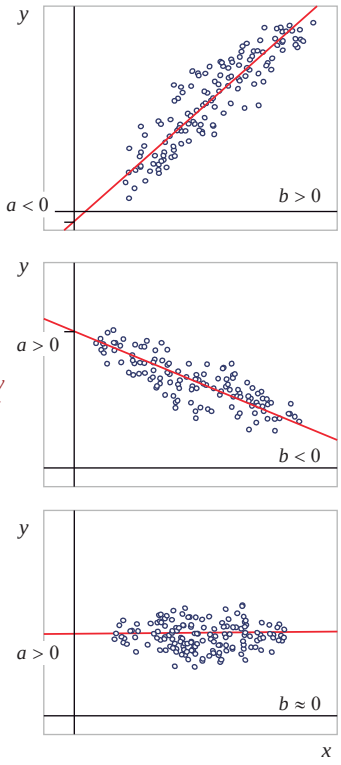
$$\hat{\beta}_0 + \hat{\beta}_1\,x = a + b\,x$$

Del componente aleatorio interesa estimar su **varianza** σ_ε^2 ya que es la responsable de la variabilidad observada en la respuesta respecto al componente determinista

$$S_R^2 = \hat{\sigma}_\varepsilon^2 \qquad = \textbf{varianza residual}$$

$\hat{y}$ será el ***valor esperado de*** y (inferencia sobre y) dado un valor de x

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\,x \to \hat{y} = a + b\,x$$



El Modelo de Regresión Lineal

Interpretación de los parámetros

$\hat{y} = a + b\,x$

- ***Coeficiente de regresión***: cuánto (y cómo) cambia y por una unidad de aumento en x
- ***Constante***: cuánto vale y para x=0 según el modelo (no siempre tiene sentido interpretarlo).

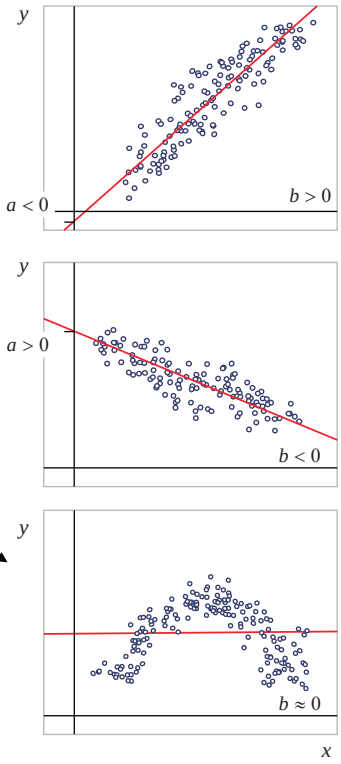
El coeficiente mas relevante es el ***coeficiente de regresión*** (la pendiente)

$b > 0$ El aumento en x va acompañado por un aumento en y
Asociación lineal directa o positiva

$b < 0$ El aumento en x va acompañado por una reducción en y
Asociación lineal inversa o negativa

$b \approx 0$ Los valores de y no tienen relación con los que toma x
Falta de asociación lineal

¡Cuidado! Un coeficiente de regresión ***b* = 0** no se traduce necesariamente en **independencia** entre las variables, solo en la **falta de asociación lineal**



Relevancia del modelo lineal

- Es el más simple (si un modelo simple caracteriza bien los datos, siempre es preferible a uno mas complejo y con más parámetros)
- Muchas relaciones no lineales se pueden transformar en lineales, por ejemplo

$$y = b_0 x^{b_1} \to \ln y = \ln b_0 + b_1 \ln x \to Y = B_0 + b_1 X$$

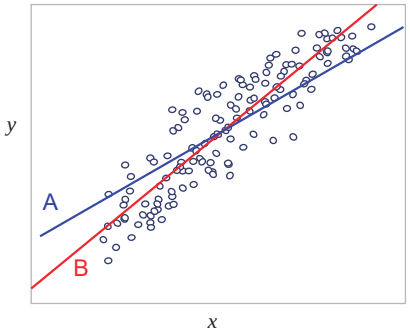
- En un intervalo de variación reducido, el cambio observado es "más o menos" lineal

Pedro Femia
pfemia@ugr.es

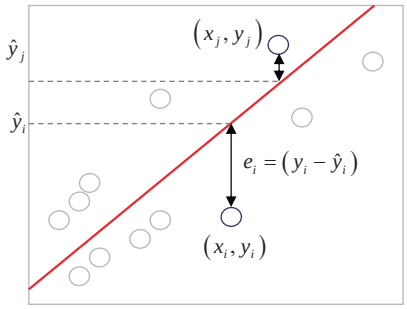
Ajuste del modelo lineal. Método de mínimos cuadrados

Ajustar el modelo lineal quiere decir **estimar** cuánto valen sus parámetros: los dos coeficientes de la recta (a y b) y la varianza de regresión (o varianza residual)

¿Cuál de los dos es mejor modelo para caracterizar la relación entre las dos variables, A o B?



Criterio de mínimos cudrados



$$e_i = (y_i - \hat{y}_i)$$

$Residuo_i$ = (*valor observado*)$_i$ – (*valor pronosticado por el modelo*)$_i$

La mejor recta, en el sentido de ser la más "*próxima*" a todos los datos es aquella en la que la suma de residuos al cuadrado es la mínima posible. Es decir, $\hat{y} = a + b\,x$ es aquella que verifica que $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ es el mínimo

Además de estimar los dos coeficientes del modelo, será necesario estimar la **varianza de regresión** s_R^2, que es la varianza de los residuos y estima a la varianza del componente aleatorio ε

$$s_R^2 = \hat{\sigma}_\varepsilon^2$$

Ajuste del modelo lineal. Método de mínimos cuadrados

Cálculos necesarios en el método de mínimos cuadrados

#	IMC	% graso
1	20,36	17,36
2	20,34	12,72
3	20,05	11,65
4	23,07	17,81
5	23,94	19,34
6	27,45	29,70
7	22,17	24,19
8	24,37	15,26
9	22,18	28,41
10	17,30	8,87
11	21,72	25,62
12	19,29	20,91

$n = 12$

Para x

$\sum x_i = 262.24$

$\sum x_i^2 = 5809.61$

$\bar{x} = 21.85$

$s_x^2 = 7.163 \quad (s_x = 2.676)$

Para y

$\sum y_i = 231.84$

$\sum y_i^2 = 4969.65$

$\bar{y} = 19.32$

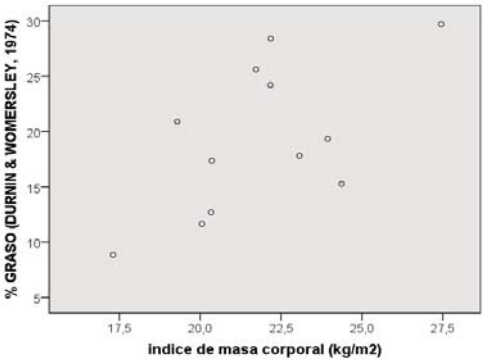
$s_y^2 = 44.591 \quad (s_y = 6.678)$

Para x e y: $\sum x_i y_i = 5186.48$

$$(xx) = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 78.79 \qquad (= (n-1)s_x^2)$$

$$(yy) = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 490.50 \qquad (= (n-1)s_y^2)$$

$$(xy) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 120.01$$



* $s_{xy} = \frac{1}{n-1}\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)$ es la ***covarianza*** entre x e y

Pedro Femia
pfemia@ugr.es

Ajuste del modelo lineal. Método de mínimos cuadrados

Estimación de los coeficientes y de la varianza de regresión

#	IMC	% graso
1	20,36	17,36
2	20,34	12,72
3	20,05	11,65
4	23,07	17,81
5	23,94	19,34
6	27,45	29,70
7	22,17	24,19
8	24,37	15,26
9	22,18	28,41
10	17,30	8,87
11	21,72	25,62
12	19,29	20,91

$n = 12$

$\sum x_i = 262.24$ $\quad \sum y_i = 231.84$

$\sum x_i^2 = 5809.61$ $\quad \sum y_i^2 = 4969.65$ $\quad \sum x_i y_i = 5186.48$

$\bar{x} = 21.85$ $\quad \bar{y} = 19.32$

$s_x^2 = 7.163$ $\quad s_y^2 = 44.591$

$(xx) = 78.79$ $\quad (yy) = 490.50$ $\quad (xy) = 120.01$

$$b = \frac{(xy)}{(xx)} = \frac{120.01}{78.79} = 1.523$$

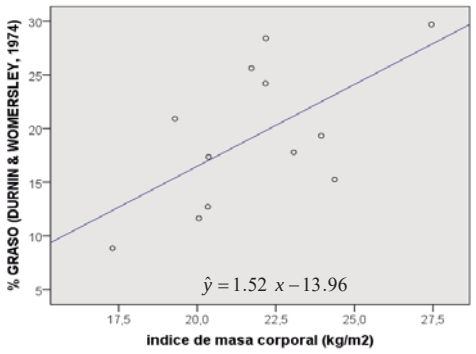
$$\bar{y} = a + b\,\bar{x} \;\rightarrow\; a = \bar{y} - b\,\bar{x}$$

$$a = 19.32 - 1.52 \times 21.85 = -13.96$$

$$\hat{y} = a + b\,x$$

$$\hat{y} = -13.96 + 1.52\,x$$

Modelo lineal estimado



Varianza de regresión

$$s_R^2 = \frac{1}{n-2}\left((yy) - \frac{(xy)^2}{(xx)}\right) = \frac{1}{12-2}\left(490.50 - \frac{120.01^2}{78.79}\right) = 30.773$$

o también:

$$s_R^2 = \frac{1}{n-2}\left((yy) - b\,(xy)\right) = \frac{1}{12-2}\left(490.50 - 1.52 \cdot 120.01\right) = 30.773$$

Inferencias con el modelo

¡Importante!

A la pregunta:		**Se responde con:**
(En cualquier caso, el paso preliminar es siempre el diagnóstico del diagrama de dispersión)		
• ¿Cuál es la relación lineal entre x e y ?	→	Expresión del ***modelo ajustado***: $\hat{y} = a + b\,x$
• ¿Es significativa la relación lineal entre x e y ?	→	***Test de regresión lineal***, o también ***Test de correlación lineal*** (en realidad son el mismo test)
• ¿Cuánto cambia y por unidad de aumento en x ?	→	IC para el ***coeficiente de regresión***: IC(β)
• ¿Cuánto vale y para un valor dado de $x = x_0$?	→	***Pronósticos*** sobre $y\|x{=}x_0$
• ¿Cuánto vale x para un valor dado de $y = y_0$?	→	¡¡ ***Pronósticos*** sobre $x\|y{=}y_0$!!
• ¿Es intensa la asociación (lineal) entre x e y ?	→	***Coeficiente de correlación lineal*** r
• ¿Son buenos los pronósticos realizados con el modelo de regresión lineal?	→	***Coeficiente de determinación*** R^2

Pedro Femia
pfemia@ugr.es

Test de regresión lineal

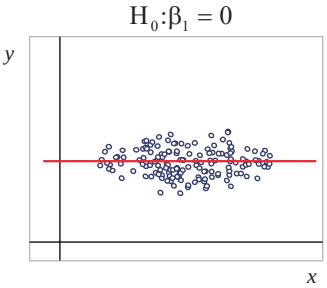
¿Hay relación lineal entre *x* e *y* ?

Para el modelo $y = \beta_0 + \beta_1 x + \varepsilon$

Hipótesis $\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$ no hay asociación lineal (la recta es horizontal)

Estadístico: $t_{exp} = \dfrac{|b|}{\sqrt{s_R^2/(xx)}} = \dfrac{|b|}{s_R}\sqrt{(xx)}$

El nivel de significación ***p*** se busca en la distribución t-Student con *n-2 g.l.*



En el ejemplo: $t_{exp} = |1.52|\sqrt{\dfrac{78.79}{30.77}} = 2.437 \quad (10\ g.l.) \longrightarrow 0.03 < p < 0.04$

Hay un cambio significativo en el % graso al cambiar el IMC

IC para el coeficiente de regresión

¿Cuánto cambia *y* por unidad de aumento en *x* ?

$$IC_{(1-\alpha)}(\beta_1) = b \pm t_{\alpha;n-2}\sqrt{\frac{s_R^2}{(xx)}}$$

Recuérdese que una vez hecho un *t-test*, el IC es inmediato:

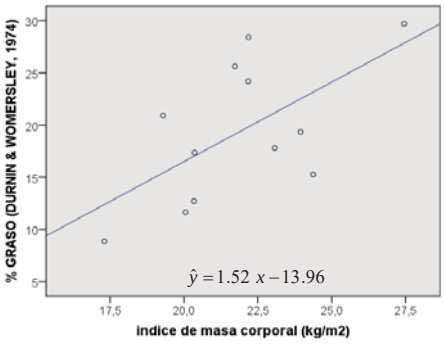
IC=numerador de la t_{exp} $\pm t_{\alpha;n-2}$ denominador de la t_{exp}

(el numerador sin el valor absoluto)

En el ejemplo:

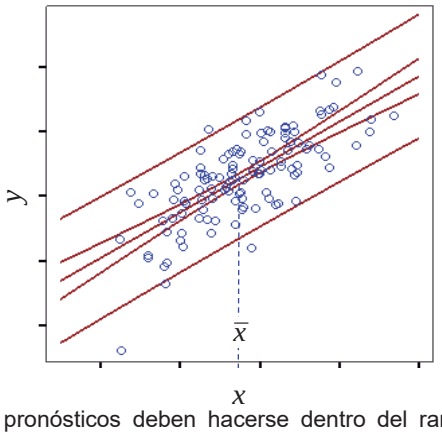
$$IC_{95\%}(\beta_1) = 1.523 \pm 2.228 \times 0.625 = (0.13;\ 2.92)$$

Por cada unidad de aumento en el IMC, el % graso aumenta (b>0) un valor que debe ser mayor a 0.13 unidades y menor a 2.92 unidades con un 95% de probabilidad (confianza)

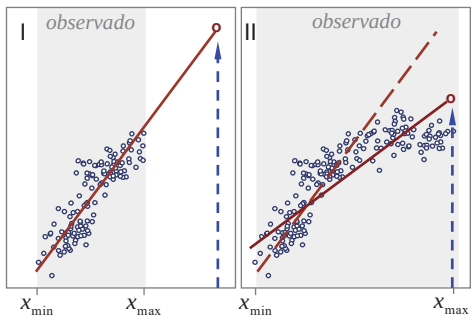


Pronósticos

¿Cuánto vale y para un valor dado de $x = x_0$?



El peligro de las extrapolaciones



- Los pronósticos deben hacerse dentro del rango observado de la variable explicativa, es decir, deben tener carácter *interpolador*, nunca *extrapolador* (mas allá del rango observado no sabemos cómo es la relación entre las variables)
- Estimación puntual de un pronóstico: Para un valor $x = x_0$ tal que $x_{min} \le x_0 \le x_{max}$ el pronóstico dado por el modelo de regresión lineal, es $\hat{y} = a + b\,x_0$
- **Intervalos de confianza.** Se puede hacer inferencia a dos niveles:
  - respecto a valores medios esperados → bandas de confianza
  - respecto a valores individuales → intervalos de predicción
- La precisión de estos pronósticos es mayor cuanto mayor sea la proximidad de x_0 a la media $\overline{x}$

Pedro Femia
pfemio@ugr.es

Pronósticos (continuación) ¿Cuánto vale y para un valor dado de $x = x_0$?

En el ejemplo, veamos diferentes situaciones:

• ¿Cuánto es el valor esperado del % graso para una persona de entre 15 y 19 años con un IMC=26?

La edad se encuentra en el rango observado
$IMC_{min} < 26 < IMC_{max}$
} El modelo SI permite estimarlo (solo damos la estimación puntual, pero sería procedente dar el intervalo de confianza)

$$\hat{y} = -13.96 + 1.52\ x$$

$$\widehat{IMC} = -13.96 + 1.52 \times 26 = 25.56$$

• ¿Cuánto es el valor esperado del % graso para una persona de entre 30 años con un IMC=26?

La edad NO se encuentra en el rango observado } El modelo NO permite estimarlo

• ¿Cuánto es el valor esperado del % graso para una persona de entre 15 y 19 años con un IMC=32?

La edad se encuentra en el rango observado
$32 > IMC_{max}$
} El modelo NO permite estimarlo

Pronósticos sobre x dado un valor de y ¿Cuánto vale x para un valor dado de $y = y_0$?

Que se pueda hacer o no depende del tipo de muestreo:

• **Si el muestreo es de tipo I** (las dos variables observadas)

Se trata de invertir el rol de las variables (como x e y son variables aleatorias se puede hacer)

Partiendo de: $(xx) = 78.79 \quad (yy) = 490.50 \quad (xy) = 120.01 \quad \bar{x} = 21.85 \quad \bar{y} = 19.32$

Regresión de y sobre x $(y|x)$

$$b = \frac{(xy)}{(xx)} = \frac{120.01}{78.79} = 1.523 \qquad \bar{y} = a + b\ \bar{x} \ \rightarrow\ a = \bar{y} - b\,\bar{x} \qquad a = 19.32 - 1.52 \times 21.85 = -13.96$$

$$\hat{y} = a + b\ x \ \rightarrow\ \hat{y} = -13.96 + 1.52\ x$$

Regresión de x sobre y $(x|y)$

$$b' = \frac{(xy)}{(yy)} = \frac{120.01}{490.50} = 0.245 \qquad \bar{x} = a' + b'\ \bar{y} \ \rightarrow\ a' = \bar{x} - b'\,\bar{y} \qquad a' = 21.85 - 0.245 \times 19.32 = 17.12$$

$$\hat{x} = a' + b'\ y \ \rightarrow\ \hat{x} = 17.12 + 0.245\ y$$

Como antes, la inferencia $\hat{x} = 17.12 + 0.245\ y_0$
solo se puede hacer si $y_{min} \le y_0 \le y_{max}$

• **Si el muestreo es de tipo II** (los valores de la variable explicativa están fijados de antemano)

No se puede elaborar el modelo de regresión de $x|y$ (ahora x no es una variable aleatoria)

El problema se aborda mediante ***calibración lineal*** (no lo vemos aquí)

Pedro Femia
pfemia@ugr.es

Correlación y correlación lineal

¿Es intensa la relación entre x e y ?

Se dice que entre dos variables (cuantitativas) hay **correlación** si hay **asociación**, es decir:

- Si valores altos (bajos) de una de las variables se acompañan con valores altos (bajos) de la otra. En este caso se dice que la **correlación** es **directa** o **positiva**
- Si valores altos (bajos) de una de las variables se acompañan de valores bajos (altos) de la otra. Ahora la **correlación** es **inversa** o **negativa**

Un ***coeficiente de correlación*** es un índice que mide la fuerza de asociación entre dos variables cuantitativas

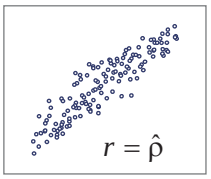
En general se representa a estos coeficientes como:
• ρ (*rho* del alfabeto griego) para aludir a la fuerza de asociación **poblacional**
• r para aludir a su **estimador muestral**: $r = \hat{\rho}$

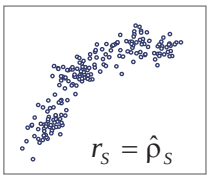
Coeficientes de correlación de Pearson y de Spearman

Si la asociación entre las variables es de tipo lineal, el coeficiente adecuado es el ***coeficiente de correlación lineal de Pearson*** **(ρ).** Este coeficiente es el más habitual y se alude a él simplemente por r

Si la asociación entre las variables es de tipo no lineal pero monotónica (siempre creciente o decreciente, sin forma de U o de ∩), el coeficiente anterior no es adecuado, pero puede calcularse el ***coeficiente de correlación de Spearman*** **(ρ_S)**

- Se alude a él como ***coeficiente de correlación no paramétrico*** (con frecuencia se usa cuando las variables no tienen distribución normal)
- La interpretación se hace igual que la del coeficiente de Pearson (quitando ahora el calificativo *lineal*)





El coeficiente de Pearson no es válido aquí

Correlación lineal

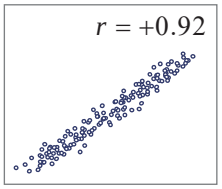
¿Es intensa la relación lineal entre x e y ?

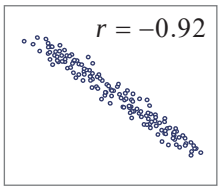
Valores posibles (válido para el coeficiente ρ_S de Spearman)

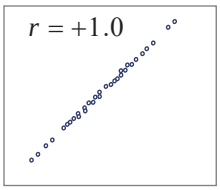
$$-1 \le \rho \le 1$$

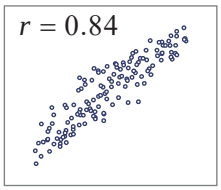
De un coeficiente de correlación se debe interpretar:
(válido para el coeficiente ρ_S de Spearman eliminando la palabra *lineal*)

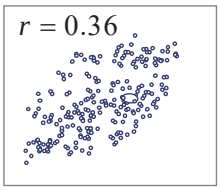
- Su signo
  - $\rho > 0$ Asociación lineal **positiva:** cuando $x \nearrow \Rightarrow y \nearrow$
  - $\rho < 0$ Asociación lineal **negativa:** cuando $x \nearrow \Rightarrow y \searrow$
- Su magnitud
  - $|\rho| \to 0$ Falta de asociación lineal
  - $|\rho| \to 1$ Asociación lineal perfecta

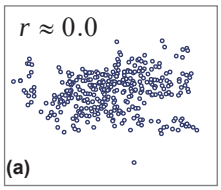


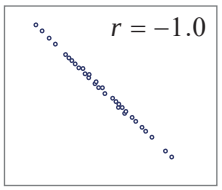




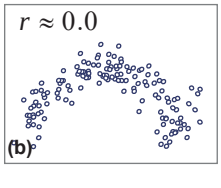








La **falta de correlación lineal** no es lo mismo que la **independencia** (esta es una condición mucho más fuerte). La variable y no depende de x en **(a)** pero si en **(b)**, donde la dependencia es no lineal. En este último caso r no es un coeficiente válido.



Ni el coeficiente de correlación de Pearson ni el de Spearman son válidos aquí.

Pedro Femia
pfemia@ugr.es

Correlación lineal

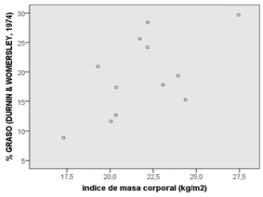
¿Es intensa la relación lineal entre *x* e *y* ?

Estimación (puntual):

$$\hat{\rho} = r \quad \text{con} \quad r = \frac{(xy)}{\sqrt{(xx)(yy)}}$$

El coeficiente de **correlación** tiene **el mismo signo** que el coeficiente de **regresión**: que es el signo de (xy) (el numerador de la *covarianza*)

En el ejemplo:



$$(xx) = 78.79$$
$$(yy) = 490.50$$
$$(xy) = 120.01$$
$$r = \frac{120.01}{\sqrt{78.79 \times 490.50}} = 0.610$$

Test de correlación:

¿Hay relación lineal entre *x* e *y* ?; Este test es **equivalente** al **test de regresión** H_0:β_1=0

Hipótesis $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$ no hay asociación lineal (la recta de regresión es horizontal, es decir, β_1=0)

Estadístico: $t_{exp} = \sqrt{\frac{(n-2)r^2}{1-r^2}}$ Se busca en la distribución t-Student con *n-2 g.l.*

En el ejemplo:

$$t_{exp} = \sqrt{\frac{(12-2)0.61^2}{1-0.61^2}} = 2.43 \ \ (10\ g.l.) \longrightarrow \ 0.03 < p < 0.04$$

Correlación de Spearman

¿Es intensa la relación lineal entre *x* e *y* ?

Estimación (puntual):

Válido para relaciones no lineales pero monótonas. Muy usado en Psicología

Se obtiene como el coeficiente de correlación de Pearson considerando los rangos: $r_S = \frac{(R^X R^Y)}{\sqrt{(R^X R^X)(R^Y R^Y)}}$

	X	Y	R_i^X	R_i^Y	$(R_i^X - R_i^Y)^2$
1	20.36	17.36	5	5	0
2	20.34	12.72	4	3	1
3	20.05	11.65	3	2	1
4	23.07	17.81	9	6	9
5	23.94	19.34	10	7	9
6	27.45	29.70	12	12	0
7	22.17	24.19	7	9	4
8	24.37	15.26	11	4	49
9	22.18	28.41	8	11	9
10	17.30	8.87	1	1	0
11	21.72	25.62	6	10	16
12	19.29	20.91	2	8	36
			78	78	134

$\sum R_i^X = \sum R_i^Y = 78.0$ Esta igualdad ocurre siempre, de hecho $\sum R_i^X = \sum R_i^Y = \frac{n(n+1)}{2}$

$\sum (R_i^X)^2 = \sum (R_i^Y)^2 = 650.0$ Esta igualdad ocurre solo si no hay empates

$\sum R_i^X R_i^Y = 583.0$

$(R^X R^Y) = 583.0 - \frac{(78)^2}{12} = 76.0$

$(R^X R^X) = 650.0 - \frac{(78)^2}{12} = 143.0 = (R^Y R^Y)$ por no haber empates

$r_S = \frac{(R^X R^Y)}{(R^X R^X)} = \frac{76}{143} = 0.531$ por no haber empates

Cuando no hay empates, se puede usar la expresión $r_S = 1 - \frac{6}{(n-1)\,n\,(n+1)} \sum (R_i^X - R_i^Y)^2 = 1 - \frac{6}{11 \times 12 \times 13} 134 = 0.531$

Test

$\begin{cases} H_0 : \rho_S = 0 \\ H_1 : \rho_S \neq 0 \end{cases}$ Si $\begin{cases} n \leq 30 \longrightarrow |r_S| \text{ En tabla de límites de significación par a el coeficiente de correlación de Spearman} \\ n > 30 \longrightarrow |r_S|\sqrt{n-1} \text{ En tabla de la distribución normal estándar} \end{cases}$

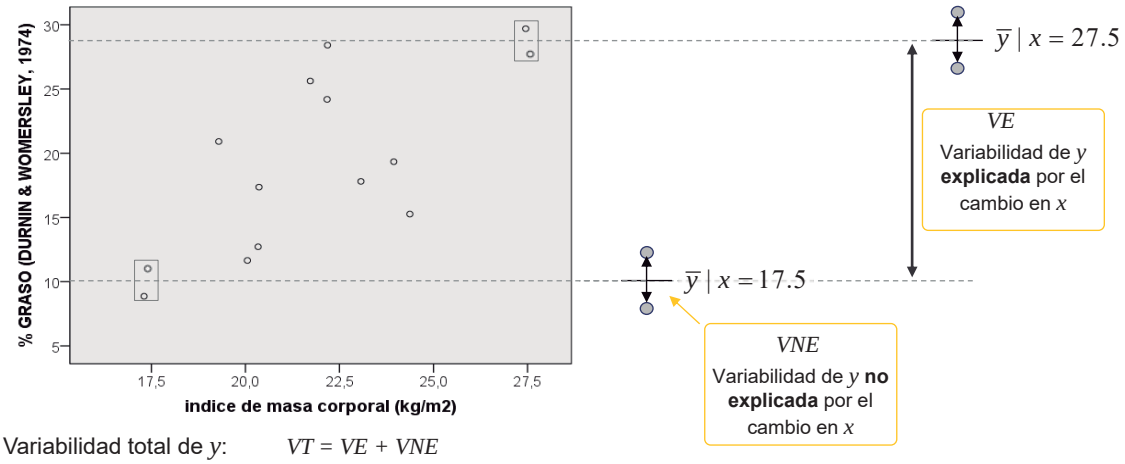
Aquí: $|r_S| = 0.531 \quad (n=12) \rightarrow r_{0.10}\,(=0.4973) < r_s\,(=0.531) < r_{0.05}\,(=0.5910) \rightarrow \ 0.05 < p < 0.10$

Pedro Femia
pfemia@ugr.es

El Coeficiente de Determinación

¿Es bueno el modelo de regresión a la hora de realizar pronósticos sobre y dado x ?

Se expresa como R^2 y se suele utilizar como un índice de la calidad de los pronósticos dados por la recta de regresión (se puede expresar en %) y habitualmente se considera como una medida de la calidad del ajuste dado por la recta de regresión



Variabilidad total de y: $VT = VE + VNE$

$R^2 = \dfrac{VE}{VT}$ El ***coeficiente de determinación*** es la proporción de variabilidad observada en y que queda explicada por su relación lineal con x

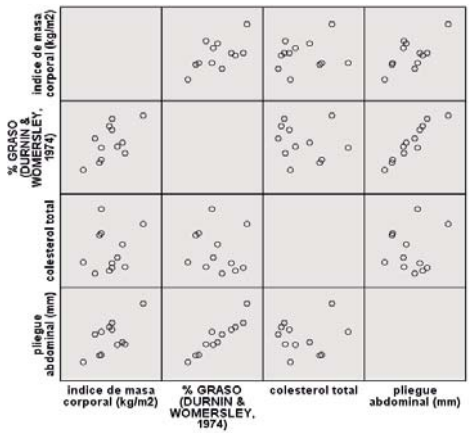
En la práctica se puede calcular como el cuadrado del coeficiente de correlación lineal: $R^2 = (r)^2$

En el ejemplo: $R^2 = (0.610)^2 = 0.373$ → El 37.3% de la variabilidad observada en el porcentaje graso viene determinada (o explicada) por su relación lineal con el IMC. El 100–37.3=62.7% restante es variabilidad no explicada

Comentarios finales

En la práctica, lo habitual es que en el análisis participen más de dos variables

Diagrama de dispersión matricial



Matriz de correlaciones

		IMC	% GRASO	colesterol total	pliegue abdominal (mm)
indice de masa corporal (kg/m2)	Correlación de Pearson	1	,610*	,163	,654*
	Sig. (bilateral)		,035	,613	,021
	N	12	12	12	12
% GRASO (DURNIN & WOMERSLEY, 1974)	Correlación de Pearson	,610*	1	-,143	,943**
	Sig. (bilateral)	,035		,658	,000
	N	12	12	12	12
colesterol total	Correlación de Pearson	,163	-,143	1	,095
	Sig. (bilateral)	,613	,658		,768
	N	12	12	12	12
pliegue abdominal (mm)	Correlación de Pearson	,654*	,943**	,095	1
	Sig. (bilateral)	,021	,000	,768	
	N	12	12	12	12

*. La correlación es significante al nivel 0,05 (bilateral).

**. La correlación es significativa al nivel 0,01 (bilateral).

La correlación de una variable consigo misma es, lógicamente, r=1

$\hat{y} = b_0 + b_1 x_1$ **Modelo de regresión lineal simple** (solo hay una variable explicativa)

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$ **Modelo de regresión lineal múltiple** (hay un conjunto de k variables ***explicativas*** o ***regresoras***)

o0o

Fin de la asignatura

Pedro Femia
pfemio@ugr.es