Matemáticas y Estadística II

Grado en Farmacia

Notas de Clase

Curso 2025–26

Índice general

Prefacio 2
Clase Expositiva 1: Presentación 4

Tema 1 **7**
2. Expositiva 2 7
2.1. Estadística descriptiva 7
2.1.1. Análisis de datos 7
2.1.2. Medidas de tendencia central 9
2.1.3. Medidas de posición 10
2.1.4. Medidas de dispersión 11
2.1.5. Varianza y Desviación típica 12
2.1.6. Disposición de los cálculos 13
2.1.7. Transformación de datos 15
2.2. Manejo de la calculadora 16
3. Expositiva 3 18
3.1. Variables aleatorias continuas 19
3.2. Inferencia estadística 22
3.3. Distribución muestral de un estadístico 22
4. Expositivas 4–5 23
4.1. Estimación de la media de la población 24
4.1.1. Conocida la varianza poblacional 24
4.1.2. Desconocida la varianza poblacional 26
4.1.3. Tamaño de la muestra 27
4.2. Estimación de una proporción 28
4.2.1. Tamaño de la muestra 28
4.3. Estimación de la varianza 29
4.4. Estimadores 31

Tema 2: Contraste de hipótesis **34**
1. Expositivas 6–10 34
1.1. Contraste para la media 34
1.1.1. Motivación 34
1.1.2. Procedimiento 35
1.2. Contraste para una proporción 36
1.3. Contraste para la varianza y la desviación típica 37
1.3.1. Contrastes unilaterales 37
1.3.2. Valor p 40
1.4. Tipos de errores 42

2. Expositivas 11–15 . . . 46
2.1. Dos medias, muestras independientes . . . 46
2.1.1. Conocidas las varianzas poblacionales . . . 47
2.1.2. Desconocidas las varianzas poblacionales, pero supuestas iguales . . . 47
2.1.3. Caso general . . . 48
2.2. Dos medias, muestras relacionadas . . . 50
2.3. Contrastes para dos proporciones . . . 52
2.3.1. Valor nulo cero . . . 52
2.3.2. Valor nulo distinto de cero . . . 53
2.3.3. Tamaño de la muestra . . . 54
2.4. La distribución F . . . 54
2.5. Contraste para dos varianzas . . . 55
2.5.1. Otros contrastes . . . 57

Tema 3: La prueba χ^2 **59**
1. Expositivas 16–18 . . . 59
1.1. Pruebas de homogeneidad . . . 59
1.2. Pruebas de independencia . . . 63
1.3. Pruebas de bondad de ajuste . . . 65
1.4. Cómo agrupar datos en intervalos . . . 68

Tema 4: Regresión y correlación **71**
1. Expositivas 19–20 . . . 71
1.1. Nociones previas . . . 71
1.1.1. Covarianza . . . 71
1.1.2. Transformación de datos . . . 72
1.1.3. Puntuaciones desviadas y tipificadas . . . 72
1.2. Regresión lineal y correlación . . . 73
1.2.1. Regresión lineal . . . 73
1.2.2. Recta de regresión . . . 73
1.2.3. Coeficiente de correlación . . . 76
1.3. Interpretación del coeficiente de correlación . . . 79
1.3.1. Varianza explicada . . . 79
2. Expositiva 21 . . . 83
2.1. Regresión no lineal . . . 83
2.1.1. Ajuste exponencial . . . 83
2.1.2. Farmacocinética: Inyección intravenosa rápida . . . 85
2.2. Decaimiento radioactivo . . . 87
2.3. Crecimiento exponencial . . . 88
2.4. Ajuste potencial (*regresión doble-log*) . . . 89
3. Expositivas 22-23 . . . 91
3.1. Contrastes de hipótesis para la regresión . . . 91
3.1.1. Contraste de hipótesis para r . . . 91
3.1.2. ANOVA . . . 92
3.1.3. Intervalos de estimación . . . 95
Apéndice . . . 97
Biografías . . . 98

Prefacio

Estas notas de *Matemáticas y Estadística II* del Grado en Farmacia continúan las de *Matemáticas para Farmacia*[*] ya publicadas y son el resultado de más de veinte años de docencia de la asignatura. La estructura fundamental del curso fue diseñada por la Prof.ª Dra. Beatriz Rodríguez Moreiras, ya retirada, y responde a un planteamiento muy claro: ofrecer unas matemáticas plenamente imbricadas con el resto de la carrera, de modo que los contenidos resulten útiles y relevantes para el estudiantado de Farmacia, sin renunciar en ningún momento al rigor conceptual.

Por mi parte, siempre he compartido y respetado este enfoque, y he procurado mantenerlo a lo largo de los años guiándome por los siguientes principios.

En primer lugar, las matemáticas constituyen la base de cualquier ciencia, y el alumnado de Farmacia debe conocer los resultados fundamentales del cálculo y de la estadística. Ahora bien, dichos contenidos no deben percibirse como un mero trámite, sino como una parte esencial de su formación científica, común a todas las disciplinas experimentales.

En segundo lugar, precisamente por esta razón, se hace especial hincapié en aquellos aspectos que pueden resultar de interés directo en técnicas de laboratorio y, muy especialmente, en farmacocinética.

Finalmente, aunque el rigor matemático no debe perderse nunca, es importante evitar contenidos sin interés formativo, excesivamente abstrusos o innecesariamente especializados para el contexto del Grado. El objetivo último es que el alumnado perciba que las matemáticas que recibe son, en lo esencial, las mismas que estudian los estudiantes de ciencias en cualquier parte del mundo, y que forman parte de un lenguaje común de la ciencia contemporánea.

En este contexto, una versión anterior de estas notas recibió en 2006-07 uno de los "VI premios a la innovación educativa y mejora de la calidad docente en la USC", por su conexión con el proyecto "Math Rules!", desarrollado en San Diego por la profesora Abbey Brown, para el uso online del software "Mathematica".

Finalmente, en una asignatura de estas características resulta imprescindible contar con un libro de texto de referencia que fije de manera clara las notaciones, los conceptos básicos y la organización de los contenidos. En nuestro caso, esta función la desempeña el manual "Estadística para Biología y Ciencias de la Salud", de Janet S. Milton, que sirve como hilo conductor del curso. Las presentes notas no pretenden sustituir dicho texto, sino complementarlo, adaptando y desarrollando aquellos aspectos que se consideran más relevantes para el Grado en Farmacia y para el enfoque aplicado de la asignatura.

Enrique Macías Virgós
Catedrático de Universidad
Santiago de Compostela, enero 2026

[*] https://hdl.handle.net/10347/44427

Licencia de uso

Se agradecerán cuantos comentarios y sugerencias se estimen oportunos. En caso de detectarse algún error o errata, se ruega comunicarlo a `quique.macias@usc.es`.

Clase Expositiva 1: Presentación

Guía Docente

La guía docente de la asignatura está disponible en el Aula Virtual de la asignatura y en la web de la Facultad. Es indispensable leerla para conocer los objetivos de la asignatura, la metodología de enseñanza y los métodos de evaluación. Aquí damos un breve resumen.

Contenido de la asignatura

El objetivo es conocer y utilizar las herramientas estadísticas necesarias para el estudio de las materias que componen el Grado de Farmacia, y que os capaciten para resolver problemas matemáticos que surgen en diferentes contextos (Biología, Química, Física, Farmacocinética). Las competencias que se van a aprender consisten en aplicar los conocimientos de Matemáticas a las ciencias farmacéuticas, para evaluar datos científicos relacionados con los medicamentos y productos sanitarios y diseñar experimentos en base a criterios estadísticos.

Bibliografía

Seguiremos sobre todo el libro de Milton, J.S., *Estadística para Biología y Ciencias de la Salud*, McGraw-Hill Interamericana, 2001.

Hay una versión electrónica disponible en
`https://www-ingebook-com.ezbusc.usc.gal/ib/NPcd/IB_Escritorio_Visualizar?cod_primaria=1000193&libro=5617`

Actividades a realizar

- Teoría, explicada en las clases expositivas (23 horas)
- Resolución de problemas y ejercicios, en las clases interactivas (10 horas)
- Laboratorios de ordenador (8 horas)
- Utilizar el aula virtual para acceder al material de la asignatura
- Tutorías y revisión del examen

Evaluación

- La asistencia a las clases expositivas e interactivas es voluntaria, aunque muy recomendable
- La asistencia a las prácticas de ordenador en laboratorio es *obligatoria* y se realizará un exámen de prácticas al final de las mismas.
- La calificación de cada estudiante será: un 20 % mediante evaluación continua (resolución de boletines de problemas, participación en el aula y en las tutorías, controles teóricos intermedios) y un 80 % la nota del examen final.

En el curso se dedica mucho tiempo a la resolución de ejercicios. Se considera un aspecto fundamental en el aprendizaje de la materia.
Se recomienda además asistir a todas las clases.

Las secciones marcadas con asterisco no son materia de examen.

Tema 1: Introducción a la Inferencia Estadística. Estimación

Capítulo 2

Expositiva 2

2.1. Estadística descriptiva

En esta clase vamos a repasar algunas ideas básicas y contenidos de Estadística descriptiva que necesitaremos de la asignatura "Matemáticas y Estadística I" del primer cuatrimestre. Son especialmente importantes:

- las nociones de media, varianza y cuasi-varianza);
- las transformaciones de datos.

2.1.1. Análisis de datos*

En diversos modelos matemáticos aparecen unos parámetros (constantes) que debemos determinar experimentalmente. Para ello tenemos que estar seguros de qué tipo de datos manejamos y cuál es su precisión. Por otra parte, al no ser exactas las mediciones tendremos que «estimar» esas constantes a partir de datos que contienen fluctuaciones aleatorias.

Tipos de datos

Al medir asociamos a cada magnitud un número. Según el tipo de escala que usemos estarán permitidas ciertas operaciones, pero otras no. Los análisis estadísticos que podemos realizar difieren también en cada caso.

Por ejemplo con datos *nominales* no tiene sentido calcular la media aritmética ni la mediana, pero sí la moda.

Hay cuatro tipos de datos o escalas (cada una es más amplia que la siguiente):

- **Datos nominales**.

  Los datos nominales son meras etiquetas para distinguir unos datos de otros. La única operación permitida es calcular los porcentajes o frecuencias para cada clase.

*Las secciones marcadas con asterisco no son materia de examen

Por ejemplo, si estamos estudiando la variable «provincia de nacimiento», podemos asignar a Ourense un `1` y a A Coruña un `0`, pero no tiene sentido calcular la media de esta variable.

- **Escalas ordinales**

  En las escalas ordinales no sólo distinguimos los datos sino que les asignamos un orden. Sin embargo, las diferencias entre los grados de la escala no tienen ningún significado.

  Ejemplo A los tres primeros equipos A,B,C de la clasificación de la liga de fútbol podemos asignarles los números `1`, `2` y `3`. Esto indica el orden entre ellos. Pero la diferencia de puntos entre A y B no tiene por qué ser la misma que entre B y C; en otras palabras, podríamos haberles asignado otros números ordenados, como `1`, `10` y `100`.

  Para este tipo de datos puede calcularse la mediana y los percentiles, pero no tiene sentido calcular la media aritmética. Por este motivo en ocasiones se prefiere que las notas de un examen estén dadas con letras o palabras (por ejemplo notable, aprobado, suspenso) y no con una escala numérica.

- **Escalas de intervalo**

  Las escalas de intervalo son el tipo de datos que encontraremos usualmente en el laboratorio, y podemos hacer cualquier cálculo con ellos (media, varianza, etc.).

  Ejemplo La temperatura medida en ^{o}C, la antigüedad medida en a.d.C., son escalas de intervalo.

  En estas escalas tiene sentido comparar las diferencias entre graduaciones (la diferencia de temperatura, el tiempo transcurrido entre dos sucesos). Sin embargo, no tiene significado físico la proporción o razón entre grados de la escala: un objeto del año 1000 no es el «doble de antiguo» que uno del año 2000, ni una temperatura de 20 ^{o}C indica que haya el «doble de calor» que en una de 10 ^{o}C (esto se entiende perfectamente al ver que si pasamos a grados Kelvin ya no se guarda la proporción, pues 293^{o}K no es el doble de 283 ^{o}K).

- **Escalas de razón**

  Las escalas de razón son de intervalo, y además tienen un cero absoluto, con significado físico.

  Ejemplo La altura en cm, el peso en kg, la temperatura en ^{o}K. Ahora sí que un objeto de 2 metros es el doble de largo que uno de 1 metro.

Precisión de los datos

La precisión se refiere al número de cifras decimales utilizadas para expresar una medida.

Si decimos que un objeto tiene una masa de 10 gramos, estamos indicando que la balanza que hemos usado estaba graduada en gramos, y que hemos observado que nuestra

medición estaba más cerca de 10 que de 9 o de 11, sin poder precisar más. Por tanto en realidad la masa del objeto puede ser cualquiera entre $9'5$ y $10'5$. En otras palabras, nuestro 10 significa $10 \pm 0'5$.

Si en cambio decimos que la masa es $10'0$ gramos es que hemos usado una balanza donde aparecían marcas para los decigramos, y hemos observado que el resultado estaba más cerca de $10'0$ que de la marca anterior ($9'9$) o de la siguiente ($10'1$). Por tanto en realidad la masa puede ser cualquiera comprendida entre $9'95$ y $10'05$. En otras palabras, $10'0$ significa $10'0 \pm 0'05$.

Estos errores en las mediciones, debidos a la sensibilidad e imprecisiones de los instrumentos, así como a los redondeos que realicemos, y los errores sistemáticos o accidentales que podamos cometer, se van propagando y creciendo a medida que hacemos operaciones aritméticas y pueden dar lugar a resultados completamente falsos. Lo mismo veremos que ocurre con algunas fórmulas, que aunque son matemáticamente "equivalentes" a otras, en la práctica pueden dar lugares a errores serios de redondeo, o necesitan muchas más operaciones aritméticas.

2.1.2. Medidas de tendencia central

Para resumir los datos de que dispongamos usaremos varios tipos de descriptores. Nos indican dónde están concentrados nuestros datos (dándonos una idea de su orden de magnitud) y cómo se distribuyen (¿lo hacen de forma simétrica?, ¿están muy agrupados?, ¿están dispersos?).

Las medidas de tendencia central indican dónde están centrados los datos.

Moda Se define la **moda** de unos datos como *el valor más frecuente*. Sirve para cualquier tipo de datos.

Ejemplo En una clase utilizamos la siguiente escala (nominal) para determinar la procedencia de los alumnos:

`1`=A Coruña, `2`=Lugo, `3`=Ourense, `4`= Pontevedra, `5`=otra.

Los porcentajes son respectivamente 38 %, 7 %, 15 %, 30 % y 10 %. Entonces la moda es `1`, es decir, el mayor porcentaje corresponde a A Coruña. Algunos datos pueden tener dos modas (bimodal).

Mediana Para datos de tipo por lo menos ordinal. Una vez ordenados los datos, la **mediana** o valor central es *la puntuación que los divide por la mitad*.

Ejemplo Un corredor de «fórmula 1» ha disputado diez grandes premios y ha quedado en los siguientes puestos:

$$\texttt{pos} = 1, 1, 3, 1, 5, 2, 3, 4, 1, 8. \tag{2.1}$$

Entonces por término medio ha quedado en el puesto `2`–`3`. Para ello basta ordenar los datos

$$\texttt{pos} = 1, 1, 1, 1, 2, 3, 3, 4, 5, 8. \tag{2.2}$$

y observar cuál (o cuáles) está en la posición central. En este ejemplo se tomaría como mediana Md = 2′5.

Media La **media aritmética** puede usarse para datos medidos en una escala de intervalo (o de razón, que es un caso particular), pero no en datos que sean sólo ordinales.

Para datos $X = \{X_1, \dots, X_n\}$ la fórmula es

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

que abreviaremos como

$$\frac{\sum_{i=1}^{n} X_i}{n}$$

o simplemente

$$\overline{X} = \frac{\sum X}{n}. \tag{2.3}$$

Hay otros tipos de medias (armónica, geométrica) que no veremos.

2.1.3. Medidas de posición

Las medidas de posición dan una indicación de la manera en que están distribuídos los datos: si están repartidos simétricamente, con qué ritmo se acumulan, etc.

Cuartiles Para datos ordenados, los cuartiles son las puntuaciones Q_1, Q_2, Q_3 que dividen los datos en cuatro partes iguales. El segundo cuartil es la mediana.

Ejemplo Las notas (ordenadas) de una clase han sido

$$1, 1, 1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 7, 7, 8, 9, 9, 10.$$

La puntuación central (la mediana) es $Q_2 = 4$. El primer cuartil es la mediana de la mitad izquierda

$$1, 1, 1, 2, 3, 3, 4, 4, 4$$

es decir $Q_1 = 3$. Análogamente el tercer cuartil es la mediana de la mitad derecha

$$5, 5, 6, 7, 7, 8, 9, 9, 10$$

es decir $Q_3 = 7$.

Percentiles Esta vez dividimos la distribución en 100 partes iguales. El percentil P_{50} es lo mismo que la mediana. Los otros cuartiles son $Q_1 = P_{25}$ y $Q_3 = P_{75}$.

Ejemplo El tercer cuartil de las posiciones `pos` en las carreras de coches de (2.2) es 4. El percentil P_{60} es 3.

Ejemplo Los diagramas de pesos y de perímetro craneal en las cartilla de salud infantil sirven a los pediatras para determinar en qué percentil se encuentra el niño durante su crecimiento (figura 2.1)

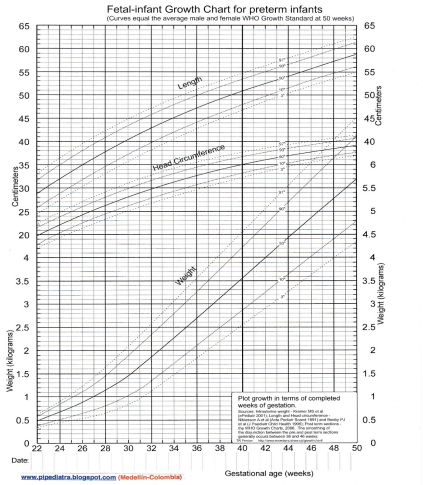


Figura 2.1: Tablas de crecimiento de bebés prematuros

(Tanis Fenton. A new growth chart for preterm babies. BMC Pediatrics (2003)

2.1.4. Medidas de dispersión

Las medidas de dispersión indican si los datos están dispersos o concentrados alrededor de un valor central.

Rango La diferencia entre los dos valores extremos. Por ejemplo para los datos `pos` en (2.2) el rango es $8 - 1 = 7$.

Amplitud intercuartil La diferencia entre Q_3 y Q_1. Para las posiciones `pos` en (2.2) la amplitud intercuartil vale $4 - 1 = 3$.

Desviación mediana Se toman las desviaciones $|X - \mathrm{Md}\,|$ en valor absoluto (es decir sin signo) respecto de la mediana, se ordenan y se calcula su mediana (=valor central).

Ejemplo Para los datos `pos` en (2.2) la mediana era $\mathrm{Md} = 2'5$. Las desviaciones (ya ordenadas) son

$$\texttt{0'5, 0'5, 0'5, 1'5, 1'5, 1'5, 1'5, 1'5, 2'5, 5'5,}$$

cuyo valor central es $1'5$.

2.1.5. Varianza y Desviación típica

Para escalas de intervalo, además de los anteriores, pueden calcularse los siguientes índices de dispersión o variabilidad.

Varianza La varianza s_n^2 es la media de los cuadrados de las desviaciones de los datos con respecto a la media

Es una medida de dispersión inspirada en la fórmula de la distancia entre dos puntos. Para datos $X = \{X_1, \ldots, X_n\}$ se calculan primero las desviaciones respecto de la media

$$x = X - \overline{X}$$

y luego se hace

$$s_n^2 = \frac{\sum x^2}{n} = \frac{\sum (X - \overline{X})^2}{n}. \tag{2.4}$$

En la práctica se usa una fórmula diferente, aunque equivalente:

$$s_n^2 = \frac{\sum X^2}{n} - (\overline{X})^2, \tag{2.5}$$

que se lee: media de los cuadrados menos cuadrado de la media.

Si queremos evitar errores de redondeo tenemos una tercera versión:

$$s_n^2 = \frac{n \sum X^2 - (\sum X)^2}{n^2}.$$

Desviación típica La raíz cuadrada de la varianza se llama *desviación típica* y se denota por s_n (sin el cuadrado).

Cuasivarianza Para inferencia estadística en vez de la varianza es preferible usar la *cuasi-varianza*

$$s_{n-1}^2 = \frac{\sum (X - \overline{X})^2}{n-1}. \tag{2.6}$$

Comparando esta fórmula con la de la varianza (2.4) se ve que la relación entre ambas es

$$s_{n-1}^2 = \frac{n\, s_n^2}{n-1}. \tag{2.7}$$

Por tanto la cuasivarianza es algo mayor que la varianza. Nótese que en la población puede considerarse que son iguales, porque para pasar de una a otra se usa $N/(N-1)$, que es prácticamente igual a 1, ya que el tamaño N de la población es muy grande.

Si queremos una fórmula directa para la cuasi-varianza, es ésta:

$$s_{n-1}^2 = \frac{n \sum X^2 - (\sum X)^2}{n(n-1)}.$$

Cuasi-desviación típica La raíz cuadrada de la cuasivarianza es la cuasi-desviación típica o *desviación típica insesgada* que denotamos por s_{n-1}.

Ejemplo Las concentraciones de un medicamento en cinco muestras de tejido han sido

$$X : 50, 60, 54, 62, 48.$$

La media es $\overline{X} = 274/5 = 54'8$. La varianza es $s_n^2 = 29'76$ y la desviación típica es $s_n = 5'45$. En cambio, la cuasi-varianza es $s_{n-1}^2 = 37'2$ y la cuasi-desviación típica es $s_{n-1} = \sqrt{37'2} = 6'1$.

Propiedades Del examen de la fórmula (2.4) se sigue que:

- la varianza y la desviación típica no pueden ser negativas: $s_n \geq 0$; lo mismo para s_{n-1}.
- pueden ser cero, pero sólo cuando todas las desviaciones $x = X - \overline{X}$ son cero, es decir si todos los datos son iguales a la media y por tanto iguales entre sí.

2.1.6. Disposición de los cálculos

Supongamos que hemos obtenido las siguientes medidas de concentración (mg/l):

$$X : 93, 95, 97, 99.$$

Los datos se disponen como en el cuadro 2.1, y se introducen en la calculadora para obtener las sumas totales.

X	X^2
93	8649
95	...
97	
99	
384	36884

Cuadro 2.1: Disposición de los cálculos de la media y la varianza

Entonces la media es

$$\overline{X} = \frac{\sum X}{n} = \frac{384}{4} = 96,$$

la varianza es

$$s_n^2 = \frac{\sum X^2}{n} - (\overline{X})^2 = \frac{36884}{4} - (96)^2 = 5$$

y la desviación típica es

$$s_n = \sqrt{5} \cong 2'24.$$

La cuasivarianza es

$$s_{n-1}^2 = \frac{4\, s_X^2}{3} = \frac{4 \times 5}{3} \cong 6'67$$

y la desviación típica insesgada es

$$s_{n-1} = \sqrt{6'7} = \cong 2'58.$$

Datos con repeticiones* Es habitual que cada dato aparezca repetido varias veces. Por ejemplo

$$X : 93, 95, 97, 93, 99, 95, 99, 93, 95, 93, 95, 93, 93, 97, 97, 97.$$

En este caso anotamos la frecuencia f con que aparece cada dato y se organizan los datos siguiendo el modelo del cuadro 2.2.

X	f	fX	X^2	fX^2
93	5	465	8649	43245
95	3	...	...	...
97	4			
99	2			
	16	1524		145232

Cuadro 2.2: Datos con repeticiones

Teniendo en cuenta las repeticiones, la media es

$$\overline{X} = \frac{5 \times 93 + 3 \times 95 + \cdots + 2 \times 99}{16} = \frac{1524}{16} = 95'25.$$

La varianza es

$$s_n^2 = \frac{145232}{16} - (95'25)^2 = 4'4375.$$

La desviación típica es

$$s_n = \sqrt{4'7333} = 2'18.$$

La cuasivarianza es

$$s_{n-1}^2 = \frac{16 \times 4'7333}{15} = 5'0488$$

y la cuasidesviación típica es

$$s_{n-1} = \sqrt{5'0488} = 2'25.$$

Por tanto, la fórmula de la media para datos con repeticiones es

$$\overline{X} = \frac{\sum fX}{n} \tag{2.8}$$

y la de la varianza

$$s_n^2 = \frac{\sum fX^2}{n} - (\overline{X})^2. \tag{2.9}$$

También puede usarse la fórmula

$$s_n^2 = \frac{n(\sum fX^2) - (\sum fX)^2}{n^2},$$

que es equivalente pero puede evitar errores de redondeo.

*Las secciones marcadas con asterisco no son materia de examen

2.1.7. Transformación de datos

Si tenemos unos datos X con media $\overline{X}$ y varianza s_n^2, en ocasiones será necesario transformarlos (por ejemplo para cambiar de unidades o de escala). Sean

$$Y = bX + a$$

los datos transformados. Entonces se tiene:

- La media se transforma igual que los datos,

$$\overline{Y} = b\overline{X} + a; \tag{2.10}$$

- La varianza sólo cambia de escala, pero no le influye el desplazamiento a (pues es una medida de dispersión). La fórmula precisa es

$$s_{n,Y}^2 = b^2 s_{n,X}^2. \tag{2.11}$$

- Para la desviación típica, al hacer la raíz cuadrada queda

$$s_{n,Y} = |b| s_{n,X};$$

nótese que la desviación típica no puede ser negativa, por eso aparece el valor absoluto de b.

Las misma fórmulas valen para la cuasivarianza y cuasidesviación típica, es decir

$$s_{n-1,Y}^2 = b^2 s_{n-1,X}^2$$

y

$$s_{n-1,Y} = |b| s_{n-1,X}.$$

Ejemplo Las siguientes temperaturas en grados Celsius

$$X : 22, 30, 26, 29, 22, 29, 23, 27, 29, 27, 29, 29, 21.$$

tienen media

$$\overline{X} = 26'38$$

y desviación típica

$$s_n = 3'25.$$

Al pasarlas a grados Fahrenheit,

$$Y = \frac{9}{5}X + 32$$

tenemos los nuevos datos

$$Y : 71'6, 86'0, 78'8, 84'2, 71'6, 84'2, 73'4, 80'6, 84'2, 80'6, 84'2, 84'2, 69'8.$$

Podemos comprobar que

$$\overline{Y} = (9/5)\overline{X} + 32 = 47'49$$

y

$$s_{n,Y} = (9/5)s_{n,X} = 5'86.$$

Puntuaciones desviadas Por las fórmulas anteriores, las puntuaciones desviadas

$$x = X - \overline{X}$$

tienen media $\overline{x} = 0$ pero la misma desviación típica que las puntuaciones directas, $s_{n,x} = s_{n,X}$.

Para verlo basta tomar $x = bX + a$, con $b = 1$ y $a = -\overline{X}$.

Puntuaciones tipificadas Si tipificamos unas puntuaciones

$$z = \frac{X - \overline{X}}{s_n}$$

pasan a tener media 0 y desviación típica 1.

Para verlo basta tomar $z = bx + a$, con $b = 1/s_n$ y $a = 0$.

2.2. Manejo de la calculadora*

Hay tres tipos de calculadoras:

- aritméticas (para hacer cuentas sencillas: $+, -, \times, /$);
- científicas, que incluyen funciones (como sen, cos, log, exp), memorias y «modos»: por ejemplo `COMP` (cálculos), `SD` o `STAT` (estadística), `LR` o `REG` (regresión lineal, regresión);
- y las calculadoras programables.

Vamos a dar una idea general de cómo usar las calculadoras científicas, aunque las teclas concretas varían con cada modelo. Es indispensable leer el manual. En internet es fácil encontrar el manual de cualquier modelo y hay websites como éste: `http://support.casio.com/en/manual/manuallist.php?cid=004`.

Modo aritmético Podemos borrar el último dato introducido (`C` *clear*) o todos (`AC` *all clear*). Normalmente hay al menos una memoria donde podemos introducir un primer dato (`Min`), lo que borra el contenido anterior; ir añadiendo más datos (`M+`) para irlos sumando; y leer el contenido de la memoria (`MR`).

Ejemplo Calcular $50'32 + 3'1 \times 2'8 = 59$. Si la calculadora tiene paréntesis, escribiremos $50.32 + (3.1 \times 2.8)$; si tiene pantalla gráfica bastará poner $50.32 + 3.1 \times 2.8$ pues respeta la prioridad en las operaciones. Con las memorias la secuencia es

$$50.32 \quad \texttt{Min} \quad 3.1 \times 2.8 \quad \texttt{M+} \quad \texttt{MR} \quad 59.$$

Modo estadístico Hay que activar el modo estadístico, que se localiza con la tecla `mode` o similar. Suele llamarse `SD` o `STAT`. Es muy importante borrar la memoria estadística, que es diferente de la memoria usual, para eliminar datos que hayan podido quedar almacenados. Se consigue, según el modelo de calculadora, por ejemplo haciendo `CLR`, `Scl` o `SCL`.

Para introducir los datos se usa la tecla `DATA` o `DT`, haciendo una secuencia del tipo

$$50 \quad \texttt{DT} \quad 60 \quad \texttt{DT} \quad 54 \quad \texttt{DT} \quad 62 \quad \texttt{DT} \quad 48 \quad \texttt{DT}.$$

*Las secciones marcadas con asterisco no son materia de examen

Ahora hay que comprobar el número de datos introducidos $\mathtt{n} = 5$ (está en el teclado o en S-SUM).

La calculadora nos proporcionará (bien directamente desde el teclado o a través de la pantalla gráfica en S-SUM o S-VAR) los valores de

$$\sum \mathtt{X},\ \sum \mathtt{X}^2,\ \overline{\mathtt{X}},\ \mathtt{s}_\mathtt{X}, \hat{\mathtt{s}}_\mathtt{X}.$$

Modo regresión lineal Se activa con LR o REG. En este modo es posible introducir simultáneamente dos variables X, Y en una secuencia del tipo

1 , 20.52 DT ... 5 , 58.72 DT

(la coma de separación de cada par de datos depende de los modelos). La calculadora nos dará los estadísticos de las dos variables (por ejemplo $\overline{\mathtt{X}}$, $\mathtt{s}_\mathtt{X}$, $\overline{\mathtt{Y}}$ y $\mathtt{s}_\mathtt{Y}$), los datos conjuntos ($\sum \mathtt{XY}$) y los parámetros de la regresión b, a y r

Calculadoras en red En caso desesperado, es posible conectarse con alguna calculadora en red, por ejemplo http://web2.0calc.com/

Capítulo 3

Expositiva 3

Comenzaremos recordando el uso que hacemos del Teorema fundamental del cálculo (Regla de Barrow) para calcular la función de distribución de una variable aleatoria usando tablas de integrales de su función de densidad. Después estudiamos los siguientes puntos del programa:

1. Población y muestra.
2. Parámetro. Estadístico.
3. Distribución de diferentes estadísticos. Teorema central del límite.
4. Estimación puntual. Propiedades de los estimadores.
5. Estimación por intervalos de confianza: conceptos básicos. Nivel de confianza.
6. Intervalos de confianza para la media, la varianza y la proporción.
7. Determinación del tamaño de la muestra.

3.1. Variables aleatorias continuas*

Si queremos conocer la altura media de la población de Galicia estudiaremos la *variable aleatoria* X = altura en cm. Es una variable *continua*.

Función de distribución Podemos calcular su *función de distribución*, es decir las probabilidades (frecuencias relativas teóricas) acumuladas:

$$F(x) = p(X \leq x).$$

Su gráfica tendrá un aspecto parecido al de la Figura 3.1.

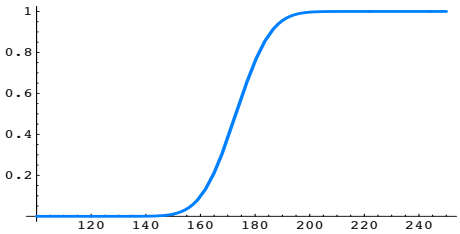


Figura 3.1: Función de distribución

Ejemplo Si X es la variable "altura" de una persona en cms., que X no supere 10 cm. es prácticamente imposible (probabilidad nula), y en cambio que no supere 190 es prácticamente seguro (probabilidad 1).

En la primera parte del curso se estudiaron también variables aleatorias discretas como la binomial, pero aquí las aproximaremos por continuas como la normal.

Función de densidad El ritmo de crecimiento de la función de distribución estará dado por su función derivada, que se llama *función de densidad*, $f(x) = F'(x)$. Su gráfica será parecida a la de la figura 3.2.

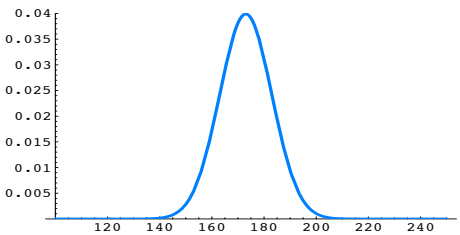


Figura 3.2: Una función de densidad, en este caso normal

Sin embargo, habrá funciones de densidad que no son simétricas, o sólo están definidas para valores positivos, como la "chi-cuadrado" de Pearson o la "F" de Snedecor.

*Las secciones marcadas con asterisco no son materia de examen

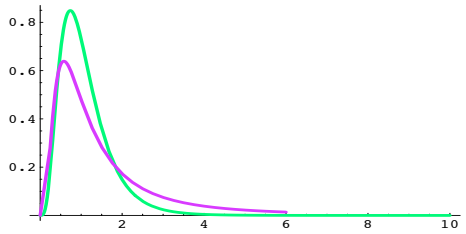


Figura 3.3: Dos curvas F de Snedecor

Área bajo la gráfica Por la regla de Barrow (Teorema fundamental del Cálculo), como $F(x)$ es una primitiva de $f(x)$, si queremos calcular el área bajo $f(x)$ en un intervalo $[a, b]$ bastará calcular $F(b) - F(a)$. En particular, como $\lim_{x\to-\infty} F(x) = 0$ tenemos

$$F(b) = \int_{-\infty}^{b} f(x)dx.$$

En otras palabras,

la probabilidad $p(X \leq b)$ de que la variable X tome valores inferiores a un valor determinado b coincide con el área a la izquierda del valor b bajo la gráfica de la función de densidad.

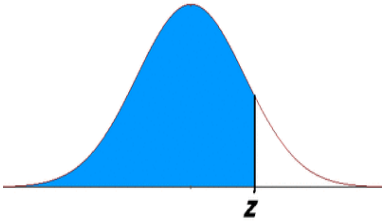


Figura 3.4: Área a la izquierda de un valor z en una función de densidad normal

Para las distribuciones más usuales calcularemos esas integrales por medio de *tablas* o con un programa informático.

Ejemplo Tenemos una variable z que sigue una distribución normal estándar. Queremos saber cuál es la probabilidad de que z valga como mucho $1'96$. Tenemos que calcular, para la curva normal $f(x)$, el área a la izquierda

$$\int_{-\infty}^{1'96} f(x)dx.$$

Recurrimos a un programa informático como EXCEL o a unas tablas y obtenemos

$$p(z \leq 1'96) = 0'975002105.$$

Esperanza Por analogía con la fórmula de la media (con valores repetidos, fórmula (2.8)), se define la *esperanza* o *valor esperado* de la variable X como

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx. \tag{3.1}$$

La media o esperanza de la población se denotará por μ.

Varianza Por analogía con la fórmula (2.5) de la varianza de unos datos, se define la *varianza* de X como

$$V(X) = E(X^2) - E(X)^2.$$

La varianza de la población se denota por σ^2.

Teorema de Chebichev* La esperanza μ es una medida de tendencia central, y la desviación típica σ es una medida de dispersión. Sea cual sea la distribución de X, al menos $3/4$ de sus valores están agrupados en el intervalo $(\mu - 2\sigma, \mu + 2\sigma)$. Más en general, tenemos el llamado «Teorema de Chebichev»:

Fuera del intervalo $\mu - k\sigma, \mu + k\sigma$ está como mucho una proporción de $1/k^2$ de los valores.

Este resultado puede mejorarse si conocemos la forma de la distribución. Por ejemplo si es «normal» entonces en el intervalo $(\mu - 2\sigma, +\mu + 2\sigma)$ está algo más del 95 % de los valores y en el intervalo $\mu \pm 3\sigma$ más del 99 %. Esto se calcula con tablas, tipificando.

La curva normal La *curva normal estándar* o campana de Gauss* es la gráfica de la función

$$z = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Cuando la variable no está tipificada usaremos las curvas normales $N(\mu, \sigma)$, de media μ y desviación típica σ, cuya fórmula es

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{\sigma})^2/2}.$$

Su gráfica está centrada en la media μ y tiene puntos de inflexión en $\mu \pm \sigma$.

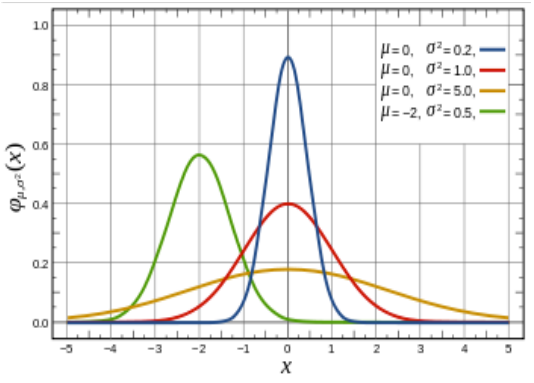


Figura 3.5: Distintas curvas normales. Tomado de Wikimedia Commons

El *Teorema central del límite* establece que cuando una variable X es el promedio de muchas variables arbitrarias (con tal de que tengan media y varianza finitas) e independientes entonces la función de densidad de X se ajusta aproximadamente a una curva normal. Por eso eso esta curva aparece en tantas situaciones diferentes.

*Las secciones marcadas con asterisco no son materia de examen

*Al final del libro tienes una pequeña biografía de cada matemático que aparece citado en este curso

3.2. Inferencia estadística

En este curso estudiaremos métodos y técnicas de *inferencia estadística* que permiten conocer, a partir de la información proporcionada por una muestra aleatoria, cuales son los parámetros de una determinada población, y ser capaces de determinarlos con un riesgo de error medible en términos de probabilidad.

Por ejemplo, si queremos conocer la media de una población no podemos medir directamente a todos los individuos de la población, por lo que debemos escoger una *muestra aleatoria* de cierto tamaño n y calcular la media de la muestra.

La media de la población se denota por μ. La media de la muestra se denota por $\overline{X}$.

Estimación puntual Si queremos conocer la altura media μ de las mujeres de Galicia, seleccionamos aleatoriamente (por sorteo) una muestra de 200 mujeres y calculamos su media. Supongamos que hemos obtenido $\overline{X} = 177$. Podemos hacer una *estimación puntual* $\mu \sim 177$ de la media de toda la población de mujeres. El problema de este tipo de estimación es que desconocemos si nuestra muestra era representativa de la población, es decir no podemos medir la probabilidad de que nuestra estimación sea acertada.

3.3. Distribución muestral de un estadístico

Otro problema es por qué hemos escogido la media muestral para estimar la media poblacional. La justificación se basa en el siguiente resultado:

Fijemos un tamaño de muestra n. Si calculásemos la media muestral en todas las muestras de ese tamaño n, tendríamos una variable teórica $\overline{X}$ (varía en cada experimento). Entonces se tiene que el valor esperado de esa variable $\bar{X}$ es μ (la media de la población). Es decir, aunque los resultados de $\overline{X}$ varíen de muestra en muestra, en conjunto oscilan alrededor de μ.

Se escribe $E(\overline{X}) = \mu$ y se dice que la media muestral es un *estimador insesgado* de la media poblacional. Por eso es adecuado usar ese *estadístico* $\overline{X}$ para estimar el *parámetro* μ.

Aparte de no tener sesgo, hay otras propiedades de los «buenos» estimadores que pueden tenerse en cuenta (pero no las veremos).

Justificación* Supongamos que hemos decidido que el tamaño de la muestra sea n, y que llamamos $X_1, \ldots, X_n$ a las medidas de cada ejemplar de la muestra. Entonces la variable $\overline{X} = (X_1 + \cdots + X_n)/n$ tiene un valor esperado de

$$E(\overline{X}) = (E(X_1) + \cdots + E(X_n))/n = n\mu/n = \mu,$$

ya que el valor esperado de cada medida X_i es μ. Hemos usado que el operador $E()$ definido en (3.1) conserva la suma y el producto por escalares.

Puedes "experimentar" la distribución muestral de un estadístico si buscas alguna página web con "sampling distribution", por ejemplo esta (necesitas un navegador que soporte Java)
`http://onlinestatbook.com/stat_sim/sampling_dist/index.html`

*Las secciones marcadas con asterisco no son materia de examen

Capítulo 4

Expositivas 4–5

Estimación por intervalos La estimación puntual que hemos visto anteriormente no nos da ninguna indicación de si la estimación que hemos hecho es buena o mala. Para remediarlo vamos a hacer *estimaciones por intervalos*. Para eso necesitaremos saber no sólo que $\overline{X}$ oscila alrededor de μ sino cómo lo hace, es decir, saber qué distribución tiene la variable $\overline{X}$.

Una estimación por intervalos es del tipo

$$\mu = \overline{X} \pm \text{error}.$$

La ventaja de este método es que podremos calcular «la probabilidad de acertar» en la estimación (*nivel de confianza*).

Necesitaremos que se cumpla alguna de estas dos condiciones:

- o bien que la muestra sea suficientemente grande (teorema central del límite)
- o bien que la variable de partida X sea normal (en este caso vale cualquier tamaño de muestra).

En estos dos casos la distribución del estadístico $\overline{X}$ será normal.

Intervalos Para dos números a, b arbitrarios y un número positivo ε, las siguientes afirmaciones son equivalentes (puedes convencerte haciendo un dibujo sobre la recta real):

- La distancia entre a y b es menor o igual que ε;
- $|a - b| \leq \varepsilon$;
- $a - \varepsilon \leq b \leq a + \varepsilon$;
- $b - \varepsilon \leq a \leq b + \varepsilon$.

Si ahora consideramos la media poblacional μ y la media $\overline{X}$ de la muestra de un experimento, decir que μ está en el intervalo $\overline{X} \pm$ error es lo mismo que decir que la distancia entre $\overline{X}$ y μ es menor que el error de estimación:

$$|\overline{X} - \mu| < \text{error}$$

y por tanto es lo mismo que decir que $\overline{X}$ está en el intervalo $\mu \pm$ error.

La probabilidad de que ocurra esto último al repetir el experimento muchas veces es

$$p\left(\mu - \text{error} \leq \overline{X} \leq \mu + \text{error}\right)$$

y puede calcularse mediante las tablas adecuadas porque $\overline{X}$ es una variable aleatoria con distribución conocida. Esta probabilidad se llama *nivel de confianza.*

Ejemplo En un experimento sobre el peso de un tumor sabemos que la media muestral $\overline{X} = 0'12$ g. Queremos hacer una estimación de la media poblacional μ con un error menor de $0'07$ g. Obtendremos un intervalo $\mu = 0'12 \pm 0'07$. Con las tablas adecuadas seremos capaces de calcular la probabilidad de que esto sea cierto. En realidad, lo que calcularemos será la probabilidad de que $\overline{X}$ caiga en el intervalo $\mu \pm 0'07$.

4.1. Estimación de la media de la población

Hay dos situaciones posibles, según que conozcamos o no la varianza de la población.

4.1.1. Conocida la varianza poblacional

Es la situación más sencilla. Supongamos que la variable X que estudiamos es normal, con media μ (que no conocemos) y desviación típica σ (que sí conocemos, quizás de experimentos anteriores). Usaremos el siguiente resultado:

Distribución muestral del estadístico $\overline{X}$: La variable teórica $\overline{X}$ aparecería al hacer muchas veces el mismo experimento y anotar cada vez el valor de la media muestral.
Si tomamos muestras de tamaño n entonces el estimador $\overline{X}$ también es normal. Tiene media $E(\overline{X}) = \mu$ y varianza $V(\overline{X}) = \sigma^2/n$. Por tanto la desviación típica de $\overline{X}$ es $\sigma/\sqrt{n}$.

Esto significa que al tipificar la variable teórica $\overline{X}$, es normal estándar $N(0,1)$:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}. \tag{4.1}$$

El intervalo de estimación tendrá la forma $\mu = \overline{X} \pm$ error, donde el error , como se ve en la fórmula 4.1, vale $z\sigma/\sqrt{n}$. Escribiremos el intervalo como

$$\mu = \overline{X} \pm z\frac{\sigma}{\sqrt{n}}. \tag{4.2}$$

Nivel de confianza El valor de z que aparece en la fórmula (4.3) depende de cuál queremos que sea el área cubierta por nuestro intervalo de confianza. Esa área (probabilidad) es el *nivel de confianza.* El área que queda fuera del intervalo se llama *nivel de significación* y se representa por α.

Como el intervalo deja colas a ambos lados, y puesto que la curva normal es simétrica, bastará buscar en las tablas el valor $z_{\alpha/2}$ que deja a su derecha un área de $\alpha/2$.

Escribiremos el intervalo como

$$(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}). \tag{4.3}$$

El *nivel de confianza* vale $1-\alpha$ y se interpreta como la probabilidad de que la estimación sea cierta. El nivel de significación α indicará que hay una cierta probabilidad de que nuestra muestra no sea representativa.

Ejemplo En un experimento con $n = 100$ hemos obtenido $\overline{X} = 177$ y queremos dar un intervalo de estimación que tenga una probabilidad de acertar del 95 %. Por experimentos anteriores sabemos que la desviación típica de la población es $\sigma = 4$.

Como el nivel de confianza es $0,95$, el nivel de significación es $\alpha = 0'05$. El valor $z_{\alpha/2}$ esn las tablas de la normal es $1,96$. Por tanto el error de estimación es

$$1'96\,\sigma/\sqrt{n} = 1'96\,\frac{4}{\sqrt{100}} = 0,784.$$

Distribución muestral del estadístico Es importante distinguir las variables X y $\overline{X}$. La variable que estudiamos en la población es X. Suponemos que es normal, tiene media $E(X) = \mu$ y varianza $V(X) = \sigma^2$. En cambio, el estimador $\overline{X}$ (media muestral) es una variable, de tipo teórico, que se calcula en cada experimento que realizamos. Tiene media $E(\overline{X}) = \mu$ y varianza $V(\overline{X}) = \sigma^2/n$.

Veamos un ejemplo sencillo.

Supongamos que nuestra población está formada solamente por tres personas A, B, C, de tallas $X = 150, 160, 170$ cm respectivamente. Entonces $\mu = 160$ y $\sigma = 8'16$ (parámetros de la población).

Imaginemos que tenemos que hacer un experimento para estimar μ, y que por motivos económicos sólo podemos tomar muestras de tamaño $n = 2$. Si la muestra obtenida en el sorteo es A, C, la media de la muestra es $\overline{X} = 160$. En cambio, si obtenemos la muestra A, B, la media será $\overline{X} = 155$. Por último, si la muestra obtenida fuese B, C el experimento daría $\overline{X} = 165$. Así, los posibles valores de $\overline{X}$ son $155, 160, 165$, que tienen media $E(\overline{X}) = 160 = \mu$.

Si calculamos la desviación típica de todos los resultados posibles, obtendremos $5'77$, que coincide con $\sigma/\sqrt{n} = 8'16/\sqrt{2}$.

Resumen del procedimiento de cálculo

Queremos estimar μ. Suponemos que la variable X es normal o que la muestra es grande. También suponemos que conocemos σ. El estimador insesgado es $\overline{X}$.

- Se fija un nivel de confianza, por ejemplo $0'95$.
- El estadístico $\overline{X}$ tiene una distribución normal. Al tipificarlo tiene una distribución z (normal estándar o $N(0,1)$).
- Se calcula el valor $z_{\alpha/2}$ que deja a su derecha un área de $\alpha/2 = 0'05/2 = 0'025$; se obtiene en las tablas $z_{0'025} = +1'96$.
- El intervalo es
$$\overline{X} \pm 1'96\frac{\sigma}{\sqrt{n}}.$$

4.1.2. Desconocida la varianza poblacional

El procedimiento anterior requería conocer σ, la desviación típica de la población. En el siguiente caso, cuando no se conoce σ, puede tomarse este otro estimador:

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \tag{4.4}$$

donde $s = s_{n-1}$ es la *cuasi-desviación típica* de la muestra, que es un estimador de σ. La curva a la que se ajusta (4.4) ya no es normal, sino que es la llamada *distribución* t de Student con $n-1$ grados de libertad.

La t de Student

La t de Student es una familia de curvas parecidas a la normal, que se usa cuando no se conoce la desviación típica de la población y sólo se conoce la de la muestra.

Las funciones de densidad t_n fueron inventadas por W. Gosset (cuyo pseudómino era «Student») en 1908 *. Tienen como ecuación

$$f(x) = c_n(1 + \frac{x^2}{n})^{-(n+1)/2},$$

donde n se llama el *número de grados de libertad*.

- La constante c_n se pone para que el área total bajo la gráfica sea 1.
- La media de t_n es cero; es simétrica y su varianza es $n/(n-2)$.
- La gráfica de t_n es parecida a la de la curva normal estándar (ver figuras 4.1 y 4.2).
- Para $n \sim 100$ grados de libertad ya se considera que t_{100} es prácticamente igual a z.

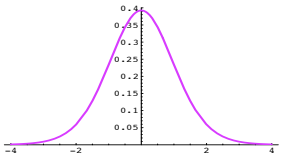


Figura 4.1: La t de Student con 15 grados de libertad

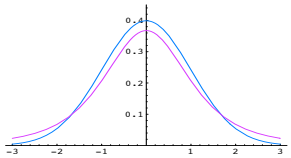


Figura 4.2: Comparación entre t_3 (violeta) y z (azul)

*Al final del libro tienes una pequeña biografía de cada matemático que aparece citado en este curso

Procedimiento de cálculo

Queremos estimar μ. Suponemos que la variable X es normal o que la muestra es grande. No conocemos σ. El estimador insesgado es $\overline{X}$.

- Tenemos una muestra por ejemplo de tamaño $n = 20$.
- Se fija un nivel de confianza, por ejemplo $0'99$.
- El estadístico $\overline{X}$ tiene una distribución t con $n - 1 = 19$ grados de libertad.
- En la tabla de t_{19} se calcula el valor $t_{\alpha/2}$ que deja a su derecha un área de $\alpha/2 = 0'01/2 = 0'005$; se obtiene en las tablas $t_{0'005} = 2'861$.
- Calculamos la media $\overline{X}$ y la cuasi-desviación típica s de la muestra.
- El intervalo es
$$\overline{X} \pm 2'861 \frac{s}{\sqrt{20}}.$$

4.1.3. Tamaño de la muestra

En ocasiones queremos limitar el error de estimación para que no sobrepase un cierto valor máximo. Para eso debemos aumentar el tamaño de la muestra. Como queremos

$$\text{error} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \text{errormax},$$

basta despejar

$$n \geq \left(\frac{z_{\alpha/2}\, \sigma}{\text{errormax}} \right)^2.$$

Para tener una idea de σ puede hacerse un *experimento piloto* con un tamaño de muestra reducido y calcular s.

Ejemplo En un experimento se quiere estimar el peso medio de un tumor en ratones sometidos a un determinado tratamiento. Se quiere que el error de estimación sea inferior a $0'01$ g y que el nivel de significación (probabilidad de acertar) sea del 95 %. Si sabemos que $\sigma = 0'07$, ¿Cuántos ratones habrá que analizar?

SOL. Como $\alpha = 0'05$ y $z_{\alpha/2} = 1'96$, será $n \geq (1'96 \times 0'07/0'01)^2 = 188'24$, es decir como mínimo habrá que usar 189 ratones.

No hace falta saber la fórmula de memoria, sólo plantear

$$\text{error} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq 0'01$$

y despejar n.

No confundas estas tres cosas:

1. la variable X que estamos estudiando en una población, que tiene su valor medio $E(X) = \mu$ y su varianza $V(X) = \sigma^2$;
2. los valores obtenidos en una muestra concreta, que tienen una media $\overline{X}$ y una varianza s^2;
3. la variable teórica $\overline{X}$ que se obtiene repitiendo muchas veces un experimento y anotando cada vez la media muestral; esta variable tiene su propia media $E(\overline{X})$ y varianza $V(\overline{X})$.

4.2. Estimación de una proporción

Supongamos que tenemos una variable nominal dicotómica (por ejemplo el sexo X con dos valores posibles `1` =«mujer», `0` =«hombre»). En una muestra de tamaño n la proporción de mujeres será $\widehat{p}$ y la de hombres $\widehat{q} = 1 - \widehat{p}$.

Nota: vamos a seguir la notación del Milton: p es la proporción en la *población* y $\widehat{p}$ será la proporción en la muestra.

Ejemplo La muestra puede ser `1, 0, 0, 1, 1, 0, 1, 1, 1, 0`. Nótese que si calculamos la «media» de la muestra nos dará $\widehat{p} = 0'6$ y si calculamos la «varianza» nos dará $\widehat{p}\widehat{q} = 0'6 \times 0'4 = 0'24$.

Queremos estimar la proporción p de mujeres en la población. El estimador $\widehat{p}$ (proporción muestral) sigue una distribución *binomial* con parámetros n, p. En efecto, podemos considerar que al estudiar la muestra hicimos n intentos o ensayos independientes, que en cada ensayo había dos posibilidades («éxito» o «fracaso») y que la probabilidad de «éxito» en cada intento era p.

Sólo consideraremos el caso en que las muestras son grandes y podemos aproximar la binomial por una normal. En ese caso se obtiene un intervalo completamente análogo al de una media, es decir

$$p = \widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}}. \tag{4.5}$$

Ejemplo En una facultad queremos estimar la proporción de mujeres. En una muestra aleatoria de 100 personas, hemos encontrado 60 mujeres y 40 hombres. Para un nivel de confianza del 95 % será $\alpha/2 = 0'025$ y en las tablas obtenemos $z_{0'025} = 1'96$. Entonces

$$p = 0'6 \pm 1'96\sqrt{0'24/100} = 0'6 \pm 0'0960,$$

con lo que obtenemos el intervalo $(0'5, 0'7)$; es decir la proporción de mujeres está entre el 50 % y el 70 % para la proporción poblacional.

Aplicación: estimación del tamaño de la población (pág. 263) Queremos estimar cuántos lobos hay en una determinada zona geográfica. Se capturan 10 ejemplares, se marcan y se liberan. Al cabo de un tiempo se capturan otros ocho ejemplares y se observa que hay dos marcados. Por tanto la proporción de marcados en la muestra es $\widehat{p} = 2/8 = 0'25$. Si hacemos la estimación $p \cong 0'25$, como la proporción de marcados en la población es $p = 10/N$, deducimos que el tamaño de la población es $N \cong 10/0'25 = 40$.

4.2.1. Tamaño de la muestra

Ya hemos visto, en algunos contrastes para una media, cómo estimar el tamaño de muestra necesario. Veámoslo en el caso de una proporción.

Supongamos que queremos que el error de estimación de *una proporción* no sobrepase un valor máximo d. Entonces basta tomar

$$n \geq \frac{z^2}{4d^2}.$$

Ejemplo (adaptado de 8.3.3 pág. 269) En un estudio sobre anemia falciforme queremos que el error de estimación para la proporción de personas enfermas no supere el 1 %. Por otro lado el nivel de confianza que necesitamos es del 99 %. Para $\alpha/2 = 0'01/2 = 0'005$ encontramos $z = \pm 2'56$. Entonces

$$n \geq (2'56)^2/(4 \times (0'01)^2) = 16384$$

por lo que necesitamos una muestra de casi 17000 personas (!)

Dato: `DISTR.NORM.ESTAND.INV(0,005)=-2,575829304`.

Justificación* Para una proporción, el error de estimación es

$$\text{error} = z\sqrt{\frac{\widehat{pq}}{n}},$$

que depende de cuánto valga la proporción muestral $\widehat{p}$. Ahora bien, el producto $\widehat{pq} = \widehat{p}(1-\widehat{p})$ no puede pasar de $1/4$ (es el valor máximo de la función $x - x^2$ en el intervalo $0 \leq x \leq 1$). Usando esto, siempre se cumple

$$\widehat{pq} \leq \frac{1}{4}.$$

con lo que

$$\text{error} \leq z\sqrt{\frac{1}{4n}} = \frac{z}{2\sqrt{n}}.$$

Si queremos que el error no sea mayor que d, nos ponemos "en el peor caso posible", es decir cuando $\widehat{pq} = 0,25$, con lo que basta considerar la desigualdad $z/2\sqrt{n} \leq d$, o lo que es lo mismo $n \geq z^2/4d^2$.

4.3. Estimación de la varianza

Las curvas χ^2_n. Cuando tenemos varias variables normales estándar independientes $z_1, \ldots, z_n$, la variable suma de cuadrados $z_1^2 + \cdots + z_n^2$ se ajusta a una curva que se llama χ^2_n (se lee *ji-cuadrado con n grados de libertad*). Su fórmula es

$$f(x) = c_n x^{\frac{n}{2}-1} e^{\frac{-x}{2}}, \quad x \geq 0.$$

Cada curva chi-cuadrado χ_n:

- sólo está definida para los valores no negativos,
- no es simétrica,
- su media es n (pero no coincide con el máximo, que es $n-2$ y su varianza es $2n$ (ver Figura 4.3).
- La constante c_n sirve para que el área total bajo la gráfica sea 1.

*Las secciones marcadas con asterisco no son materia de examen

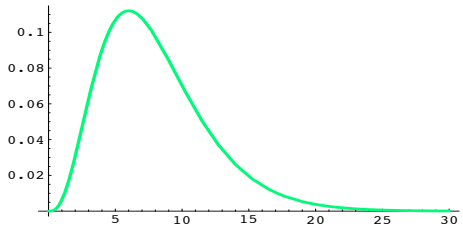


Figura 4.3: χ^2 con 8 grados de libertad

Esta distribución fue descubierta por el matemático inglés K. Pearson en 1900 [†].

Para valores de n altos ($n \geq 30$) podemos usar las tablas de la curva normal de acuerdo con la siguiente transformación (ver Sixto Ríos, Métodos estadísticos, 1967, pág. 237):

$$z \cong \sqrt{2\chi_n^2} + \sqrt{2n-1}.$$

Estimador insesgado de la varianza

Para estimar la varianza de la población σ^2 no es adecuado usar la varianza de la muestra. Usaremos la cuasivarianza de la muestra.

En efecto, el estimador s_n^2 oscila alrededor de $E(s_n^2) = \frac{n-1}{n}\sigma^2$ (teorema de Fisher). En cambio si tomamos la *cuasivarianza muestral* obtenemos

$$E(s_{n-1}^2) = \sigma^2.$$

Nótese que en la *población* (que tiene un tamaño grande N) la cuasi-varianza sería $\frac{N\sigma^2}{N-1} \cong \sigma^2$ porque $\frac{N}{N-1} \cong 1$. Por tanto la diferencia entre varianza y cuasivarianza sólo es importante en muestras pequeñas.

Intervalo de estimación para σ^2 El resultado teórico que necesitamos es:

Si la variable X que estamos estudiando es normal, con varianza σ^2, entonces el estadístico

$$\frac{(n-1)s^2}{\sigma^2}$$

tiene una distribución χ^2 con $n-1$ grados de libertad. Aquí, s es la cuasi-desviación típica de la muestra, que varía de experimento en experimento.

Por tanto debemos encontrar los valores

$$a = \chi^2_{1-\alpha/2}, \quad b = \chi^2_{\alpha/2}$$

correspondientes al nivel de significación α y tomar

$$a \leq \frac{(n-1)s^2}{\sigma^2} \leq b \tag{4.6}$$

lo que despejando σ^2 nos da el intervalo

$$\frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a}. \tag{4.7}$$

[†]Al final del libro tienes una pequeña biografía de cada matemático que aparece citado en este curso

Nótese que por la relación entre varianza y cuasi-varianza tenemos

$$(n-1)s^2 = \sum (X - \overline{X})^2 = ns_n^2.$$

Ejemplo (p. 256) En una raza de perros interesa que las características sean homogéneas, por lo que queremos estimar la varianza σ^2 de la altura. Con ello comprobaremos si ha sido efectivo cierto procedimiento de cría selectiva. En una muestra de 15 perros observamos una cuasi-varianza $\widehat{s}^2 = 0'21$. Usaremos un nivel de confianza del 90 %.

SOL.: Como $\alpha = 0'10$, en las tablas de la curva χ^2 con $n - 1 = 14$ grados de libertad obtenemos los valores $b = \chi^2_{0'05} = 23'685$ y $a = \chi^2_{0'95} = 6'571$. Entonces de (4.6) obtenemos

$$6'571 \leq \frac{14 \times 0'21}{\sigma^2} \leq 23'685$$

o lo que es lo mismo,

$$\frac{14 \times 0'21}{23'685} \leq \sigma^2 \leq \frac{14 \times 0'21}{6'571}$$

con lo que tenemos

$$0'1241 \leq \sigma^2 \leq 0'4474.$$

Es decir la varianza de la raza de perros (en la población) está con mucha certeza en el intervalo $(0'12, 0'45)$. Si necesitásemos un intervalo para la desviación típica tomaríamos la raíz cuadrada,

$$\sqrt{0'1241} \leq \sigma^2 \leq \sqrt{0'4474},$$

es decir

$$0'35 \leq \sigma \leq 0'67.$$

4.4. Estimadores*

Estimador de la media Para estimar la media poblacional μ hemos usado la media muestral $\overline{X}$. Decimos que $\overline{X}$ es un *estimador* de μ. Aunque no lo veremos en este curso, el motivo para escoger un estimador en vez de otro es que tenga determinadas buenas propiedades; por ejemplo, hemos dicho que si hacemos un experimento reiteradas veces, los distintos resultados $\overline{X}$ oscilan alrededor de μ. En términos técnicos, se dice que el valor esperado de $\overline{X}$ es μ, y se escribe $E(\overline{X}) = \mu$. Otra forma de decirlo es que $\overline{X}$ es un estimador *sin sesgo* del parámetro μ.

Un cálculo parecido nos muestra que la "varianza" del estimador $\overline{X}$ a lo largo de los experimentos es $V(\overline{X}) = \sigma^2/n$.

Estimador de la varianza Ahora supongamos que queremos estimar la varianza de la población

$$\sigma^2 = V(X) = E(X^2) - E(X)^2$$

y usásemos la varianza de la muestra

$$s_n^2 = \frac{\sum X_i^2}{n} - \overline{X}^2.$$

*Las secciones marcadas con asterisco no son materia de examen

El valor esperado de este estimador sería

$$E(s_n^2) = \frac{n-1}{n}\sigma^2$$

así que NO es un estimador insesgado de la varianza. Por eso se usa la cuasivarianza muestral, ya que el cambio

$$s_{n-1}^2 = \frac{n}{n-1}s_n^2$$

corrige el sesgo.

Otras propiedades Otras propiedades interesantes de un estimador, además de no tener sesgo, son que sea *consistente* (aproxima mejor el parámetro cuanto mayor sea el tamaño de la muestra) y *eficiente* (no tiene demasiada variabilidad).

Tema 2: Contraste de hipótesis

Capítulo 1

Expositivas 6–10

- Hipótesis estadística.
- Tipos de error. Nivel crítico o p-valor. Potencia de un contraste.
- Contrastes con una muestra: para una media, para una varianza y para una proporción.
- Contrastes con dos muestras: comparación de dos medias; comparación de dos varianzas; comparación de dos proporciones.

Podemos descargar el software *R* en `http://www.r-project.org/`. Es un programa gratuito para cálculo y gráficas de Estadística. Funciona con UNIX, Windows y MacOS. Por ahora nos servirá para calcular los valores de las tablas.

Retomamos las ideas que hemos estado usando en la estimación por intervalos, pero desde otro punto de vista.

1.1. Contraste de hipótesis para la media poblacional

1.1.1. Motivación

Ejemplo básico Estudiamos el crecimiento anual de los abetos. Creemos que el valor medio de esta variable es $\mu = 7'25$. Sin embargo en una muestra de 50 árboles se obtuvo $\overline{X} = 7'27$ (y $s = 0'03$). ¿Es este resultado compatible con nuestra suposición?

Queremos *contrastar una hipótesis* sobre la población: ¿es $\mu = 7'25$? La llamaremos *hipótesis nula* H_0 (porque contiene el signo "="). Fijamos un *nivel de confianza* (por ejemplo del 95 %). Esto significa que consideramos muy improbable un resultado que se obtenga en menos del 5 % de los experimentos. Ahora realizamos el experimento y debemos decidir si el resultado obtenido en la muestra entra dentro de lo «probable» o es «muy improbable». En este último caso deberíamos reconocer que nuestra hipótesis de partida no parece cierta, y tendríamos que rechazarla.

Usamos el estimador $\overline{X}$ que ya conocemos. Si H_0 fuese cierta, el estadístico

$$\frac{\overline{X} - 7'25}{s/\sqrt{50}}$$

seguiría una distribución t_{49}. Como $t_{\alpha/2} = 2'01$ (tablas), el estadístico oscila alrededor de cero, y estará comprendido, con probabilidad $0'95$, entre los valores $t = -2'01$ y $t = +2'01$.

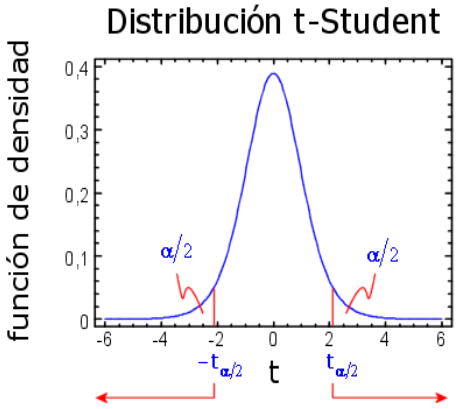


Figura 1.1: tomado de http://e-stadistica.bio.ucm.es

Ahora lo calculamos en nuestra muestra y obtenemos

$$\frac{7'27 - 7'25}{0'03/\sqrt{50}} = 4'71,$$

valor que está fuera del *intervalo de aceptación* $(-2'01, +2'01)$. Por tanto la diferencia entre la hipótesis $\mu = 7'25$ y el valor muestral $\overline{X} = 7'27$ es *significativa*: no parece deberse sólo a la oscilación «razonable» del estadístico. En consecuencia *rechazamos* H_0.

Nótese que siempre queda la pequeña posibilidad de que nos hayamos equivocado, es decir que realmente nos haya tocado una de esas muestras del 5 % cuyos árboles tienen un crecimiento excesivamente rápido o anormalmente lento. Pero nuestra hipótesis en principio hay que rechazarla porque los datos experimentales no la apoyan.

1.1.2. Procedimiento

El ejemplo anterior es un contraste de hipótesis *bilateral* para la media poblacional μ. La hipótesis nula (siempre debe contener el signo "=") es

$$H_0 \colon \mu = \mu_0;$$

la hipótesis alternativa es

$$H_1 \colon \mu \neq \mu_0.$$

El estadístico es

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \tag{1.1}$$

y sigue una distribución t con $n - 1$ grados de libertad. Se fija un nivel de significación α y la región de aceptación está comprendida entre los valores $\pm t_{\alpha/2}$.

En todos los contrastes los pasos a dar son:

- Establecer una hipótesis científica que nos interesa contrastar.
- Disponer los cálculos: población, variable, parámetro, hipótesis nula y alternativa, estadístico, tipo de distribución, nivel de confianza.
- Calcular con las tablas los valores donde empieza la región crítica, es decir aquel valor a partir del los cual consideramos que el resultado es muy improbable y va contra la hipótesis nula.
- Cálculo del valor muestral, aceptación o rechazo de H_0.
- Conclusiones para nuestra hipótesis inicial, interpretación.

Ejemplo (Adaptado de Samuels et al. 6.3.16.) Creemos que el nivel medio de hemoglobina en pacientes que están en la etapa final de una enfermedad renal tratados con el medicamento "Epoetina alfa" es de 10 g/dl. En un experimento con 101 pacientes obtenemos una media de 10,3 g/dl con una $s = 0,9$. ¿Es aceptable nuestra hipótesis con un nivel de confianza del 95 %?

Ponemos $H_0 : \mu = 10$, $H_1 : \mu \neq 10$, y usamos el estadístico

$$\frac{\bar{X} - 10}{s/\sqrt{101}}$$

que sigue una distribución t_{100}. Para un nivel de significación $\alpha = 5\,\%$ es $t_{0,025} = 1,984$. El valor del estadístico en el experimento es

$$t = \frac{10,3 - 10}{0,9/\sqrt{101}} = 3,35,$$

por tanto rechazamos H_0.

Interpretación: si H_0 fuese cierta, la probabilidad de obtener ese resultado experimental es menor que el 5 %, por tanto consideramos que nuestra hipótesis no tiene suficiente evidencia.

1.2. Contraste de hipótesis para una proporción

Llamamos p a la proporción en la población y $\hat{p}$ en la muestra.

Para un contraste sobre proporciones estableceremos una hipótesis nula (por ejemplo $H_0 : p = p_0$) y usaremos el estadístico

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}. \tag{1.2}$$

Para n grande sigue una distribución normal.

Ejemplo 8.4.7 Creemos que el 85 % de los niños con dolor torácico presentan un electrocardiograma normal. En una muestra de 139 niños esa proporción fue de 123/139, es decir de más del 88 %. ¿Podemos rechazar la hipótesis $p = 0'85$ con un nivel de significación del 10 %?

Sol.: Ponemos $H_0\colon p = 0'85$ y $H_1\colon p \neq 0'85$. Como el contraste es bilateral, buscamos el punto $z_{\alpha/2} = z_{0'05} = 1'645$. En la muestra el estadístico vale

$$\frac{0'885 - 0'85}{\sqrt{0'85 \times 0'15/139}} = 1'15.$$

Como está comprendido entre $-1'645$ y $+1'645$ aceptamos H_0.

Con el software R haríamos `qnorm(0.05)` para obtener el valor `-1.644854`.

1.3. Contraste de hipótesis para la varianza y la desviación típica poblacionales

Empezamos con un contraste bilateral. La hipótesis nula es

$$H_0\colon \sigma = \sigma_0$$

y la alternativa es

$$H_1\colon \sigma \neq \sigma_0.$$

El estadístico, que ya conocemos, es

$$\frac{(n-1)s^2}{\sigma_0^2} \tag{1.3}$$

que sigue una distribución χ^2 con $n-1$ grados de libertad.

Si H_0 es cierta, el estadístico oscilará alrededor del valor promedio (puede comprobarse que es $n-1$). La región de aceptación correspondiente al nivel de confianza $1-\alpha$ estará limitada por los dos valores $a = \chi^2_{1-\alpha/2}$ y $b = \chi^2_{\alpha/2}$.

Ejemplo (adaptado de Milton, 7.2.1) Un criador de perros trata de acercar mediante crianza selectiva la varianza de la altura de los animales al valor $0'25$. Contrastar la hipótesis $\sigma^2 = 0'25$ con un nivel de significación del 5 %, si en una muestra aleatoria de 15 perros se obtuvo una cuasi-varianza $s^2 = 0'21$.

SOL.: Usamos el estadístico $14s^2/0'25 = \chi^2_{14}$. Es un contraste bilateral, por lo que en las tablas obtenemos los valores

$$b = \chi^2_{0'025} = 26'12, \quad a = \chi^2_{0,975} = 5'63.$$

En el experimento obtenemos $\chi^2 = 11'76$, por lo que aceptamos H_0.

Interpretación: aunque en el experimento hemos obtenido un valor inferior a la hipótesis, la diferencia no es suficiente para rechazarla.

Cuando el contraste sea unilateral (ver la siguiente subsección 1.3.1) habrá que reformular H_0 y H_1 y buscar los valores χ^2_α ó $\chi^2_{1-\alpha}$, según el caso.

1.3.1. Contrastes unilaterales

En ocasiones la región de rechazo o *región crítica* ocupará las dos colas (contraste *bilateral*), pero en otras estará toda concentrada en el lado derecho o izquierdo (contraste *unilateral*).

Ejemplo (6.5.4) La Consellería de Pesca sólo permite la extracción de almeja si el número medio de bacterias por cm^3 en el agua no pasa de 70. Se hizo un muestreo en 9 lugares de una ría y se obtuvo un recuento de $\overline{X} = 71'7$, con $s = 2'3$. ¿Que decisión deben tomar los inspectores (con un nivel de confianza del 99 %)?

Sol.: Aunque el valor observado es ligeramente más grande que el permitido, esto puede ser una peculiaridad de las zonas de donde se han tomado las muestras de agua y no ser representativo de toda la ría. Se trata de ver si la diferencia es excesiva o no. Como hipótesis nula ponemos

$$H_0 : \mu \leq 70$$

(la que contiene el signo '=") y como alternativa la contraria: $H_1 : \mu > 70$. El estadístico es el dado en (1.1), que sigue una t con 8 grados de libertad.

Para decidir qué tipo de contraste debemos tener en cuenta (unilateral, bilateral, por la derecha, por la izquierda.) hay que mirar qué resultados del experimento nos llevarían a rechazar H_0, es decir, qué valores del estadístico van completamente en contra de la hipótesis nula y no son solamente pequeñas variaciones.

En el ejemplo razonamos de la siguiente forma. Un resultado muestral pequeño (por ejemplo $\overline{X} = 65$) sería un indicio favorable a H_0 mientras que un valor muestral muy alto (por ejemplo $\overline{X} = 80$) sería un indicio contrario a H_0. Nótese que en el primer caso el estadístico daría un valor negativo, en el segundo daría un valor muy positivo. Los valores ligeramante positivos pueden ser debidos a peculiaridades de la muestra, y no ser significativos para la población. Por tanto en este ejemplo *son los valores muy positivos del estadístico los que favorecen la hipótesis alternativa.* Así que la región crítica se concentra en la parte derecha. Se trata de un *contraste unilateral derecho.*

Podemos razonar de otra manera. Si la hipótesis nula es que $\mu \leq 70$, es decir $\mu - 70 \leq 0$, y realizamos un experimento, como la media muestral $\overline{X}$ es un estimador de μ, es de esperar qu $\overline{X} - 70$ sea negativo, o como mucho algo positivo, pero no demasiado positivo. Por tanto la región crítica se concentra en la parte derecha.

El siguiente paso es determinar con precisión la región crítica, es decir, la región de rechazo de H_0. Para ello nos dicen que el *nivel de significación* es $\alpha = 0,01$. Es decir consideramos improbable un resultado que sólo pueda darse en uno de cada 100 experimentos.

La región crítica corresponde al valor de las tablas que deje fuera un área igual al nivel de significación.

En las tablas de la distribución t_8 vemos que el valor que deja a su derecha un área de $0'01$ es $2'896$, es decir $p(t \geq \mathbf{2'896}) = 0'01$. Esto significa que un resultado más alto de $2'896$ es improbable (probablidad inferior a $0,01$).

Al calcular el estadístico en nuestra muestra obtenemos

$$\frac{71,7 - 70}{2'3/\sqrt{9}} = 2'22$$

que no supera el borde de la región crítica, porque $2'09 < 2'896$. Por tanto aceptamos H_0, es decir la diferencia entre el valor muestral y el poblacional no es significativa. Como aceptamos H_0, las almejas son aptas para el consumo.

Con el software *R* para calcular el valor de las tablas escribiríamos `qt(0.99,8)` y nos devolvería el valor `2.896459`.

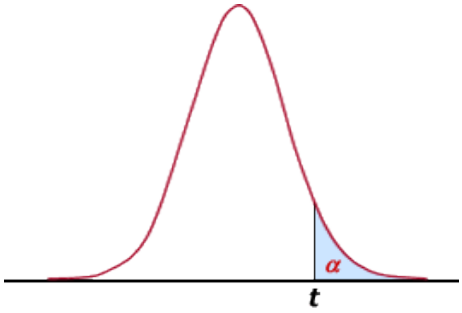


Figura 1.2: tomado de http://www.chemiasoft.com/

Nota: Si hubiésemos puesto la hipótesis nula $H_0\colon \mu \geq 70$ y la alternativa $H_1 : \mu < 70$ el contraste sería unilateral izquierdo y aceptaríamos H_0, con lo que prohibiríamos el consumo. El matiz es el siguiente, y depende de lo que diga la normativa: en un caso se permite el consumo si los valores del sondeo no son muy superiores a 70. En el segundo se permite el consumo sólo si son claramente inferiores a 70.

Ejemplo (adaptado de Samuels et al. 9.3.1) Contraste unilateral para una proporción. En una población de niños con edades comprendidas entre los 3 y 5 años creemos que al menos el 4 % tienen deficiencia de hierro. Contrastar esta hipótesis con un nivel de confianza del 99 % sabiendo que en una muestra de 848 niños hemos encontrado deficiencia de hierro en 32 de ellos.

SOL: $H_0\colon p \geq 0'04$, $H_1\colon p < 0'04$. Estadístico

$$\frac{\hat{p} - 0'04}{\sqrt{0'04 \times 0'96/848}} = z.$$

Para $\alpha = 1\,\%$ calculamos $z_\alpha = 2'33$. Es un contraste unilateral *izquierdo* ya que los valores muestrales muy inferiores a $p_0 = 0,04$ nos llevarán a descartar H_0. En el experimento es $\hat{p} = 32/848 = 3,77\,\%$, luego $z = -0'34$. Como este valor está por encima de $-2'33$ aceptamos H_0.

Interpretación: aunque el valor obtenido es inferior a 4 %, la diferencia no es significativa, por lo que podemos aceptar que en la población la proporción de niños con deficiencia de hierro es por lo menos del 4 %.

Ejemplo 7.2.2, pág. 256 El nivel de calcio en la sangre de los mamíferos no debe sobrepasar una desviación típica σ de 1 ml/100 mg. En otro caso habría problemas de coagulación. Las nueve pruebas realizadas en un laboratorio animal arrojaron un valor $s = 2$. ¿Podemos afirmar que en este laboratorio la desviación típica σ es más alta de lo deseable? (Usar un nivel de significación del 5 %).

Sol.: Vamos a contrastar la hipótesis nula $H_0\colon \sigma \leq 1$ frente la alternativa será $H_1\colon \sigma > 1$. El estadístico de contraste, de acuerdo con (1.3), es

$$\frac{8s^2}{1^2}.$$

Los valores altos de s favorecen la hipótesis alternativa, y eso corresponde a valores altos del estadístico. Por tanto el contraste es *unilateral derecho*.

En las tablas de χ^2_8 obtenemos que el punto que deja a su derecha un área del 5 % es 15′5, es decir $p(\chi^2 \geq 15'5) = 0'05$. En la muestra tenemos $s = 2$, por tanto el estadístico vale 32. Está en la región crítica y en consecuencia *rechazamos* H_0.

Otra manera de hacerlo sería calcular directamente el área a la derecha del valor muestral, es decir $p = p(\chi^2 \geq 32) = 0'00009$. Como $p < \alpha = 0'05$ deducimos que la muestra cae dentro de la región crítica, y por tanto rechazamos H_0.

La interpretación es que, si H_0 fuese cierta, el valor muestral es demasiado alto como para estar entre los resultados que serían "probables" (el 95 % de todos). En otras palabras, con mucha seguridad, σ en este laboratorio supera el valor máximo de $\sigma = 1$.

Con R el borde de la región crítica se buscaría como `qchisq(0.95,8)` y daría 15′50731. El área a la derecha de 16 se calcula como `1-pchisq(16,8)` y vale $9'314161e-05$.

1.3.2. Valor p

En el ejemplo de la pesca de la página 38 podíamos haber procedido de otra manera, sin calcular el borde de la región crítica. Primero se calcula el estadístico en la muestra, 2′22, y después el área p a su derecha, que según las tablas está entre 0′05 y 0′025, con lo que $p > \alpha = 0'01$. Al ser p más grande que el área de la región crítica se deduce que 2′22 no está en la región crítica, con lo que aceptamos H_0.

Con el software *R* el área a la derecha de 2′09 se calcularía como `1-pt(2.22,8)` y devuelve el valor `0.02859105`.

El valor p es el área que corresponde a los valores que están más allá que los del experimento. Representa la probabilidad de que en una muestra aparezca un resultado como el de experimento o aún más alejado. Por tanto si p es muy pequeño tendremos que rechazar H_0 pues significa que hemos obtenido un resultado improbable.

Cuando *no* nos den un nivel de significación α podemos proceder directamente de la siguiente manera:

- Establecer la hipótesis científica que nos interesa contrastar.
- Disponer los cálculos: población, variable, parámetro, hipótesis nula y alternativa, estadístico, tipo de distribución.
- Cálculo del valor muestral en el experimento y el área p que corresponde a ese valor
- Si p es pequeña, rechazamos H_0, porque significa que el valor obtenido en el experimento es improbable. Si p es grande, significa que el resultado entra dentro de lo que hemos considerado probable.
- Conclusiones para nuestra hipótesis inicial, interpretación.

En un contraste, evaluamos la evidencia en contra de H_0. La distribución del estadístico nos permite saber cuánto puede esperarse que el resultado del experimento se desvíe de H_0 por culpa del azar al escoger la muestra. El valor p nos indica si el resultado es compatible con H_0 (si el valor p es alto, no nos hemos desviado demasiado) o el resultado es improbable

(valor p pequeño, nos hemos desviado mucho). En resumen, el valor p es la probabilidad de obtener, si H_0 fuese cierta, un resultado tan desviado o más que el que hemos obtenido.

El interés del valor p es que no hace falta elegir un nivel de significación concreto. Esta elección depende del contexto (si la posibilidad de cometer un error al rechazar una hipótesis puede tener consecuencias serias) y suele estar influída por las expectativas del investigador (si no está muy convencido querrá una evidencia muy alta). En la práctica profesional, se publica el valor p del experimento, de modo que cada científico pueda evaluar el resultado en función de su propio nivel de exigencia.

Ejemplo. Se realiza un estudio para estimar la proporción de mujeres que han padecido en el último año una enfermedad de transmisión sexual. En una muestra aleatoria de 200 mujeres, 30 declararon haber padecido una enfermedad de ese tipo.

¿Puede aceptarse que la proporción en la población no sobrepasa el 10 %?

SOL: Es un contraste para una proporción, $H_0 : p \leq 0'10$. No se da ningún nivel de significación, por lo que hay que calcular el valor p.

Los datos del problema son $n = 200$, $\hat{p} = 30/200 = 0'15$. El estadístico vale

$$\frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0'15 - 0'10}{\sqrt{0'10 \times 0'90/200}} = 2'36.$$

Como es un contraste unilateral derecho ($H_1 : p > 0'10$), hay que ver si el valor muestral se ha ido demasiado a la derecha. Para eso calculamos, en la curva normal, el área a la derecha de $z = 2'36$. Vemos que vale $p = 0'00914$, que es un valor p muy pequeño (por ejemplo más pequeño que el 1 %). En consecuencia rechazamos H_0.

La interpretación es que el resultado del experimento (15 %) es demasiado alto como para considerarlo compatible con la hipótesis de que en la población la proporción no pasa del 10 %.

En general, lo que le interesa al investigador es la hipótesis *alternativa*: este medicamento es mejor, estas dos poblaciones son diferentes, este tratamiento es efectivo. La hipótesis nula representa lo contrario: no hay mejoría, no hay diferencias, no tiene efecto. Cuando rechazamos H_0 es porque observamos un efecto real, es decir, no debido únicamente a una fluctuación aleatoria. Decimos que hay una evidencia *estadísticamente significativa*.

Diferencia significativa En muchas ocasiones cuado se acepta H_0 suele decirse: hubo diferencias, pero no fueron *significativas*. Esto quiere decir que no encontramos evidencia de que la variación no se deba más que al azar.

Si rechazamos H_0, es porque el resultado del experimento es tan improbable que no se debe sólo a la fluctación achacable al muestreo. En este caso el estadístico tiene un valor muy alto (medido con el valor p o porque cae en la región de rechazo), pero esto no significa que sea de mucha *magnitud* (ya que medimos en desviaciones típicas, no en unidades físicas.

Ejemplo En un estudio sobre la diferencia de peso entre hombres y mujeres, se obtuvieron valores medios de 87 y 71 Kg. El valor del estadístico t fue de 0,93, que no es muy alto y corresponde a un área a la derecha del 45 %, por lo que se acepta H_0. Pero que la diferencia no sea *significativa* no quiere decir que sea grande o pequeña en tamaño (en este caso son 16 Kg.).

1.4. Tipos de errores

(págs. 224 y 239) El propósito de un experimento es seleccionar una muestra y decidir si los datos tienden a apoyar o a refutar nuestra hipótesis de investigación.

En un contraste de hipótesis la decisión se toma proponiendo una *hipótesis nula*, observando el valor de algún *estadístico de contraste* cuya distribución de probabilidad conocemos, y descartando la hipótesis si el valor obtenido en el experimento es improbable.

Algunos autores comparan esto con lo que pasa en un juicio: mientras no se demuestre lo contrario, el acusado es inocente. El juez sólo lo considerará culpable si se acumulan suficientes evidencias para rechazar la presunción de inocencia. Rechazaremos la hipótesis nula sólo si hay suficientes evidencias en contra.

Pero no podemos estudiar la población entera, por lo que comprobamos la hipótesis sólo en una muestra. La consecuencia es que nunca podemos probar o refutar la hipótesis con certeza absoluta. De este modo siempre es posible cometer un error, independientemente de la decisión que adoptemos

	H_0 **cierta**	H_0 **falsa**
Aceptamos H_0	OK $(1-\alpha)$	Error tipo II (β)
Rechazamos H_0	Error tipo I (α)	OK Potencia $(1-\beta)$

Cuadro 1.1: Posibles errores en el contraste de hipótesis

Error de tipo I: A veces, debido al azar, la muestra no es representativa de la población. Cuando rechazamos H_0 porque el estadístico cae en la región crítica es posible que, después de todo, H_0 fuese cierta, pero que hayamos tenido la mala suerte de trabajar con una muestra extrema.

Los resultados de la muestra no reflejan la realidad de la población, y nos llevan a una inferencia errónea. Decimos que hemos cometido un *error de tipo I*.

Se llama también un *falso positivo*, porque nos parece que hay una diferencia donde no la hay. Esto ocurrirá en un determinado tanto por ciento de las muestras (igual al nivel de significación α). De todos modos, cuando H_0 sea cierta, la mayor parte de las veces la aceptaremos (nivel de confianza $1-\alpha$).

Ejemplo. Estamos comparando la eficacia de dos medicamentos. El medicamento A es muy barato, pero el B es extremadamente caro. Si la hipótesis nula es "los dos medicamentos son igual de eficaces", un error de tipo I puede costarle mucho dinero a los pacientes (usan un medicamento caro sin motivo). Esto se evitará poniendo un nivel de significación pequeño.

Error de tipo II: Por otra parte cabe la posibilidad de que aceptemos una H_0 falsa. En este caso cometeremos un *error de tipo II*.

Se llama también un *falso negativo* porque no apreciamos una diferencia que realmente existe. La probabilidad de que esto ocurra se llama β y es más difícil de calcular.

En la Figura 1.3 está dibujada una posible situación: a la izquierda la distribución que estamos suponiendo en una H_0 falsa y a la derecha la distribución que ocurre en realidad; la probabilidad β está marcada en amarillo (probabilidad de caer en la región de aceptación, pero medida en la auténtica distribución, no en la que suponemos nosotros).

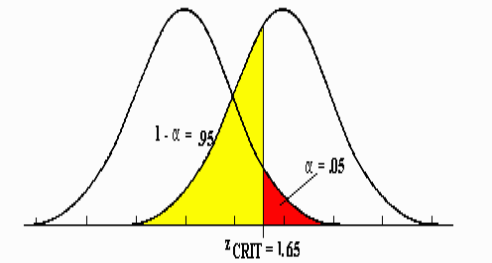


Figura 1.3: Tomado de http://www.psychstat.missouristate.edu/

Ejemplo. A veces las consecuencias de un error de tipo I pueden no ser muy serias, pero sí las de un error de tipo II. Por ejemplo, si dos medicamentos son igual de efectivos y además cuestan lo mismo, puede interesarnos contrastar la hipótesis nula "tienen los mismos efectos secundarios". Ahora bien, supongamos que se sabe que el medicamento A ha sido usado durante décadas sin que haya habido noticias de efectos secundarios; en cambio, hay alguna sospecha de que el medicamento B puede causar graves efectos secundarios en algunos pacientes. En este caso,

– un error de tipo I (rechazar H_0 siendo cierta, creer que hay efectos secundarios cuando no los hay) no tendrá grandes consecuencias en los consumidores,

– pero un error de tipo II (confundirse y aceptar H_0 siendo falsa: decir que no hay efectos cuando sí los hay) puede tener graves consecuencias desde el punto de vista de la salud pública.

El valor de β depende de una serie de factores (tamaño de la muestra, nivel de significación, etc.). Como puede verse en la figura, hacer decrecer α puede hacer que aumente β. También se ve que el error de tipo II es importante cuando el valor real está cerca del valor que estamos contrastando, o si el intervalo de aceptación es demasiado amplio. Al diseñar un experimento deberemos analizar los costes relativos de cada tipo de error. Si el error β puede tener consecuencias, lo correcto sería aumentar el nivel de significación o aumentar el tamaño de la muestra.

La frecuencia $1-\beta$ con que se rechaza una H_0 falsa se llama *potencia* del contraste.

Como la potencia es la probabilidad de no cometer un error de tipo II, es deseable que sea alta. La potencia de un contraste depende de muchos factores. En general aumenta si se toman muestras más grandes.

Ejemplo (adaptado de Samuels et al. 7.9.1) Se realiza una prueba médica para detectar una enfermedad que sufre el 1 % de la población. Imaginemos que H_0 es: *la persona está sana*. Realizamos la prueba a 10.000 personas y obtenemos estos resultados:

	Tiene la enfermedad	No tiene la enfermedad	
Test positivo	800	4950	
Test negativo	200	94.050	
	1.000	99.000	

Como vemos, la probabilidad de un falso positivo (es decir, que la prueba dé positiva en una persona sana) es $4950/99000 = 0'05$, por tanto el nivel de significación es $\alpha = 5\,\%$. Por otro lado, la proporción de falsos negativos es $\beta = 200/1000 = 0'20$, por lo que la potencia de este contraste sería del 80 %.

En otras palabras, esta prueba tiene un 80 % de probabilidad de detectar la enfermedad si la persona la tiene, y un 95 % de probabilidad de indicar la ausencia de enfermedad en una persona que efectivamente está sana.

Curiosamente, la proporción de resultados erróneos entre los positivos es altísima, $4950/5750 = 0'86$. Esto se debe a que la frecuencia de la enfermedad (prevalencia) es muy baja.

Por tanto, en un experimento bien diseñado procuraremos que las probabilidades de cometer un error de tipo I o de tipo II sean bajas. Pero eso no nos garantiza que la mayor parte de las veces acertemos al rechazar o aceptar H_0.

Ejemplo (adaptado de Samuels et al. 7.3.4) La quimoterapia es el tratamiento habitual para tratar cierto tipo de cáncer. Realizamos un estudio para comparar su eficacia con un nuevo tratamiento por inmunoterapia. Como hipótesis nula pondremos que la inmunoterapia no es efectiva. Estudiar las conecuencias de cometer un error de tipo I y de tipo II.

SOL.: Si la inmunoterapia no es efectiva (o sea si H_0 es cierta) pero concluimos erróneamente que lo es (falso positivo), cometemos un error de tipo I: llevará a un uso innecesario de una terapia inefectiva y potencialmente peligrosa. Un aspecto bueno es que la probabilidad de cometer un error de este tipo la controlamos nosotros, fijando el nivel de significación α.

Si la inmunoterapia es efectiva (o sea si H_0 no es cierta) pero no lo detectamos (es decir, aceptamos H_0), es un falso negativo o error de tipo II. Se continuará usando el tratamiento antiguo hasta que alguien proporcione evidencia convincente de la eficacia del nuevo. La probabilidad de cometer este tipo de error es menos controlable. Quizás debamos repetir el experimento con una muestra más grande, si sigue interesándonos el problema.

Ejemplo Contraste de hipótesis para la media poblacional μ. En rojo están representados los posibles valores de β (probabilidad de un error de tipo II). Cuanto más cerca se encuentre el verdadero valor del que hemos supuesto en la hiptesis nula, mayor será la probabilidad un error tipo II. Pero como el verdadero valor de μ es desconocido, en la mayor parte de los casos no se puede calcular el error de tipo II. Para estimarlo pueden usarse las llamadas *curvas ROC Receiver Operating Characteristic*.

En azul está marcada α. la probabilidad del error de tipo I.

Usualmente, se diseñan los contrastes de tal manera que la probabilidad α sea del 5 % ($0'05$), aunque a veces se usan el 10 % ($0'1$) o el 1 % ($0'01$). El recurso para aumentar la potencia del contraste, esto es, para disminuir β, es aumentar α, o aumentar el tamaño de la muestra, lo que en la práctica conlleva que el estudio sea más caro.

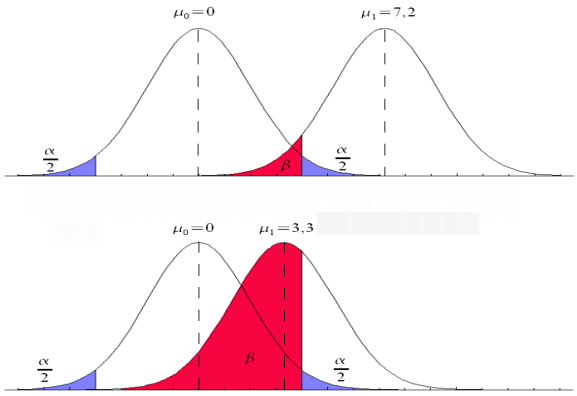


Figura 1.4: tomado de https://commons.wikimedia.org/wiki/File:Beta-Fehler.png

Capítulo 2

Expositivas 11–15

Queremos comparar las medias de dos poblaciones. Tomamos una muestra en cada una de las poblaciones. Usaremos métodos diferentes según que las muestras que tomamos estén emparejadas o sean independientes entre sí.

Las muestras son *independientes* si han sido elegidas independientemente una de otra, es decir hemos hecho un sorteo diferente en cada población, o hemos seleccionado cada muestra sin tener en cuenta la otra. Puede que a veces tengan el mismo tamaño y otras veces tamaños diferentes.

En cambio, las muestras son *emparejadas* o *relacionadas* si hay un emparejamiento natural entre cada elemento de una muestra y otro elemento de la otra muestra. Por ejemplo hermanos gemelos, marido y mujer, madre e hijo. Otras muchas veces esto ocurre cuando los datos se recogen dos veces de los mismos sujetos o se dan dos tratamientos a las mismas personas.

2.1. Contraste para dos medias, muestras independientes

Decimos que dos muestras son independientes cuando cada una ha sido obtenida aleatoriamente y sin ninguna relación de emparejamiento con la otra.

Cuando las muestras son independientes, la distribución del estadístico es aproximadamente t, Va a haber tres casos posibles:

En los dos primeros usamos estadísticos algo más sencillos:

- cuando conocemos las varianzas poblacionales σ_1^2, σ_2^2 (lo veremos en 2.1.1);
- cuando no las conocemos pero podemos suponer que son iguales (lo veremos en 2.1.2).

en el caso general, que veremos en el apartado 2.1.3), el número de grados de libertad puede ser difícil de calcular.

2.1.1. Conocidas las varianzas poblacionales

Si se conocen las varianzas σ_1^2 y σ_2^2 de las poblaciones, usaremos el estadístico

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

que tiene una distribución normal estándar z.

Ejemplo (adaptado de 9.1.3, pág. 291) Tenemos dos poblaciones (corredoras y nadadoras olímpicas), con varianzas $\sigma_1^2 = 16$ y $\sigma_2^2 = 18$. En la primera se selecciona una muestra de tamaño 20 y se obtiene una media de $\overline{X}_1 = 15'5$. En la segunda se tiene $n_2 = 25$ y $\overline{X}_2 = 9'8$. Queremos contrastar la hipótesis $\mu_1 = \mu_2 + 5$.

Ponemos

$$H_0 \colon \mu_1 - \mu_2 = 5, \quad H_1 \colon \mu_1 - \mu_2 \neq 5.$$

Para un nivel de confianza del 95 % es un contraste bilateral, cuya región crítica está limitada por $z = \pm 1'96$. En nuestra muestra el estadístico vale

$$\frac{(15'5 - 9'7) - (5)}{\sqrt{16/20 + 18/25}} = 0'8/1'23 = 0'65$$

que está en la región de aceptación de la hipótesis nula.

Dato: `DISTR.NORM.ESTAND.INV(0,025)=-1,959963985`.

2.1.2. Desconocidas las varianzas poblacionales, pero supuestas iguales

(pág. 302) A veces no conocemos la varianzas de las poblaciones pero podemos asumir que son iguales. Esta suposición se comprueba con un contraste F para dos varianzas, que veremos más adelante (en el apartado 2.5). Otras veces se decide usando la regla práctica que aparece más abajo.

En este caso primero se calcula la *cuasi-varianza muestral conjunta*

$$s_p^2 = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}$$

que es la media ponderada de las cuasi-varianzas de las muestras (nótese que el denominador es la suma de $n_1 - 1$ y $n_2 - 1$).

Entonces el estadístico

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{s_p \cdot \sqrt{(1/n_1 + 1/n_2)}} \tag{2.1}$$

tiene una distribución t con $n_1 + n_2 - 2$ grados de libertad.

Como regla práctica (pág. 295), cuando no dispongamos de las tablas de F, aceptaremos que $\sigma_1^2 = \sigma_2^2$ siempre que se cumpla

$$\frac{1}{2} \leq \frac{\hat{s}_1^2}{\hat{s}_2^2} \leq 2,$$

es decir, ninguna cuasi-varianza muestral es más del doble de la otra.

Ejemplo (9.3.3, pág. 304) Medimos la distancia que recorren volando los murciélagos en busca de alimento. Sospechamos que las hembras vuelan más que los machos.

En la muestra, los datos de las hembras son $n_1 = 25$, $\overline{X}_1 = 205$ metros, $\hat{s}_1 = 100$; los de los machos $n_2 = 11$, $\overline{X}_2 = 135$, $\hat{s}_2 = 95$. Creemos que $\mu_1 > \mu_2$ en la población.

Sol.: Las hipótesis son $H_0\colon \mu_1 \leq \mu_2$, $H_1\colon \mu_1 > \mu_2$. Con el estadístico (2.1) el contraste es unilateral derecho, es decir los valores muy positivos del estimador corresponden a H_1. Para un nivel de significación $\alpha = 10\,\%$ la región crítica comienza en el valor $1'30$ (usamos la distribución t con 34 grados de libertad).

La varianza conjunta conjunta es

$$s_p^2 = \frac{24 \times (100)^2 + 10 \times (95)^2}{24 + 10} = 9713'24$$

y el valor del estadístico en la muestra es $70/35'66 = 1'96$ (cuidado, en la fórmula del estadístico aparece s_p, no su cuadrado).

Por tanto rechazamos H_0 ya que ese valor está en la región crítica. Interpretación: el experimento apoya la hipótesis de que la distancia recorrida por término medio por las hembras es mayor que la de los machos.

Dato: $\texttt{DISTR.T.INV}(0,2;34) = 1,306951587$ (en algunas versiones Excel devuelve para la distribución t el valor para dos colas; si queremos un contraste unilateral debemos darle el doble de probabilidad).

Intervalo de confianza Aunque muchas veces no lo diremos explícitamente, siempre es posible calcular un intervalo de confianza una vez que nos dan un estadístico como por ejemplo (2.1). Debemos, salvo casos especiales, ser capaces de deducirlo, ya que el error de estimación, es decir, la distancia entre la media poblacional (o la diferencia de medias) y la media muestral es igual a t veces la desviación típica del estimador.

Por tanto

$$(\mu_1 - \mu_2) = (\overline{X}_1 - \overline{X}_2) \pm \text{error}$$

donde

$$\text{error} = t_{\alpha/2} \cdot s_p \cdot \sqrt{(1/n_1 + 1/n_2)}.$$

Aunque en un problema hayamos hecho un contraste unilateral usando t_α, si después hacemos un intervalo debemos usar el valor $t_{\alpha/2}$ que corresponde a un contraste *bilateral* , ya que el intervalo es del tipo $\pm$.

2.1.3. Caso general

Falta ver el caso más habitual, en que ni conocemos las varianzas poblacionales, ni podemos suponer que sean iguales.

El estadístico que usaremos es

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}}$$

donde $(\mu_1 - \mu_2)_0$ es el valor que ponemos en la hipótesis nula. Este estadístico tiene una distribución t de «Student», pero su número de grados de libertad γ es complicado de

calcular; usaremos la siguiente fórmula (pág. 309):

$$\gamma \leq \frac{\left[\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2\right]^2}{\frac{[\hat{s}_1^2/n_1]^2}{n_1-1} + \frac{[\hat{s}_2^2/n_2]^2}{n_2-1}}.$$

Esta prueba fue inventada por B.L. Welch en 1947.

Ejemplo (Ejercicio 9.4.2, pág. 311) Un isótopo radioactivo (`Sr-90`) se acumula en los huesos a través de la leche de vaca consumida. Se quiere conocer si el nivel de isótopo en los niños es diferente que en los adultos. Para ello se toman:

- Una muestra aleatoria de $n_1 = 121$ niños. En ellos se obtiene una concentración media de $\overline{X}_1 = 2'6$ picocurios/gramo, con una cuasi-desviación típica de $\hat{s}_1 = 1'2$;
- otra muestra aleatoria de $n_2 = 61$ adultos. Para éstos se tiene $\overline{X}_2 = 0'4$ y $\hat{s}_2 = 0'11$.

En este caso está claro que las dos muestras son completamente independientes, no hay relación entre los adultos y los niños elegidos. Planteamos las hipótesis

$$H_0\colon \mu_1 = \mu_2, \quad H_1\colon \mu_1 \neq \mu_2,$$

donde llamamos μ_1 a la concentración media en la población de niños y μ_2 en la población de adultos. Podemos escribir las hipótesis de esta otra forma:

$$H_0\colon \mu_1 - \mu_2 = 0, \quad H_1\colon \mu_1 - \mu_2 \neq 0.$$

Cálculos del ejemplo En el ejemplo el número de grados de libertad es

$$\gamma \leq \frac{\left[(1'2)^2/121 + (0'11)^2/61\right]^2}{\frac{[(1'2)^2/121]^2}{120} + \frac{[(0'11)^2/61]^2}{60}} = \frac{0'00014639}{1'1809 \times 10^{-6}} = 123'965$$

por lo que tomaremos t con 123 grados de libertad, que es prácticamente igual a la normal estándar z (última línea de las tablas de la t).

Para un nivel de confianza del 95 % será $\alpha/2 = 0'025$ y el valor correspondiente en la tabla es $t = \pm 1'96$.

Por otro lado el valor del estadístico es

$$\frac{(2'6 - 0'4) - (0)}{\sqrt{(1'2)^2/121 + (0'11)^2/61}} = \frac{2'2}{0'109996} = 20$$

que supera con mucho el valor tabulado $t_\alpha = 1'96$, por lo que rechazamos H_0, es decir concluimos que la media de los niños es diferente de la de los adultos.

Contrastes unilaterales Si hubiésemos hecho un contraste unilateral, por ejemplo para comprobar que la media de los niños es *mayor* que la de los adultos ($\mu_1 > \mu_2$), pondríamos

$$H_0\colon \mu_1 \leq \mu_2, \quad H_1\colon \mu_1 > \mu_2,$$

o escrito de otro modo

$$H_0\colon \mu_1 - \mu_2 \leq 0, \quad H_1\colon \mu_1 - \mu_2 > 0.$$

Como en la fórmula del estadístico aparece $\overline{X}_1 - \overline{X}_2$, los valores positivos del estadístico son favorables a H_1, por tanto se trata de un contraste unilateral derecho. En las tablas encontramos $t_{0'95} = 1'645$ con lo que rechazamos H_0, es decir concluímos que la media de los niños es significativamente mayor.

Nota El «curio» (abreviación Ci) es una antigua unidad de radiactividad que representa la cantidad de material en la que se desintegran $3'7 \times 10^{10}$ átomos por segundo, más o menos la actividad de 1 gramo de `Ra-226`. Ha sido sustituída por el becquerel (abreviación Bq), que equivale a una desintegración nuclear por segundo. Por tanto $1\,\text{Ci} = 3'7 \times 10^{10}\,\text{Bq}$.

El prefijo «pico» significa 10^{-12}.

Ejemplo 7.2.1,2, 3,4 (adaptado de Samuels et al., Fundamentos de Estadística para las Ciencias de la Vida)

El abuso de sustancias como el pegamento puede producir trastornos neurológicos. Se midieron las concentraciones de un derivado del tolueno en un grupo de 6 ratas y se compararon con los de 5 ratas de un grupo control.

La concentración media en el primer grupo ($\overline{X}_1 = 540,8$ ng/g,) fue sustancialmente mayor que la del segundo grupo ($\overline{X}_2 = 444,2$ ng/g). Se tiene $s_1 = 66,1$ y $s_2 = 69,6$.

Puede inferirse que la diferencia observada indica un fenómeno bilógico real, o es sólo debida al azar?

2.2. Contraste de hipótesis para dos medias, muestras relacionadas

Decimos que dos muestras que están «relacionadas», o «emparejadas», cuando cada observación de una muestra está emparejada o relacionada con una observación de la otra muestra.

Por ejemplo, se trata de

- dos análisis realizados a la misma persona en dos momentos distintos (medidas repetidas)
- las puntuaciones de dos hermanos gemelos
- los datos de unas madres y sus hijas.

Se reconocen porque "se hace únicamente un sorteo" y la segunda *muestra* se deduce de la primera. Sin embargo técnicamente hay dos *poblaciones* (antes/después, gemelo primero/gemelo segundo, madres/hijas, ...)

Estadístico En el caso de muestras emparejadas trabajaremos con las diferencias $D = X_1 - X_2$ de las puntuaciones, lo que reducirá el problema a una sola muestra.

Si el contraste es bilateral, la hipótesis nula $H_0\colon \mu_1 = \mu_2$ se convierte en decidir si la media poblacional de las diferencias es cero: $H_0\colon \mu_D = 0$. La hipótesis alternativa es $H_1\colon \mu_D \neq 0$. También podemos hacer contrastes unilaterales.

Usaremos el estadístico

$$\frac{\overline{D} - (\mu_D)_0}{\hat{s}_D/\sqrt{n}}$$

que sigue una distribución t con $n-1$ grados de libertad.

Nota Cuando tenemos dos series de datos X_1, X_2, la media de las diferencias es lo mismo que la diferencia de las medias:

$$\overline{D} = \frac{\sum(X_1 - X_2)}{n} = \frac{\sum X_1 - \sum X_2}{n} = \overline{X}_1 - \overline{X}_2.$$

Sin embargo la varianza de las diferencias no es la diferencia de las varianzas, por eso s_D hay que calcularla o nos la tienen que dar.

Ejemplo. 9.5.1 pág. 314 Queremos investigar el efecto del ejercicio físico en el nivel de colesterol. Se toman muestras de sangre de 11 personas dos veces: una antes y otra después de un programa de ejercicios. Se obtienen los datos siguientes:

Nivel previo en mg/dl	Nivel posterior en mg/dl	D = Diferencia
182	198	-16
...	...	...
262	226	+36
232	210	+22

¿Es aceptable la hipótesis de que el ejercicio ha disminuido significativamente el nivel de colesterol?

Sol.: Tenemos $n = 11$. Calculamos la media de las diferencias y obtenemos $\overline{D} = 33'2$, mientras que la cuasi-desviación típica de las diferencias es $\hat{s}_D = 51'1$. La hipótesis a contrastar es $\mu_D > 0$. Por tanto ponemos $H_0\colon \mu_D \leq 0$ y $H_1\colon \mu_D > 0$.

Usaremos un nivel de confianza del 90 %. Es un contraste unilateral derecho con $\alpha = 0'1$. En la curva t con 10 g.l. obtenemos que $t_{0'10} = 1'372$. En la muestra el estadístico vale $2'15$. Como $2'15 > 1'812$ rechazamos H_0, es decir la diferencia media observada es suficientemente alta como para que sea aceptable suponer que el ejercicio físico disminuye significativamente el nivel de colesterol.

Nota Aunque las muestras estén emparejadas hay dos *poblaciones* diferentes: las personas que no hacen ejercicio físico y las que sí lo hacen.

Valor p En el ejemplo anterior podíamos haber razonado de otro modo. El nivel de significación es $\alpha = 0'1$ y se trata de un contraste unilateral derecho. En la curva t_{10} el área a la derecha del valor muestral $2'15$ vale $p = 0'0285$ que es menor que $\alpha = 0'10$. Por tanto el valor muestral está en la región crítica. Es decir, conociendo el área p correspondiente al experimento podemos ver si es "grande" o pequeña" y deducir si el resultado que hemos obtenido entra dentro de lo aceptable o es improbable.

Con Excel el área a la derecha se calcularía como `=DISTR.T(2,15;10;1)` y devuelve el valor `0,028532234`.

Intervalo de confianza En el ejemplo anterior, un intervalo de confianza para la diferencia de medias poblacionales es

$$\mu_D = \overline{D} \pm t_{\alpha/2}\,\frac{\hat{s}_D}{\sqrt{n}}.$$

Una forma de ver la relación entre contraste de hipótesis (bilaterales) e intervalos de confianza es la siguiente: si hubiésemos hecho los cálculos obtendríamos el intervalo

$$33'2 \pm 27'9$$

que tiene los límites $(5'3; 61'1)$. Este intervalo no contiene el valor $\mu_D = 0$, lo que significa que, a la vista del valor de la media de la muestra, no es razonable pensar que en la población la media sea 0.

2.3. Contrastes para dos proporciones

Los contrastes para dos proporciones son análogos a los de dos medias. Supondremos que las muestras son independientes y grandes ($n \geq 30$), para poder usar la aproximación de la binomial a la normal.

Siguiendo el libro de Milton (pág. 282) llamaremos *valor nulo* el que aparece en la hipótesis nula. Por ejemplo si es

$$H_0 \colon p_1 - p_2 \geq 0'3$$

entonces el valor nulo es $0'3$.

2.3.1. Valor nulo cero

Cuando el valor nulo es cero usaremos el siguiente método (pág. 283). Primero calculamos la *proporción conjunta*

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

que es la media ponderada de $\hat{p}_1$ y $\hat{p}_2$ (proporciones en las muestras). Es decir, es la proporción que se obtiene si juntamos las muestras.

Entonces el siguiente estadístico es aproximadamente normal z:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

Ejemplo Estudiamos la efectividad de un medicamento contra la insuficiencia renal. En 34 pacientes tratados con el medicamento sólo uno sufrió insuficiencia, es decir una proporción de $\hat{p}_1 = 1/34 = 0'03$; en cambio, en 38 pacientes a los que se dió un placebo hubo 10 que sufrieron insuficiencia renal, es decir $\hat{p}_2 = 10/38 = 0'26$.

Nuestra hipótesis es que, a la vista de los datos de esta muestra, en el conjunto de pacientes la proporción de casos de insuficiencia de los que no toman el medicamento (p_2) supera a la de los pacientes que lo toman (p_1).

Entonces ponemos

$$H_0 \colon p_1 \geq p_2, \quad H_1 \colon p_1 < p_2,$$

o lo que es lo mismo

$$H_0 \colon p_1 - p_2 \geq 0, \quad H_1 \colon p_1 - p_2 < 0.$$

Este "cero" en la hipótesis nula es lo que hemos llamado *valor nulo*.

Vemos que es un contraste unilateral izquierdo (valores muy negativos del estadístico nos llevan a rechazar la hipótesis nula H_0).

Calculamos la proporción conjunta

$$\hat{p} = \frac{34 \times 0'03 + 38 \times 0'26}{34 + 38} = \frac{1 + 10}{34 + 38} = 0'15.$$

El valor muestral del estadístico es

$$\frac{(0'03 - 0'26) - (0)}{\sqrt{(0'15)(1 - 0'15)(1/34 + 1/38)}} = \frac{-0'23}{0,084} = -2'73.$$

El valor p será el área a su izquierda

$$p = p(z \leq -2'73) = 0'003.$$

Cumple $p < \alpha = 0'01$ por lo que rechazamos H_0 con un nivel de significación del 1 %. Podemos afirmar que la proporción de casos de insuficiencia renal en los pacientes que toman la medicación es menor que la proporción en los que no la toman, con un nivel de confianza del 99 %.

2.3.2. Valor nulo distinto de cero

El "valor nulo" $(p_1 - p_2)_0$ puede ser distinto de cero, como en el siguiente ejemplo.

Ejemplo (adaptado de 8.5.2, pág. 276) En el mismo ejemplo del apartado 2.3.1, ahora queremos contrastar la hipótesis de que, a la vista de los datos de esta muestra, la proporción p_2 de casos de insuficiencia de los que no toman el medicamento supera *en más de un* 10 % a la de los pacientes que lo toman. Es decir creemos que $p_2 > p_1 + 0'1$.

En este caso se usa el estadístico

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2}}, \tag{2.2}$$

que tiene una distribución aproximadamente normal z (pág. 281).

Cálculos Para evitar después complicaciones con los signos es mejor escribir la pregunta como $p_1 - p_2 < -0'1$. Ponemos

$$H_0 \colon p_1 - p_2 \geq -0'1, \quad H_1 \colon p_1 - p_2 < -0'1.$$

Este valor $-0'1$ que va a aparecer en la diferencia $p_1 - p_2$ de la hipótesis nula se llama *valor nulo*. En este ejemplo, y tal y como hemos escrito las diferencias, es negativo. En otros será positivo. Si fuese cero tendríamos que usar el contraste del apartado 2.3.1.

Para decidir si el contraste es bilateral o no, razonamos de la siguiente forma: si H_0 fuese cierta, la diferencia $p_1 - p_2$ no podría ser muy negativa, por lo que el valor del estadístico (2.2) no debe ser muy negativo. Por tanto la región de rechazo de H_0 (es decir la región *crítica*) corresponde a los valores muy negativos. Es un contraste unilateral izquierdo.

El estadístico en la muestra vale

$$\frac{(0'03 - 0'26) - (-0'1)}{\sqrt{0'03 \times 0'97/34 + 0'26 \times 0'74/38}} = -0'13/0'077 = -0'17.$$

Se trata ahora de ver si es suficientemente negativo como para caer en la región crítica. Como no nos dan α, razonamos con el valor p, que es el área a la izquierda del experimento. En las tablas vemos que vale

$$p = p(z \leq -0'17) = 0,4325$$

que es muy grande; así que el resultado del experimento está poco alejado del cero, por lo que aceptamos H_0.

Interpretación: No podemos afirmar que la proporción de insuficiencias renales en los pacientes no tratados supere en más de un 10 % a la de los pacientes medicados (en la población).

p-valor Un p-valor pequeño nos indica que si H_0 fuese cierta, habríamos obtenido un resultado que sólo se daría en muy pocos experimentos. Esto es un fuerte indicio de que nuestra hipótesis nula no es correcta. En el caso anterior el p-valor es grande, por lo que es un resultado perfectamente compatible con H_0.

2.3.3. Tamaño de la muestra

Ya hemos visto, en algunos contrastes para una media y una proporción, cómo estimar el tamaño de muestra necesario. Veámoslo en el caso de dos proporciones (adaptado de 8.3 pág. 267 y del ejercicio 8.5.8).

Supongamos que queremos que el error de estimación de *dos proporciones* no sobrepase un valor máximo d. Entonces basta tomar

$$n \geq \frac{z^2}{2d^2}.$$

Justificación* Para dos proporciones, el error de estimación es $z \cdot s_{\text{estim}}$, donde

$$s_{\text{estim}} = \sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2}.$$

Como $n \leq n_1$ y $n \leq n_2$, la varianza del estimador es

$$s^2_{\text{estim}} \leq (1/4)(1/n + 1/n) = 1/2n,$$

porque sabemos que $\hat{p}\hat{q} \leq 1/4$.

Por tanto necesitamos que, en el peor de los casso posibes, sea

$$z \cdot \sqrt{\frac{1}{2n}} \leq d.$$

Así se obtiene un tamaño de muestra que es el doble que en el caso de una sola proporción.

Podeis consultar la página web
`http://stattrek.com/tutorials/ap-statistics-tutorial.aspx`,
en concreto la sección sobre "Statistical Inference"

2.4. La distribución F

Necesitamos una nueva familia de funciones de densidad, que usaremos para hacer contrastes sobre varianzas. Se llaman las curvas F y fueron descubiertas por el matemático G.W. Snedecor (1881– 1974).

Estas curvas aparecen en la siguiente situación: sea X una variable χ^2 (chi-cuadrado) con m grados de libertad y sea Y otra variable χ^2 con n grados de libertad. Si hacemos el cociente

$$\frac{X/m}{Y/n}$$

aparece una curva que tiene la siguiente fórmula

$$F(x) = c_{m,n}x^{m/2-1}(n + mx)^{-(m+n)/2}.$$

Se llama curva F con m grados de libertad en el numerador y n grados de libertad en el denominador.

Tiene un valor esperado de $n/(n-2)$.

*Las secciones marcadas con asterisco no son materia de examen

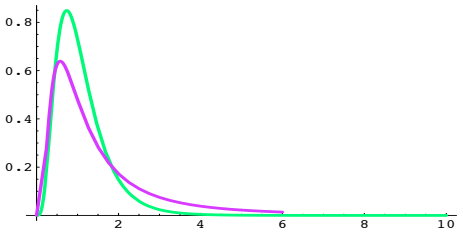


Figura 2.1: Dos curvas F con $(10, 23)$ y $(15, 4)$ grados de libertad

Nota Como hay tantas curvas F, sus tablas se organizan por páginas, una página para cada una de las área más usuales ($0'99$, $0'975$, etc.). Las áreas pequeñas pueden deducirse de las grandes gracias a la siguiente fórmula:

$$ {}_{\alpha}F_{m,n} = \frac{1}{{}_{1-\alpha}F_{n,m}}. $$

Ejemplo En una F con 8 grados de libertad en el numerador y 12 grados en el denominador, queremos calcular el valor que deja a su izquierda un área de $0'05$. Para ello buscamos el área $0,95$ en una F con $(12, 8)$ grados (nótese que los invertimos), obteniendo en la tabla el valor $3,28$. Entonces

$$ {}_{0'05}F_{8,12} = 1/(3'28) = 0'30. $$

Excel Por supuesto podemos calcular los valores anteriores con el ordenador (ojo, comprobar si da áreas a la derecha o la izquierda):
`DISTR.F.INV(0,95;8;12)=0,304512355`
`DISTR.F.INV(0,05;12;8)=3,283939006`.

2.5. Contraste para dos varianzas

En algunos contrastes sobre dos medias (apartado 2.1.2) suponíamos que las varianzas de las dos poblaciones eran iguales. Teníamos una regla práctica para comparar las varianzas σ_1^2 y σ_2^2 a partir de sus estimadores muestrales $\hat{s}_1^2$ y $\hat{s}_2^2$. En realidad debería hacerse un contraste de hipótesis

$$ H_0 \colon \sigma_1^2 = \sigma_2^2, \quad H_1 \colon \sigma_1^2 \neq \sigma_2^2, $$

usando como estadístico (pág. 297)

$$ \hat{s}_1^2/\hat{s}_2^2 $$

que sigue una curva F con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

Ejemplo En un problema sobre murciélagos (9.3.3), tenemos que las cuasivarianzas de las dos muestras (machos y hembras) son $\hat{s}_1^2 = (100)^2 = 10000$ y $\hat{s}_2^2 = (95)^2 = 9025$. El estadístico $\hat{s}_1^2/\hat{s}_2^2$ sigue una distribución F con $(24, 10)$ grados de libertad. El contraste es bilateral. Para $\alpha = 10\,\%$ tenemos los valores $F_{0,05} = 0,44$ y $F_{0,95} = 2,74$. Como el estadístico vale $\hat{s}_1^2/\hat{s}_2^2 = 10000/9025 = 1'11$ aceptamos H_0, es decir podemos suponer que

las varianzas poblacionales son iguales.

Datos: `DISTR.F.INV(0,05;24;10)=2,737247653`,
`DISTR.F.INV(0,95;24;10)=0,443510346`.

Ejemplo (9.2.2 y 9.2.4, pág. 297) Comparamos el nivel de un medicamento en personas jóvenes o mayores. Queremos contrastar la hipótesis $H_0\colon \sigma_1^2 = \sigma_2^2$ a partir de los siguientes datos:

- $n_1 = 41$, $\hat{s}_1 = 0,102$, $\quad n_2 = 29$, $\hat{s}_2 = 0,068$.

Como no nos dan el nivel de significación α, en este problema calculamos el valor p en vez de los límites de la región crítica. Para una distribución F con $(40, 28)$ grados de libertad obtenemos que el valor muestral es

$$\hat{s}_1^2/\hat{s}_2^2 = (0,102)^2/(0,068)^2 = 2,25.$$

Esta puntuación deja a su izquierda un área de $0,0136$. Como el contraste es bilateral, rechazamos H_0 con un nivel de significación del $10\,\%$ porque se cumple $p < \alpha/2 = 0,05$. Escribiríamos $p = 2 \times 0,0136 < \alpha$.

Dato: `DISTR.F(2,25;40;28)=0,013596365`.

Nota Los contrastes para dos desviaciones típicas serían exactamente iguales y se usaría el mismo estadístico $\hat{s}_1^2/\hat{s}_2^2$ (con cuadrados), pero la hipótesis se escribiría $H_0\colon \sigma_1 = \sigma_2$ (sin cuadrados) o cualquier otra variante (como "$\geq$" o "$\leq$").

Ejemplo 9.3.1 pág. 301 y 9.3.4 pág. 305. Estudiamos la aparición de angina de pecho con el ejercicio. En un experimento con animales se hicieron dos grupos de 9 ratas. A uno se les dió un placebo y al otro un fármaco experimental. Después de la actividad física se midió el tiempo de recuperación obteniéndose los siguientes datos:

- Grupo placebo: $n_1 = 9$, $\overline{X}_1 = 329$ segundos, $\hat{s}_1 = 45$ segundos;
- Grupo fármaco: $n_2 = 9$, $\overline{X}_2 = 283$ segundos, $\hat{s}_2 = 43$ segundos.

Se pide:

1. Hacer una prueba F y comprobar que podemos suponer que las varianzas poblacionales son iguales.
2. ¿Podemos concluir que el fármaco reduce significativamente el tiempo de recuperación? Usar un nivel de confianza del $95\,\%$.
3. Dar un intervalo de confianza para la diferencia de tiempos medios de recuperación.

2.5.1. Otros contrastes*

Podríamos hacer también contrastes del tipo

$$H_0\colon \sigma_1 = 5\,\sigma_2$$

(y todas sus variantes), aunque no los veremos en este curso. En ese caso el estadístico a usar es

$$F = \frac{\hat{s}_1^2/\sigma_1^2}{\hat{s}_2^2/\sigma_2^2},$$

es decir

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2}/\frac{\sigma_1^2}{\sigma_2^2} = \frac{\hat{s}_1^2}{\hat{s}_2^2}/25.$$

Nivel de significación y valor p Hagamos un repaso (extracto de un artículo reciente en la revista *Nature* firmado por más de 800 científicos):

Si la diferencia obtenida en un experimento no es "significativa" no podemos afirmar que entre los dos grupos no hay diferencias. Es decir, aceptar H_0 no es lo mismo que "demostrar" H_0. Si el valor p es más grande que un valor de referencia como $0'05$ no debemos concluir que hemos "probado" que no hay diferencias entre dos poblaciones, o que un tratamiento no tiene efecto, o que no hay asociación entre dos variables.

Por ejemplo, aunque el intervalo de confianza incluya el 0 (lo que nos llevaría a aceptar H_0 en un contraste bilateral), eso no quiere decir que no contenga también valores positivos o negativos bastante alejados, que en otro estudio serían aceptados como significativos. Debería recordarse que todos los valores del intervalo son razonablemente compatibles con los datos (con el nivel de confianza usado) y no dar por "demostrado" ningún valor concreto. Como ya sabemos, que el intervalo de los valores "más compatibles" no significa que los valores fuera del intervalo sean incompatibes, sólo son "menos compatibles". Por otro lado,no todos los valores en un intervalo son igual de compatibles con los datos. El más compatible es el valor central, y lo son menos cuanto más alejados están.

Del mismo modo, si encontramos una diferencia significativa entre dos muestras no es seguro que sea cierta H_1.

En resumen, el hecho de que algo supere o no un valor de referencia no nos asegura que sea "cierto" o no. Debe decirse "nuestros resultados son más compatibles con que si hay un efecto no es importante".

Con esto damos por terminadas las lecciones 8 y 9 del libro.

*Las secciones marcadas con asterisco no son materia de examen

Tema 3: La prueba χ^2

Capítulo 1

Expositivas 16–18

- Contrastes para datos categóricos: tablas de contingencia. Contrastes de homogeneidad. Contrastes de independencia .
- Pruebas de bondad de ajuste.

Este tema corresponde en parte a la Lección 12 del libro.

La prueba de chi-cuadrado es un *test* de hipótesis que se utiliza para determinar si existe una diferencia estadísticamente significativa entre las frecuencias esperadas y las frecuencias observadas en una o más categorías de una *tabla de contingencia*.

Como regla general, las observaciones se clasifican en categorías mutuamente excluyentes. La hipótesis nula es que no hay diferencias entre las clases, en la población. El propósito de la prueba es calcular la probabilidad de las frecuencias observadas en el experimento, si la hipótesis nula fuese verdadera.

Estudaremos *pruebas de homogeneidad* (Sección 1.1) y *pruebas de independencia* (Sección 1.2). El procedimiento en ambas es similar.

1.1. Pruebas de homogeneidad

En ocasiones interesa decidir si dos o más grupos distintos se comportan de la misma manera respecto de una variable dada. Por ejemplo, en algunas enfermedades pueden ser determinantes el sexo, el grupo sanguíneo o la edad. Para ello veremos en este capítulo una *prueba de homogeneidad* basada en la distribución χ^2. Es una generalización de los contrastes para dos proporciones que hemos visto en el capítulo anterior.

En las pruebas χ^2 de homogeneidad, una variable (por ejemplo el género) está dividida en categorías, y contrastaremos la hipótesis de que la distribución de otra variable (por ejemplo la frecuencia de aparición de una enfermedad) es homogénea, es decir, no varía entre las distintas categorías.

Ejemplo (12.1.1 y 12.1.2 pág. 441) Para probar una nueva vacuna contra la hepatitis se toman 1083 voluntarios de los que se vacuna sólo a una parte (549). Al cabo de un tiempo se observan los siguientes casos de enfermedad.

		vacunados		
		sí	no	
hepatitis	sí	11	70	
	no	538	464	
		549		1083

El cuadro con los datos se llama una *tabla de contingencia* (la palabra *contingencia* significa "algo que puede suceder o no suceder"). En él se entrecruza la información de las dos variables, que pueden ser nominales, ordinales o de intervalo, pero deben estar clasificadas en *categorías*.

Notación

Las frecuencias de las celdas se denotarán por n_{ij} (i número de fila, j número de columna). El total o *marginal* de la fila i se denota $n_{i\bullet}$. Análogamente el total de la columna j será $n_{\bullet j}$. El número total de sujetos es n.

En el ejemplo anterior tendríamos $n_{12} = 70$ y $n_{21} = 538$; por otra parte $n_{1\bullet} = 81$ y $n_{\bullet 2} = 534$. El número total de sujetos es $n = 1083$.

Podemos completar el cuadro deduciendo los valores marginales que faltan:

		vacunados		
		sí	no	
hepatitis	sí	11	70	81
	no	538	464	1002
		549	534	1083

Idea teórica Si la distribución de casos de hepatitis fuese la misma entre los vacunados y los no vacunados (es decir si la vacuna *no* fuese efectiva), entonces la proporción de casos de hepatitis en cada grupo debería ser sensiblemente igual a la proporción de casos de hepatitis en el total de los sujetos.

Por ejemplo la proporción de casos de hepatitis entre los no vacunados, que es

$$n_{12}/n_{\bullet 2} = 70/534 = 0'1310\ldots$$

debería ser parecida a la proporción total de casos de hepatitis, que es

$$n_{1\bullet}/n = 81/1083 = 0'0747.$$

Al comparar las dos fracciones,

$$\frac{n_{12}}{n_{\bullet 2}} \sim \frac{n_{1\bullet}}{n},$$

vemos que *si hubiese homogeneidad* el valor de n_{12} debería ser parecido a

$$n_{1\bullet} \times n_{\bullet 2}/n = 81 \times 534/1083 = 39'94,$$

que llamaremos "valor estimado" $\hat{E}_{12} = 39'94$; pero en realidad el valor observado es mayor, $n_{12} = 70$. Haríamos lo mismo en las demás casillas.

El estadístico χ^2 que vamos a estudiar nos dirá si es significativa la diferencia entre

- la *frecuencia experimental* observada n_{ij}
- y la *frecuencia teórica* $\hat{E}_{ij}$ que esperaríamos si la distribucion fuese homogénea. En este caso la proporción de una casilla en relación al total de su columna debería ser sensiblemente igual a la relación entre el total de la fila y el total de sujetos:

$$\frac{\hat{E}_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n}.$$

Contraste Como hipótesis nula ponemos que, en la población, las frecuencias experimentales coinciden con las teóricas, es decir que la variable "hepatitis" se distribuye de manera similar en las dos poblaciones (vacunados y no vacunados):

$$H_0: \text{frec. teor.} = \text{frec. experim. (distribución homogénea)}$$
$$H_1: \text{frec. teor.} \neq \text{frec. experim.}$$

Si aceptamos H_0 significará que la vacuna es ineficaz, ya que *no influye* en la aparición o no de casos de hepatitis.

Tendremos que comparar cada casilla n_{ij} con la estimación teórica

$$\hat{E}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

donde el numerador es el producto de las dos marginales correspondientes a la casilla (o sea el total de su fila por el total de su columna), y el denominador es el número total de datos. Y sumaremos todas las casillas.

El estadístico

$$\sum_{i,j} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{1.1}$$

tiene una distribución aproximadamente χ^2 con $(f-1)(c-1)$ grados de libertad, donde f es el número de filas y c el número de columnas.

Es muy importante darse cuenta de que este contraste es *unilateral derecho* pues cualquier diferencia (positiva o negativa) entre la frecuencia experimental y la frecuencia teórica hace que aumente el valor del estadístico (debido a los cuadrados).

Limitaciones de la prueba χ^2(*) Como la distribución del estadístico sólo es aproximadamente χ^2, hay que tener cuidado de que se cumplan algunas condiciones. Por experiencia se recomienda que los valores de las frecuencias teóricas esperadas sean mayores que 5,

$$\hat{E}_{ij} > 5.$$

Si en algunas casillas es $\hat{E}_{ij} \leq 5$, se recomienda que esto sólo pase en pocas casillas, como mucho en *un quinto* de ellas.

*Las secciones marcadas con asterisco no son materia de examen

Si hay demasiadas casillas pequeñas, una solución es agrupar varias categorías en una sola, para tener casillas mayores.

Por último, si las muestras son demasiado grandes, la prueba χ^2 tiende a rechazar H_0, aunque sea cierta.

(Adaptado de P. Hoel, Introducción a la estadística matemática, Ariel 1976)

Cálculos En el ejemplo de la página 59 anotamos todas las estimaciones teóricas (nótese que los totales marginales no cambian):

		vacunados: sí	vacunados: no	
hepatitis	sí	11 *41.06*	70 *39.94*	81
	no	538 *507.94*	464 *494.06*	1002
		549	534	1083

El estadístico en la muestra vale

$$\frac{(11-41'06)^2}{41'6}+\frac{(70-39'94)^2}{39'94}+\frac{(538-507'94)^2}{507'94}+\frac{(464-494'06)^2}{494'06}$$
$$=48,24.$$

En una χ^2 con g.l. $=(f-1)\times(c-1)=1$, el área a la derecha de $48'24$ es prácticamente cero, por tanto *rechazamos* H_0. Es decir, las diferencias entre las frecuencias observadas y las teóricas se han ido acumulando hasta dar un valor muy alto. Esto significa que *sí que hay diferencias significativas* en cada casilla, por tanto la distribución de casos de hepatitis *no* es homogénea. Conclusión: hay una diferencia entre estar vacunado o no a la hora de contraer la hepatitis. La vacuna sí es eficaz.

Dato: `DISTR.CHI(48,2418767;1)=3,76756E-12`

En la prueba χ^2, se rechaza H_0 cuando se han encontrado diferencias significativas entre el valor observado y el teórico; el reparto teórico corresponde a un reparto homogéneo (es decir, puramente "proporcional" o "por regla de tres"). Rechazar H_0 se interpreta como que el reparto *no* es homogéneo.

La homogeneidad consiste en que la distribución (frecuencias) de una variable "B" es similar sea cual sea el nivel que examinemos de la otra. Si la variable "A" no influye en "B", y la proporción de B es, en general, por ejemplo del 50 %, esperaremos encontrar en cada grupo de A la misma proporción de B, no otra diferente.

Nota El problema anterior puede hacerse con un contraste para dos proporciones (tomando $\hat{p}_1=11/549$ la proporción de casos de hepatitis en los vacunados y $\hat{p}_2=70/534$ en los no vacunados). Sin embargo la prueba χ^2 servirá también cuando haya más de dos grupos o poblaciones o niveles de la variable a estudio, como en el ejemplo siguiente.

Ejemplo (12.2.1 p. 451, 12.2.2 p. 454) Queremos determinar si existe asociación o no entre el grupo sanguíneo y la aparición de úlceras duodenales. Tomamos una muestra de 1301 pacientes con úlcera y otra de 6313 personas sanas de control, y analizamos sus grupos sanguíneos. Obtenemos la siguiente tabla:

	Grupo sanguíneo				
	0	A	B	AB	
Pacientes	698	472	102	29	1301
Controles	2892	2625	570	226	6313

El valor del estadístico es

$$\frac{(698-613'42)^2}{613'42}+\cdots+\frac{(226-221'43)^2}{221'43}=29'12.$$

Como el valor p en χ^2_3 es muy pequeño, $p<0'001$, rechazamos H_0 y concluímos que la distribución de grupos sanguíneos no parece homogénea entre las dos poblaciones.

```
DISTR.CHI(29,12;(2-1)*(4-1))=6,8031E{-08}
```

Nota* Usando que $\sum \hat{E}_{ij}=\sum n_{ij}=n$ no es difícil ver que la fórmula (1.1) coincide con la siguiente:

$$\sum_{i,j}\frac{n_{ij}^2}{\hat{E}_{ij}}-n.$$

1.2. Pruebas de independencia

El mismo tipo de contraste χ^2 servirá para hacer *pruebas de independencia*. Cuando veamos el tema sobre "Regresión" aprenderemos a establecer fórmulas sencillas que relacionen dos variables de intervalo (por ejemplo el peso y la altura). En otros casos las variables serán nominales o estarán organizadas en categorías y querremos saber al menos si tienen alguna relación entre sí o bien son independientes. Para ello veremos en este apartado una *prueba de independencia* basada en la distribución χ^2.

Diferencia entre pruebas de homogeneidad e independencia

La diferencia entre una *prueba de homogeneidad* y una *prueba de independencia* está en cómo se seleccionan las muestras. En una prueba de *homogeneidad* se fija de antemano el tamaño de cada categoría de una de las variables. Por tanto tenemos varias muestras predeterminadas.

Por ejemplo, en la Sección 1.1 habíamos fijado de antemano cuántos voluntarios íbamos a vacunar (549) y cuántos no (534). Lo que queríamos era conocer si la proporción de casos de hepatitis era la misma en ambas poblaciones (vacunados / no vacunados).

En cambio, en una prueba de *independencia* tenemos una sola muestra. Se selecciona una muestra total de tamaño n, pero no se fijan los totales marginales de ninguna de las dos variables. Después se *clasifican* los sujetos en función de las categorías de cada variable. Pero el procedimiento es el mismo en los dos casos

El contraste permite medir si es significativa la asociación entre las dos variables.

Ejemplo (adaptado de pág. 456) Creemos que hay una relación entre el número de cloroplastos de las hojas de los árboles y el nivel de SO_2 en al aire. Se seleccionan 60 árboles y se clasifican en función del nivel de dióxido de azufre de su zona y el nivel de cloroplastos de sus hojas. Se obtienen los siguientes datos:

		Nivel de cloroplastos		
		Alto	Normal	Bajo
	Alto	5	4	13
Nivel de SO_2	Normal	5	10	5
	Bajo	7	9	2

60

SOL.: Calculamos las marginales y con ellas las frecuencias teóricas si no hubiese relación (reparto proporcional):

6.23	*8.43*	*7.33*	*22*
5.67	*7.67*	*6.67*	*20*
5.10	*6.90*	*6.00*	*18*
17	*23*	*20*	*60*

El estadístico vale $\chi^2_{\text{exp}} = 12'17$, y el área a su derecha en las tablas con g.l. $= 4$ es $0'01 < p < 0'02$. Por lo tanto el experimento no es conclusivo, ya que podríamos aceptar o rechazar H_0 según el nivel de significación que nos parezca razonable. Con un nivel $\alpha = 5\,\%$ rechazaríamos H_0, pero con $\alpha = 1\,\%$ aceptaríamos H_0.

Recuerda que el nivel de significación mide la probabilidad de cometer un error de tipo I, es decir, este nivel define el riesgo que quieres tomar con tus resultados. El riesgo se refiere a rechazar una H_0 que en realidad es cierta. Si quieres correr un riesgo pequeño, un valor $p > 0'01$ nos lleva a que no hay influencia del SO_2; es una conclusión conservadora.

Idea teórica* Recordemos que la probabilidad de que dos sucesos A y B ocurran simultáneamente en la población es

$$p(A \cap B) = p(A_{|B}) \times p(B),$$

donde $p(A_{|B})$ es la *probabilidad condicionada*. Pero si A y B son independientes entonces queda

$$p(A \cap B) = p(A) \times p(B). \tag{1.2}$$

Tomemos por ejemplo las variables X «nivel de cloroplastos» e Y «nivel de dióxido de azufre». La casilla de abajo a la izquierda ($n_{31} = 7$) representa el número de casos del suceso intersección

$$A \cap B = (X \text{ vale Alto}) \text{ y } (Y \text{ vale Bajo}).$$

La estimación de $p(A \cap B)$ será

$$\frac{n_{31}}{n} \tag{1.3}$$

donde $n_{31} = 7$ y $n = 60$. Si no hay asociación entre esas variables, este valor observado debería ser próximo a la estimación de $p(A) \times p(B)$ que es

$$\frac{n_{3\bullet}}{n} \times \frac{n_{\bullet 1}}{n} \tag{1.4}$$

donde $n_{3\bullet} = 18$ $n_{\bullet 1} = 17$.

*Las secciones marcadas con asterisco no son materia de examen

A partir de (1.3) y (1.4), simplificando una n del denominador queda que debemos comparar n_{31} con

$$\hat{E}_{11} = \frac{n_{3\bullet} \times n_{\bullet 1}}{n}$$

y lo mismo para las demás casillas. Por tanto la estimación teórica es la misma que en las pruebas de homogeneidad.

Ejemplo. Adaptado de Samuels et al. 10.5.3 Queremos ver si hay relación entre el tono del pelo y el color de los ojos. Tomamos 6800 hombres y los clasificamos de acuerdo con los dos criterios:

		Tono del pelo			
		Castaño	Negro	Rubio	Pelirrojo
	Marrones	438	288	115	16
Color de ojos	Verdes	1387	746	946	53
	Azules	807	189	1768	47

6800

Contrastamos la hipótesis H_0: el tono del pelo y el color de los ojos son independientes. En la tabla χ^2 con 6 grados de libertad obtenemos $p < 0,0001$, con lo que rechazamos H_0. Hay una evidencia muy grande de que sí hay relación entre las variables.

Suele cometerse el error de decir "las variables son independientes" al rechazar H_0, y es todo lo contrario. Si rechazamos H_0 es que hay asociación entre las variables, es decir son *dependientes.*

1.3. Pruebas de bondad de ajuste*

Muchos de los contrastes que hemos visto hasta ahora exigen que la variable estudiada tenga una distribución normal. Necesitamos una prueba que nos permita estar seguros de que ésto es cierto.

En otras ocasiones nos interesará saber si determinados datos se ajustan razonablemente a una curva dada. Para ello se utilizan las llamadas pruebas de «bondad de ajuste».

Veremos la más conocida, la prueba χ^2 de Pearson que es similar a las pruebas de homogeneidad. Esta prueba sólo puede usarse para muestras grandes y requiere que los datos estén agrupados en categorías (en la sección 1.4 puedes ver cómo agrupar datos en intervalos).

Ejemplo (adaptado del libro J. Amón, *Estadística para Psicólogos*). Los datos del Cuadro 1.1 dela página 66 están agrupados en $r = 5$ intervalos y corresponden a pesos de niños (en Kg.). Queremos ver si se ajustan a una curva normal $N(\mu, \sigma)$ de media y desviación típica desconocidas

En primer lugar tenemos que estimar los parámetros media μ y desviación típica σ de la población. Para ello usamos sus estimadores insesgados, $\mu \cong \overline{X} = 9'06$ y $\sigma \cong \hat{s} = 3'96$, que se calculan a partir de los datos como hemos explicado en la Lección 1.

El número de parámetros que hemos tenido que estimar es $k = 2$. En otras ocasiones ya conoceremos de antemano alguno de estos parámetros y no tendremos que estimarlos.

*Las secciones marcadas con asterisco no son materia de examen

X peso en Kg. intervalo	frecuencia
2–4	8
5–7	10
8–10	14
11–13	9
14–16	9

Cuadro 1.1: Cuadro de pesos y frecuencias

Podemos escribir

$$H_0\colon f = N(\mu,\sigma), \quad H_1\colon f \neq N(\mu,\sigma).$$

para indicar que queremos contrastar la hipótesis de que la distribución de nuestra variable es normal.

Cálculo de las frecuencias teóricas Veamos ahora cómo se calcularía la frecuencia teórica correspondiente a un intervalo, por ejemplo el intervalo $8-10$, si H_0 fuese cierta. En primer lugar los límites reales de ese intervalo debido a la precisión de los datos son $(7'5, 10'5)$. Al tipificarlos queda

$$\frac{7'5 - 9'06}{3'96} = -0'39, \quad \frac{10'5 - 9'06}{3'96} = +0'36.$$

Por tanto la proporción de casos en ese intervalo es en teoría, si se ajustan a la normal,

$$\begin{aligned} p(7'5 \leq N \leq 10'5) &= p(-0'39 \leq z \leq 0'36) \\ &= 0,6406 - 0,3483 = 0'2923 \end{aligned}$$

(este valor se obtiene a partir de las tablas de la normal).

Como en total hay $n = 50$ casos, en ese intervalo $8-10$ debería haber en teoría una frecuencia absoluta de

$$f_{\text{teor}} = 0'2923 \times 50 = 14'615.$$

Datos: `DISTR.NORM.ESTAND(-0,39)=0,348268273`
`DISTR.NORM.ESTAND(0,36)=0,640576433`

Análogamente rellenaríamos los otros intervalos:

intervalo	f	f_{teor}
2–4	8	6.255
5–7	10	11.160
8–10	14	14.615
11–13	9	11.400
14–16	9	6.570
	50	50

La suma de las frecuencias teóricas es la misma que la de las experimentales, es decir $n = 50$; sólo cambia el reparto entre intervalos.

Nota: En el primer intervalo y en el último hemos tenido que hacer un pequeño ajuste para que la suma total sea 50. En el primero, en vez del intervalo $(1'5, 4'5)$ hemos tomado $(-\infty, 4'5)$. En el último, en vez de $(13'5, 16'5)$ hemos calculado el área correspondiente a $(13'5, +\infty)$. Si no lo hiciésemos así, el área total no sería 1.

Prueba χ^2 de Pearson Ahora debemos decidir si si las diferencias obervadas entre f y f_{teor} son significativas o no, mediante el estadístico adecuado.

El estadístico es

$$\sum_r \frac{(f - f_{\text{teor}})^2}{f_{\text{teor}}}$$

y tiene una distribución aproximadamente χ^2 con $r - k - 1$ grados de libertad. Recordemos que

- r es el número de categorías o intervalos en que hemos agrupado los datos;
- k es el número de parámetros que hemos tenido que estimar para precisar la distribución teórica.

Es un contraste *unilateral* derecho.

En el ejemplo

$$\frac{(8 - 6'255)^2}{6'255} + \cdots + \frac{(9 - 6'57)^2}{6'57} = 2'035.$$

Para $5 - 2 - 1 = 2$ grados de libertad, el área a la derecha de este valor muestral es

$$p = p(\chi^2 \geq 2'035) = 0'36,$$

(área a la derecha) que es un valor alto, con lo que aceptamos H_0. Interpretación: las diferencias entre los datos experimentales y la distribución teórica no son significativas, por lo que podemos aceptar que nuestra variable se ajusta bien a una curva normal de media $\mu = 9'06$ y desviación típica $\sigma = 3'96$.

Dato: `DISTR.CHI(2,035;2)=0,361497555`

Ejemplo Hemos lanzado cien veces tres monedas y hemos contado en cada caso el número de caras. Obtenemos los datos del Cuadro 1.2 de la página 67:

número de caras X	frecuencia f
0	20
1	40
2	25
3	15

Cuadro 1.2: Cuadro de lanzamiento de monedas

¿Se ajustan estas frecuencias a una distribución normal con desviación típica $\sigma = 0'75$? (estamos usando la aproximación normal de la binomial). Esto tiene interés por ejemplo para saber si las monedas están trucadas.

Sol.: En este caso sólo tenemos que estimar la media μ a partir de la media muestral. Por tanto $k = 1$. Si calculamos obtenemos

$$\overline{X} = \frac{\sum fX}{n} = 1'35.$$

(Recordemos que la media de una variable binomial debería ser $np = 3 \times 0,5 = 1,5$, pero suponemos que lo hemos olvidado). El número de categorías es $r = 4$.

Ahora tenemos que calcular las frecuencias teóricas. Este problema es interesante porque la variable X es discreta (número de caras) y la distribución normal es continua. Por tanto tenemos que hacer una «corrección por continuidad» que consiste en lo siguiente: cada valor de X, por ejemplo $X = 1$, en realidad corresponde a un intervalo $(0'5, 1'5)$. De este modo podemos calcular la probabilidad

$$\begin{aligned} p(0'5 \leq N \leq 1'5) &= p(1'13 \leq z \leq 0'2) \\ &= 0'5792 - 0'1285 = 0'4507 \end{aligned}$$

(este último valor se obtiene con las tablas) y en consecuencia el número esperado de casos en ese intervalo es

$$0'4507 \times n = 45'07.$$

Haciendo lo mismo en los demás intervalos tenemos

intervalo	f	f_{teor}
$(-\infty, 0'5)$	20	12.85
$(0'5, 1'5)$	40	45.07
$(1'5, 2'5)$	25	35.81
$(2'5, +\infty)$	15	6.26

Al calcular el estadístico obtenemos

$$\frac{(20 - 12'85)^2}{12'85} + \cdots + \frac{(15 - 6'26)^2}{6'26} = 20'01$$

y en una χ^2 con $r - k - 1 = 2$ grados de libertad el valor p (el área a la derecha) es prácticamente 0. Por tanto rechazamos H_0, es decir nuestros datos no se ajustan a la normal de media $\mu = 1'35$ (estimada) y desviación típica $\sigma = 0'75$.

1.4. Cómo agrupar datos en intervalos*

Podemos agrupar unos datos en intervalos, para que sean más manejables. Perderemos en precisión pero ganaremos en sencillez. También, si queremos usar un contraste χ^2, debemos tenerlos agrupados en categorías (por ejemplo ligero /medio/ pesado).

*Las secciones marcadas con asterisco no son materia de examen

Ejemplo Supongamos que tenemos $n = 50$ datos (pesos en Kg.)

$$\begin{aligned} X: 11, 9, 11, 8, 3, 8, 3, 6, 10, 5, 6, 5, 16, 12, 13, \\ 15, 4, 15, 9, 3, 9, 12, 2, 2, 10, 7, 11, 11, 6, 6, 10, 12, 5, \\ 14, 11, 14, 16, 7, 8, 10, 9, 10, 16, 15, 4, 16, 2, 5, 9, 9. \end{aligned}$$

Seguiremos el siguiente procedimiento:

- Decidir cuántos intervalos usar. Una regla puede ser usar $\sqrt{n}$ intervalos, y como mucho 10. En nuestro ejemplo vamos a tomar 7 intervalos.
- Calcular el rango real de los datos, teniendo en cuenta su precisión. Es decir como el dato más pequeño es 2 y el mayor es 16 y tienen una precisión de $\pm 0'5$, debemos cubrir un rango desde $1'5$ hasta $16'5$, es decir $16'5 - 1'5 = 15$ unidades.
- La longitud de cada intervalo sería entonces $15/7 = 2'14$. Pero debe tener el mismo número de decimales que los datos, en este caso ninguno, así que redondeamos *hacia arriba* y tomamos intervalos de longitud 3 (redondeamos hacia arriba porque con siete intervalos de longitud 2 no cubriríamos todo el rango).
- Va a haber un exceso de $7 \times 3 - 15 = 6$ unidades (7 intervalos de longitud 3 para cubrir un rango de 15). Podemos repartirlo a ambos lados, por ejemplo 3 unidades antes de $1'5$ y otras 3 al final. Entonces los siete intervalos serán

$$(-2'5, 1'5], (1'5, 4'5], \ldots, (13'5, 16'5], (16'5, 19'5]$$

- Cada intervalo se representa (con guiones) de acuerdo con el número de decimales de los datos, es decir, el intervalo $(1'5, 4'5]$ se representa como $2 - 4$, el intervalo $(16'5, 19'5]$ como $17 - 19$, etc.
- Ahora hacemos el recuento de datos en cada intervalo. Obtenemos el Cuadro (1.1) de la página 66. En él no aparecen los dos intervalos extremos ya que no contienen datos.

Nota Hay varias reglas posibles para decidir cuántos intervalos tomar. Las más conocidas son usar $3'5\, s_X\, \sqrt[3]{n}$ (Scott) o usar $2\,(\mathrm{IQ})\,\sqrt[3]{n}$ (Friedman-Diaconis), donde $\mathrm{IQ} = Q_3 - Q_1$ es la amplitud intercuartil.

Muchas veces se cita la "regla de Sturges" $1 + \log_2 n$, pero desde 1995 se sabe que la demostración original contiene un error.

Tema 4: Regresión y correlación

Capítulo 1

Expositivas 19–20

1. Regresión: método de mínimos cuadrados, rectas de regresión.
2. Varianza total. Varianza residual y varianza explicada.
3. Correlación: coeficiente de correlación lineal.
4. Otros modelos de regresión: El modelo exponencial y el modelo potencial.
5. Contraste de hipótesis para los parámetros de la regresión.

Este Tema 4 corresponde a la Lección 11 del libro.

Cuando dos variables X, Y parecen estar asociadas, en ocasiones es posible encontrar una fórmula sencilla que las relacione, al menos aproximadamente. En este Capítulo estudiaremos la "Regresión lineal", donde la fórmula que se busca es la de una recta.

1.1. Nociones previas

Veamos algunos asuntos previos necesarios para este tema.

1.1.1. Covarianza

La *covarianza* de dos variables X, Y se define como

$$s_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n}.$$

En la práctica se usa esta otra fórmula equivalente:

$$s_{XY} = \frac{\sum XY}{n} - (\overline{X})(\overline{Y}), \tag{1.1}$$

que se lee: media de los productos menos producto de las medias.

Nota: La covarianza de una variable consigo misma es la varianza:

$$s_{XX} = \frac{\sum X^2}{n} - (\overline{X})^2 = s_X^2.$$

Nota: La covarianza S_{XY} de dos variables puede ser negativa.

Nota: Hay que tener cuidado en no confundir $\sum X^2$ (suma de los cuadrados) con $(\sum X)^2$ (cuadrado de la suma). Tampoco se debe confundir $\sum XY$ (suma de los productos) con $(\sum X)(\sum Y)$ (producto de las sumas).

1.1.2. Transformación de datos

Si tenemos unos datos X con media $\overline{X}$ y desviación típica s_X, en ocasiones será necesario transformarlos (por ejemplo para cambiar de unidades o de escala).

Sean

$$\widehat{Y} = bX + a$$

los datos transformados. Entonces se tiene:

- La *media* se transforma igual que los datos,

$$\overline{\widehat{Y}} = b\overline{X} + a.$$

- La *desviación típica*, como es una medida de dispersión, sólo cambia de escala con el valor absoluto de b, pero no le influye ni el signo de b ni el desplazamiento a. La fórmula exacta es

$$s_{\widehat{Y}} = |b| s_X. \tag{1.2}$$

  Recordemos que la desviación típica no puede ser negativa.

1.1.3. Puntuaciones desviadas y tipificadas

Como aplicación de las fórmulas anteriores se tiene:

- si tomamos puntuaciones *desviadas* respecto de la media, $x = X - \overline{X}$, entonces se cumple

$$\overline{x} = 0, \quad s_x = s_X;$$

- si tomamos puntuaciones *tipificadas*, $z = (X - \overline{X})/s_X = x/s_x$ entonces se cumple

$$\overline{z} = 0, \quad s_z = 1;$$

  nótese que $s_z^2 = (\sum z^2)/n$;

- finalmente para la covarianza tenemos

$$s_{XY} = s_{xy} = (\sum xy)/n.$$

1.2. Regresión lineal y correlación

Empezaremos estudiando variables entre las que existe una relación *lineal*, es decir una fórmula sencilla como $Y = bX + a$.

1.2.1. Regresión lineal

Hemos recolectado los siguientes datos correspondientes a tiempo transcurrido (en segundos) y velocidad de caída (m/s) de un objeto (Cuadro 1.1).

t	1	2	3	4	5
v	20.52	29.14	36.76	47.80	58.72

Cuadro 1.1: Velocidad de caída de un objeto

Si los representamos gráficamente (Figura 1.1) veremos que la «nube de puntos» (diagrama de dispersión) está prácticamente alineada, lo que sugiere que existe una relación del tipo $v = bt + a$ (ecuación de una recta) entre las dos variables, y que las discrepancias se deben a errores de precisión.

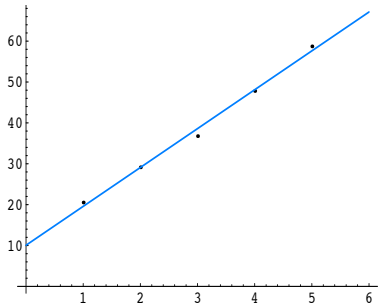


Figura 1.1: Tiempos (s) y velocidades de caída (m/s) de un objeto

El problema matemático será determinar la recta «que más se parece» a los datos.

1.2.2. Recta de regresión

Tenemos unos pares de observaciones

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

de dos variables X e Y, queremos encontrar una relación del tipo

$$\widehat{Y} = bX + a.$$

Distinguimos $\widehat{Y}$ (la estimación que vamos a hacer teóricamente) de Y (el valor que hemos obtenido experimentalmente). Llamaremos *errores de estimación* (o «valores residuales») a las diferencias

$$e = Y - \widehat{Y}.$$

Queremos encontrar una recta («recta de regresión») que cumpla las siguientes condiciones (método de mínimos cuadrados):

- la media de los errores es cero, $\overline{e} = 0$. Esto quiere decir que unos errores se compensan con otros;
- la varianza s_e^2 de los errores es mínima. Esto quiere decir que los errores están lo más concentrados posible.

Nota: En general no se puede conseguir que $s_e^2 = 0$, pues ésto significaría que todos los errores son iguales entre sí e iguales a la media $\overline{e} = 0$, es decir que todos los errores son $e = 0$, lo que sólo es cierto si los puntos de los datos están perfectamente alineados.

Con las dos condiciones anteriores, la *recta de regresión* existe y es única. Se obtienen las siguientes fórmulas:

- La pendiente de la recta de regresión es

$$b = \frac{s_{XY}}{s_X^2},$$

- La ordenada en el origen es

$$a = \overline{Y} - b\overline{X},$$

donde s_{XY} es la covarianza, s_X^2 es la varianza de X, y $\overline{X}$, $\overline{Y}$ son las medias.

Nota: La fórmula para a significa que la recta de ajuste siempre pasa por el punto de coordenadas $(\overline{X}, \overline{Y})$. Entonces la ecuación de la recta puede escribirse también en la forma punto-pendiente:

$$(\widehat{Y} - \overline{Y}) = b(X - \overline{X}).$$

Nota: Como la media de los errores es cero, las puntuaciones estimadas y las experimentales tienen la misma media: $\overline{\widehat{Y}} = \overline{Y}$. En cambio no tienen la misma varianza, como veremos más adelante.

Nota: La fórmula de b introduce muchos errores de redondeo, ya que hemos dividido entre n en seis sitios distintos (en las medias, la covarianza y la varianza). Una fórmula diferente, pero completamente equivalente, que evita esos errores de redondeo, es la siguiente:

$$b = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}.$$

Nota El término *regresión* fue introducido por Francis Galton en 1889. Estudiando la altura de padres e hijos en más de mil familias llegó a la conclusión de que los padres muy altos tenían hijos que heredaban parte de esta altura, pero que revelaban una tendencia a *regresar* a la media.

$X=t$	$Y=v$	X^2	XY	Y^2
1	20.52	1	20.52	
2	29.14	4	58.28	⋮
3	36.76	⋮	⋮	
4	47.80			
5	58.72			
15	192.94	55	673.88	8354.39

Cuadro 1.2: Disposición de los cálculos para la regresión

Disposición de los cálculos Retomemos los datos de velocidades y tiempos del Cuadro 1.1 de la página 73. Los datos se disponen como en el Cuadro 1.2 y se introducen en la calculadora para obtener las sumas totales.

Entonces (hay que escribir las fórmulas de cada cosa)

$$\begin{aligned}
\overline{X} &= \frac{\sum X}{n} = \frac{15}{5} = 3,\\
\overline{Y} &= \frac{\sum Y}{n} = \frac{192'94}{5} = 38'59,\\
s_X^2 &= \frac{\sum X^2}{n} - (\overline{X})^2 = \frac{55}{5} - (3)^2 = 2,\\
s_{XY} &= \frac{\sum XY}{n} - (\overline{X})(\overline{Y}) = \frac{673'88}{5} - (3)(38'59) = 19'01.
\end{aligned}$$

Obtenemos

$$b = \frac{s_{XY}}{s_X^2} = \frac{19'01}{2} = 9'51$$

y

$$a = \overline{Y} - b\overline{X} = 38'59 - 9'51 \times 3 = 10'07,$$

con lo que la ecuación de la regresión de v sobre t es

$$\widehat{v} = 9'51t + 10'07.$$

Nota: Deben explicitarse siempre todas las fórmulas para detectar más facilmente posibles errores de cálculo.

Estimación La fórmula en el problema anterior es bien conocida, $v = gt + v_0$, donde g es la aceleración debida a la gravedad y v_0 es la velocidad inicial.

Podemos comparar los valores estimados por esta fórmula con los valores experimentales v de la velocidad: por ejemplo si $t = 3$ segundos, la velocidad estimada es $\widehat{v} = 38'6$ m/s aunque la observada es $v = 36'76$ m/s.

También podemos estimar que al cabo de $t = 10$ segundos la velocidad de caída será $\widehat{v} \cong 105'17$ m/s.

Mínimos cuadrados* Veamos cómo se obtienen las fórmulas de la regresión. Como $Y = \widehat{Y} + e$, la media de las estimaciones $\widehat{Y}$ es la misma que la de las observaciones Y (porque

*Las secciones marcadas con asterisco no son materia de examen

la media de los errores es cero). Como $\widehat{Y} = bX + a$, por la fórmula de transformación de datos que vimos en el apartado 1.1.2 tiene que ser

$$\overline{Y} = \overline{\widehat{Y}} = b\overline{X} + a.$$

De ahí deducimos la fórmula de $a = \overline{Y} - b\overline{X}$.

Por otra parte, como $\bar{e} = 0$, la varianza de los errores es, por la fórmula de la varianza,

$$s_e^2 = \frac{\sum e^2}{n}. \tag{1.3}$$

Usando la fórmula de a, queda

$$e = Y - \widehat{Y} = Y - (bX + a) = (Y - \overline{Y}) - b(X - \overline{X})$$

Para abreviar llamemos $x = X - \bar{X}$ e $y = Y - \bar{Y}$ a las puntuaciones desviadas. Entonces

$$e = y - bx.$$

Ahora tenemos

$$e^2 = (y - bx)^2 = y^2 - 2bxy + b^2x^2$$

luego de la fórmula (1.3) se deduce

$$s_e^2 = \frac{1}{n}(\sum y^2 - 2b\sum xy + b^2\sum x^2) = s_y^2 - 2bs_{xy} + b^2s_x^2, \tag{1.4}$$

es decir

$$s_e^2 = s_Y^2 - 2bs_{XY} + b^2s_X^2. \tag{1.5}$$

Por tanto la varianza de los errores depende de b. Queremos que esa varianza sea mínima. Si lo vemos como una función de variable b, vamos a buscar un mínimo absoluto en función de b. Al derivar queda

$$\frac{ds_e^2}{db} = -2s_{XY} + 2bs_X^2.$$

Al igualar a cero,

$$0 = -s_{XY} + bs_X^2,$$

queda

$$b = \frac{s_{XY}}{s_X^2},$$

que es la fórmula buscada.

Se comprueba que hemos obtenido un mínimo, porque la derivada segunda de s_e^2 es $2s_X^2 > 0$.

1.2.3. Coeficiente de correlación

El *coeficiente de correlación de Pearson* sirve para indicar si el ajuste *lineal* que hemos realizado es bueno o malo.

Viene dado por la fórmula

$$r = \frac{s_{XY}}{s_X s_Y}. \tag{1.6}$$

Nota: Esta fórmula introduce mucho errores de redondeo, porque se divide varias veces po n (al calcular las medias, las varianzas y la covarianza). Una fórmula equivalente sin ese problema de redondeo sería

$$r = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\left(\sum X^2\right) - \left(\sum X\right)^2}\ \sqrt{n\left(\sum Y^2\right) - \left(\sum Y\right)^2}}. \tag{1.7}$$

Propiedades

- El coeficiente de correlación r tiene el mismo signo que b y que la covarianza.

  El motivo es porque

  $$r = b\,\frac{s_X}{s_Y}. \tag{1.8}$$

  y las desviaciones típicas siempre son positivas.

- El coeficiente de correlación siempre está comprendido entre -1 y $+1$:

  $$-1 \leq r \leq +1.$$

  Lo comprobaremos más adelante. A veces los errores de redondeo en un cálculo pueden dar lugar a que esto no se cumpla. En ese caso, la solución es emplear la fórmula (1.7).

Interpretación

- Cuando $r = \pm 1$ nuestros datos se ajustan perfectamente a la recta. Además, como el signo de r es el mismo que el de b,
  - una $r = +1$ quiere decir que el ajuste es perfecto y positivo (al crecer X crece Y),
  - mientras que si $r = -1$ el ajuste es perfecto pero $b < 0$, es decir al crecer X disminuye Y.
- Una $r \sim 0$ significa que el ajuste es malo.

Uso de la calculadora En una calculadora científica que tenga dos variables (Figura 1.2) pueden meterse directamente los pares de datos (X, Y). Se reconoce porque no tiene marcada en las teclas ninguna variable; a veces vienen marcadas las dos. Para eso hay que ponerla en modo "regresión" (suele llamarse `REG`), y en particular "regresión lineal". La calculadora nos dará los valores de las medias $\bar{X}$, $\bar{Y}$, las desviaciones típicas s_X, s_Y, y también los coeficientes b y a de la recta de regresión, así como el coeficiente de correlación de Pearson r.

En cambio, una calculadora de una sola variable (Figura 1.3) se reconoce porque en algunas teclas está marcada una sola variable, por ejemplo $\sum X$ y $\sum X^2$. En ella hay que meter por separado los valores de X (para obtener $\bar{X}$ y s_X), y después los de Y (para obtener $\bar{Y}$ y s_Y). La suma de la columna de los productos XY del problema de la página 75 se obtiene calculando a mano

$$1 \times 2'52 + 2 \times 29'14 + \cdots + 5 \times 58'72.$$

Figura 1.2: Una calculadora con dos variables tiene modo "regresión".

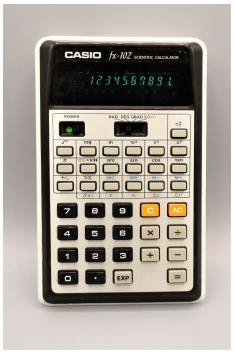


Figura 1.3: Una calculadora de una sola variable tiene modo "estadística" pero no tiene "regresión".

Pero en un problema como el de la página 78 ninguna calculadora es de gran utilidad, ya que no es posible disponer de los datos originales y ya nos dan las sumas hechas.

Consigue el manual de tu calculadora y comprueba que sabes hacer con ella los cálculos necesarios. Hay muchos websites donde puedes conseguirlo, por ejemplo este es el de la casa Casio
`https://www.casio-europe.com/es/asistencia/manuales/`

Ejemplo de cálculo: Regresión lineal (11.2.3, p. 403) La enfermedad de Crohn es una enfermedad crónica que provoca inflamación del intestino. Se ha pedido a unos pacientes que anoten diariamente información sobre varias variables clínicas, lo que permite calcular un índice numérico Y para valorar la fase en que está la enfermedad. Pero este índice es complicado de calcular y se ha ideado uno más sencillo X. Los datos que manejamos a partir de 106 observaciones son:

$$\sum X = 366'1; \quad \sum Y = 12623; \quad \sum X^2 = 2435'63; \quad \sum XY = 75989'6.$$

a) Obtener una fórmula lineal para predecir el antiguo índice Y a partir del nuevo X.

b) ¿Qué valor de Y estimaríamos para $X = 5'5$?

SOL.: Las fórmulas necesarias son

$$\text{medias} \quad \bar{X} = \frac{\sum X}{n} = \frac{366'1}{106} = 3'4538, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{12623}{106} = 119'085,$$

$$\text{varianza} \quad S_X^2 = \frac{\sum X^2}{n} - (\bar{X})^2 = \frac{2435'63}{106} - (3'4538)^2 = 11'049,$$

(no se calcula la varianza de Y porque no nos piden el coeficiente de correlación).

$$\text{desviación típica} \quad s_X = \sqrt{11'049} = 3'324,$$

$$\text{covarianza} \quad s_{XY} = \frac{\sum XY}{n} - (\bar{X}\bar{Y}) = \frac{75989'6}{106} - (3'4538)(119'085) = 305'587,$$

$$\text{pendiente de la recta} \quad b = \frac{s_{XY}}{s_X^2} = \frac{305'587}{11'049} = 27'657,$$

$$\text{ordenada en el origen} \quad a = \bar{Y} - b\bar{X} = 119'085 - (27'657)(3'4538) = 23'563.$$

En resumen, la recta de regresión es

$$\widehat{Y} = bX + a = 27'657X + 23'563.$$

El valor estimado para $X = 5'5$ será

$$\widehat{Y} = (27'657)(5'5) + 23'563 = 175'68.$$

1.3. Interpretación del coeficiente de correlación

Recordemos que el *coeficiente de correlación de Pearson* r sirve para indicar si el ajuste *lineal* que hemos realizado es bueno o malo.

Viene dado por la fórmula

$$r = \frac{s_{XY}}{s_X s_Y}. \tag{1.9}$$

1.3.1. Varianza explicada

Vamos a profundizar en la interpretación del coeficiente de correlación r.

Cada puntuación experimental se descompone como

$$Y = \widehat{Y} + e$$

donde $\widehat{Y} = bX + a$ es la estimación dada por la recta de regresión y $e = Y - \widehat{Y}$ es el error de estimación.

Al calcular las varianzas se comprueba, por unos cálculos análogos a los que hicimos al encontrar las fórmulas de mínimos cuadrados, que

$$s_Y^2 = s_{\widehat{Y}}^2 + s_e^2. \tag{1.10}$$

Esto se interpreta así: la variabilidad s_Y^2 de la *variable dependiente* Y tiene dos partes:

- una es la variabilidad de las estimaciones $\widehat{Y}$. Esta parte se llama *variabilidad explicada*, porque se deduce directamente de la variable independiente X. De hecho, por la fórmula de la regresión y las fórmulas que vimos para transformación de datos se tiene
$$s_{\widehat{Y}}^2 = b^2 s_X^2;$$

- la otra parte es la variabilidad s_e^2 debida a los errores de estimación, que no controlamos, o que dependen de otras variables distintas de X. Es la parte que hicimos que valiese lo mínimo posible al obtener las fórmulas de b y de a.

Por tanto *el ajuste de la recta es mejor o peor según sea la proporción* $s^2_{\widehat{Y}}/s^2_Y$ *entre la varianza explicada y la varianza total.*

Pero por la relación entre $r = S_{XY}/(S_X S_Y)$ y $b = S_{XY}/S_X^2$ tenemos que $b = rS_Y/S_X$. Por tanto,

la proporción de varianza explicada por la regresión es

$$\frac{s^2_{\widehat{Y}}}{s^2_Y} = \frac{b^2 s^2_X}{s^2_Y} = \frac{r^2 s^2_Y}{s^2_Y} = r^2,$$

que es el cuadrado del coeficiente de correlación de Pearson. Se llama *coeficiente de determinación.*

En resumen

- el coeficiente de determinación r^2 representa la proporción entre la varianza $S^2_{\widehat{Y}}$ "explicada" por la regresión y la varianza total S^2_Y;
- como el porcentaje de varianza explicada está entre 0 % y 100 %, siempre se tendrá $0 \leq r^2 \leq 1$;
- cuando $r^2 = 1$ (es decir $r = \pm 1$), significa que el ajuste es perfecto, pues el 100 % de la varianza estaría explicada por la regresión y no habría errores de estimación;
- el signo de r es el mismo que el de b:
  - una $r = 1$ es un ajuste perfecto positivo (la recta tiene pendiente positiva, es decir, cuanto mayor sea X mayor es Y);
  - una $r = -1$ es un ajuste perfecto pero negativo (la recta tiene pendiente negativa, es decir, a mayor X menor Y);
- por otro lado, una r próxima a cero indica un ajuste malo, pues casi toda la varianza procede del error.

Ejemplo En el problema anterior de la página 78, si fuese $\sum Y^2 = 2475309'428$, calcular el coeficiente de determinación e interpretarlo.

SOL.: La varianza de Y es

$$S_Y^2 = \frac{\sum Y^2}{n} - (\bar{Y})^2 = \frac{2475309,428}{106} - (119'085)^2 = 9170'738.$$

La desviaciones típicas son

$$S_X = \sqrt{11'049} = 3'324, \quad S_Y = \sqrt{9170'738} = 95'764.$$

Por tanto el coeficiente de correlación de Pearson es

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{305'587}{3'324 \cdot 95'764} = 0'96.$$

El coeficiente de determinación será

$$r^2 = (0'96)^2 = 0'9216.$$

Esto significa que el $92'16\,\%$ de la varianza total está explicada por la regresión (es decir, por la variable X), mientras que el $7'84\,\%$ restante es varianza residual, debida a otros factores. Aparentemente, el ajuste es muy bueno.

Varianza de los errores Ya vimos en (1.10) que la varianza de los errores de estimación es

$$s_e^2 = s_Y^2 - s_{\widehat{Y}}^2.$$

Y como $s_{\widehat{Y}}^2 = r^2 s_Y^2$ (por el cálculo del coeficiente de determinación), queda

$$s_e^2 = (1 - r^2)s_Y^2. \tag{1.11}$$

Ejemplo (11.2.1 pág. 396) En las aves acuáticas creemos que hay una relación sencilla entre el número X de horas de luz diurna en el momento de la reproducción y la duración Y en días del período de cría. Obtenemos los siguientes datos:

X	12.8	13.9	14.1	14.7	15.0	15.1	16.0	16.5	16.6	17.2	17.9
Y	110	54	98	50	67	58	52	50	43	15	28

Se calcula $b = -15'11$, $a = 290'06$ y $r = -0'85$. El coeficiente de determinación (ver apartado 1.3.1) es $r^2 = 0'73$, lo que se significa que un $73\,\%$ de la variabilidad de Y se debe a X (varianza *explicada*), quedando otro $27\,\%$ debido a otros factores (varianza de los errores o *residual*).

Además podemos hacer predicciones. Si por ejemplo la reproducción se inició con un fotoperíodo de $14'5$ horas predecimos que la cría de los polluelos durará

$$\hat{Y} = -15'11 \times 14'5 + 290'06 = 70'97$$

es decir aproximadamente 71 días.

Justificación* Vamos a comprobar la fórmula (1.10),

$$s_Y^2 = s_{\widehat{Y}}^2 + s_e^2,$$

que dice que la varianza de las puntuaciones Y (varianza *total* s_Y^2) es la suma de la varianza de las puntuaciones $\widehat{Y}$ (varianza *explicada* $s_{\widehat{Y}}^2$) y de la varianza de los errores (varianza *residual* s_e^2).

Empezamos considerando las variables *desviadas* respecto de la media, es decir

$$x = X - \bar{X}, \quad y = Y - \bar{Y}, \quad \widehat{y} = \widehat{Y} - \bar{\widehat{Y}}$$

(usamos puntuaciones desviadas porque tienen media cero, lo que simplifica los cálculos).

Sabemos que $\bar{\widehat{Y}} = \bar{Y}$ porque $\bar{e} = 0$. Entonces

$$y = \widehat{y} + e.$$

*Las secciones marcadas con asterisco no son materia de examen

Además, por las fórmulas de transformación de datos, sabemos que las varianzas no cambian,

$$s_x^2 = s_X^2, \quad s_y^2 = s_Y^2, \quad s_{\widehat{y}}^2 = s_{\widehat{Y}}^2,$$

por lo que tenemos que demostrar la fórmula

$$s_y^2 = s_{\widehat{y}}^2 + s_e^2. \tag{1.12}$$

Tenemos

$$\widehat{y} = \widehat{Y} - \bar{Y} = (bX + a) - (b\bar{X} + a) = bx,$$

es decir la recta de regresión en coordenadas desviadas pasa por el origen,

$$\widehat{y} = bx.$$

Esto tiene sentido porque habíamos visto la *interpretación del coeficiente* $a = \bar{Y} - b\bar{X}$*: la recta de regresión pasa por el punto formado por las medias.*

Ahora vamos a ver la *interpretación del coeficiente b: las variables* X *y* e *son independientes, es decir, la covarianza* $s_{Xe} = 0$.

Si desviamos, tenemos $S_{Xe} = S_{xe}$. Ahora bien,

$$\begin{aligned} S_{xe} &= \frac{\sum xe}{n} = \frac{\sum x(y - \widehat{y})}{n} = \frac{\sum xy - x\widehat{y}}{n} \\ &= \frac{\sum xy}{n} - \frac{\sum xbx}{n} = S_{xy} - bS_x^2 = 0. \end{aligned}$$

porque

$$b = \frac{S_{XY}}{S_X^2} = \frac{S_{xy}}{S_x^2}.$$

Como consecuencia,

$$s_{\widehat{y}e} = \frac{\sum \widehat{y}e}{n} = \frac{\sum bxe}{n} = bs_{xe} = 0.$$

Finalmente,

$$\begin{aligned} s_y^2 &= \frac{\sum y^2}{n} = \frac{\sum (\widehat{y} + e)^2}{n} = \frac{\sum \widehat{y}^2}{n} + 2\frac{\sum \widehat{y}e}{n} + \frac{\sum e^2}{n} \\ &= s_{\widehat{y}}^2 + 2s_{\widehat{y}e} + s_e^2 \\ &= s_{\widehat{y}} + s_e^2, \end{aligned}$$

que es la fórmula (1.12) que queríamos demostrar.

Capítulo 2

Expositiva 21

2.1. Regresión no lineal

En ocasiones la relación entre dos variables X e Y no es tan sencilla como una recta y necesitamos ajustar los datos a algún otro tipo de curva.
Estudiaremos primero, por su importancia en Farmacocinética, el *ajuste exponencial*.

2.1.1. Ajuste exponencial

Ejemplo básico Al poner una inyección intravenosa, el modelo teórico para la concentración $C = C(t)$ del medicamento en sangre viene dado por la fórmula

$$C = C_0 e^{-kt}, \tag{2.1}$$

donde C_0 es la concentración inicial y k se llama la constante de eliminación del medicamento (ver la Figura 2.1). En la Sección 2.1.2 puedes ver cómo se deduce esta formula. Ahora estudiaremos técnicas estadísticas para corregir los errores experimentales y poder estimar esas constantes.

Puedes visitar la página web
`http://vam.anest.ufl.edu/maren/index.html`
para una simulación de este modelo.

Nota Es posible conocer C_0 y k si disponemos de dos medidas experimentales *exactas* y resolvemos un sistema de dos ecuaciones con dos incógnitas. Pero en la práctica sólo tendremos varias medidas aproximadas.

Nota Desconocemos la concentración inicial C_0, porque aunque sabemos la cantidad inyectada M_0, no conocemos el "volumen aparente de distribución" V en que se difunde el medicamento. Es $C_0 = M_0/V$.

Debemos estimar la curva exponencial $\widehat{C} = C_0 e^{-kt}$ que mejor se ajusta a nuestros datos.

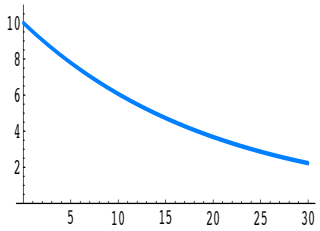


Figura 2.1: Concentración plasmática de un medicamento, $C_0 = 10$, $k = 0'05$, semivida 14.

Procedimiento para la regresión

Tomando logaritmos en la fórmula (2.1),

$$\ln \widehat{C} = -kt + \ln C_0$$

lo convertimos en un ajuste lineal, con

$$X = t, \quad Y = \ln C.$$

De los parámetros de la regresión $b = -k$ y $a = \ln C_0$ deducimos

$$k = -b, \quad C_0 = e^a.$$

Ejemplo Inyectamos por i.v. $M_0 = 125$ mg de un medicamento. Las concentraciones plasmáticas a medida que pasa el tiempo son (cuadro 2.1):

$X = t$	C	$Y = \ln C$	X^2	XY
1	5.0	1.6094	1	
2	3.0	1.0986	4	
3	2.0	0.6931	9	
4	1.5	0.4055	16	
10		3.8067	30	7.5080

Cuadro 2.1: Tiempos (h) y concentraciones (mg/l) tras una i.v.

Obtenemos $b = -0'40$ y $a = 1'96$, por tanto $k = 0'40$ y $C_0 = e^{1'96} \cong 7'1$.

Papel semi-logarítmico* Si representamos tiempos y concentraciones en un papel «semi-logarítmico» (como el de la Fig. 2.2) se obtiene directamente la representación lineal (pues en el eje vertical están marcados en realidad los logaritmos de las concentraciones). Con una regla puede obtenerse aproximadamente la recta de ajuste, y sobre el mismo papel estimar C_0 (ordenada en el origen) y k (la pendiente cambiada de signo).

En la página web `http://incompetech.com/beta/plainGraphPaper` tienes la posibilidad de generar e imprimir todo tipo de papel técnico.

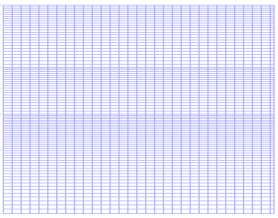


Figura 2.2: Papel semilogarítmico

Otros logaritmos* En el ajuste exponencial podemos usar logaritmos decimales (`log`) en vez de logaritmos neperianos (`ln`). Para ello, hay que tener en cuenta que para cualquier número $x > 0$ se tiene $\ln x \cong 2'30 \log x$, donde aparece la constante

$$\ln 10 = 2'30259\ldots.$$

Por lo mismo se tiene $\log e = 1/2'30$. Por eso, si usamos logaritmos decimales tendríamos

$$\log \widehat{C} = -kt \log e + \log C_0$$

y al hacer un ajuste lineal con $X = t$, $Y = \log C$, obtendremos unos parámetros de regresión $b_L = -k/2'30$ y $a_L = \log C_0$, es decir

$$k = -2'30 b_L, \quad C_0 = 10^{a_L} = e^{2'30 a_L}.$$

2.1.2. Farmacocinética: Inyección intravenosa rápida

Bolus Un *bolus* es una dosis grande de medicamento, administrada casi siempre al principio de un tratamiento, para aumentar la concentración en sangre hasta un nivel terapéutico. El término *bolus* denota una dosis de efecto rápido, en contraposición a *basal*, que se refiere a una actuación lenta en pequeñas dosis o a perfusión continua.

Inyectamos una determinada cantidad M_0 (en mg) de medicamento por vía intravenosa (i.v.) Suponemos para simplificar que la difusión en el torrente sanguíneo es instantánea (en realidad tarda unos minutos).

Volumen de distribución El medicamento va a difundirse (suponemos que uniformemente) en órganos o tejidos que ocupan un cierto volumen V (en litros); se le llama el *volumen aparente de distribución*. No podemos conocerlo directamente, pero sí estimarlo, como veremos. De hecho no es un volumen real, pues depende de la intensidad de la unión del fármaco con los tejidos o con el plasma sanguíneo.

Aclaramiento El paciente va a ir eliminando el medicamento por distintos medios (sudor, orina, heces, leche, saliva, metabolismo), a un ritmo que tampoco conocemos. El volumen de plasma depurado por unidad de tiempo se llama *aclaramiento*.

Semivida

La semivida de un medicamento es el tiempo en que la concentración en sangre después de una i.v. cae a la mitad. La fórmula es

$$t_{1/2} = \frac{\ln 2}{k}.$$

*Las secciones marcadas con asterisco no son materia de examen

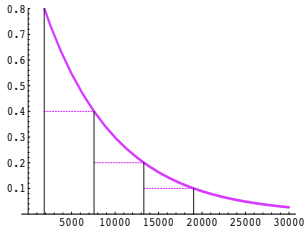


Figura 2.3: Semivida

Muchas veces se usa incorrectamente el término *vida media*. En realidad, no debe confundirse la semivida (denotada por $t_{1/2}$, en inglés HALF TIME), con la *vida media* (en inglés AVERAGE LIFE) que es $1/k$. La semivida sirve también para escribir la concentración en base 2,

$$C = C_0 2^{-t/t_{1/2}}.$$

En la página web
`http://divulgacioncientificadecientificos.blogspot.com.es/p/libro-book.html`
puedes bajarte el libro "CIENCIA y además lo entiendo!!!". El capítulo 45 se titula " ¿Por qué se habla de la semivida de un isótopo y no de su vida entera? (autor: Enrique Macías Virgós)"

Justificación El tiempo necesario para pasar de una concentración cualquiera C a la mitad es siempre el mismo (*semivida*).

En efecto, si en el momento t teníamos una concentración $C = C(t) = C_0 e^{-kt}$, al pasar un tiempo s y tener la mitad será

$$C/2 = C(t+s) = C_0 e^{-k(t+s)} = C_0 e^{-kt} e^{-ks} = C e^{-ks}$$

de donde deducimos que $1/2 = e^{-ks}$, es decir, $ks = -\ln(1/2) = \ln 2$. Queda

$$s = (\ln 2)/k.$$

medicamento	**k**	$\mathbf{t_{1/2}}$
Acetaminofeno (Paracetamol)	0.277	2.50
Diazepam (Valium)	0.021	33
Digoxina	0.0161	43
Gentamicina	0.347	2
Lidocaína	0.39	1.8
Teofilina	0.126	5.5

Cuadro 2.2: Constante de eliminación k (en 1/h) y semivida $t_{1/2}$ (en h).

Modelo matemático* Para hacer un modelo matemático de una inyección intravenosa, consideremos primero la cantidad total $M = M(t)$ de medicamento en el organismo, que

*Las secciones marcadas con asterisco no son materia de examen

va a ir cambiando con el tiempo. No podemos conocerla directamente, pero sí podremos conocer la concentración plasmática, sin más que extraer una muestra de sangre:

$$C(t) = \frac{M(t)}{V} \quad \text{(en mg/l)}.$$

Veamos cómo varía M en un breve intervalo de tiempo $[t, t + \Delta t]$. Si el paciente excreta con un aclaramiento γ l/h, entonces

$$\Delta M = +0 - \gamma \times \Delta t \times C$$

donde C es la concentración del líquido eliminado. Por tanto

$$\frac{\Delta M}{\Delta t} = -\gamma C$$

y tomando el límite cuando $\Delta t \to 0$ queda

$$M'(t) = -\gamma C(t) = -\gamma \frac{M(t)}{V}.$$

El cociente $k = \gamma/V$ se llama la *constante de eliminación* del medicamento, y se mide en $1/\,\text{h}$. El producto $\gamma = kV$ se llama *constante de aclaramiento*.

Por tanto la función M y su derivada cumplen $M'(t) = -kM(t)$, que es una ecuación diferencial lineal, homogénea, con coeficientes constantes. La solución general es $M(t) = \text{cte.} \cdot e^{-kt}$ y tomando $t = 0$ queda cte. $= M(0)$. Así tenemos $M(t) = M_0 e^{-kt}$, donde M_0 es la cantidad inicial de medicamento, es decir, la que hemos inyectado. Dividiendo todo por V, obtenemos la concentración,

$$C(t) = C_0 e^{-kt}.$$

Este modelo matemático se llama *decaimiento exponencial*. Es el mismo que aparece en otros muchos fenómenos físicos, por ejemplo la desintegración radioactiva.

PROBLEMA* Administramos un medicamento por vía intravenosa (semi-vida 20 min., volumen aparente de distribución $V = 15$ l.) ¿Cuál debe ser la cantidad de medicamento inyectada para que durante 8 horas la concentración plasmática sea superior a 20 microgramos/ml?

PROBLEMA* Se administran 500 mg de un antibiótico por vía intravenosa. A las 6 horas la concentración plasmática es de $10\,\text{mg}/\,\text{l}$. Seis horas después vueve a extraerse sangre y la concentración ha bajado a $4\,\text{mg}/\,\text{l}$. Calcular el volumen de distribución y la semi-vida.

2.2. Decaimiento radioactivo*

Algunos átomos (como por ejemplo el isótopo C^{14} del carbono) tienen un núcleo inestable y emiten espontáneamente una partícula beta (es decir, un neutrón se descompone

*Las secciones marcadas con asterisco no son materia de examen

en un protón y un electrón, y este último se emite), con lo que pasan a un estado estable (nitrógeno N^{14}). Este fenómeno radioactivo es aleatorio, así que la cantidad $M = M(t)$ de átomos radioactivos cambia a un ritmo proporcional a la cantidad de átomos radioactivos que aún quedan, o sea

$$dM/dt = -kM,$$

donde $k > 0$ (*constante de desintegración*) es un indicador de la frecuencia con que el fenómeno se produce en un momento dado.

Ejemplo* El carbono C^{14} tiene una semivida de 5730 años. Por tanto su constante de desintegración es $k = \ln 2/s = 0'000120968$. La vida media $1/k = 8266'65$ es el tiempo que por término medio tarda en descomponerse un átomo dado.

PROBLEMA* La semivida del Plutonio 238 es $87'74$ años; al emitir una partícula alfa (formada por dos protones y dos neutrones) se convierte en uranio U^{234}. En 1971, la misión Apolo XIV dejó en la Luna un pequeño reactor nuclear con $3'35$ kg de Pu^{238} ¿Qué cantidad de plutonio continúa siendo radioactiva en el año actual?

PROBLEMA* El becquerel (Bq) es una unidad que mide la radioactividad en términos de M', el ritmo de decaimiento radiactivo (1 Bq= un átomo decae cada segundo). Tras el accidente de la central nuclear de Chernobil en 1986, el nivel de contaminación del suelo por Cesio 137 en Bielorrusia era de 37 kBq/m^2 ¿Cuál será en la actualidad?

Semividas de algunos isótopos radioactivos

C^{14}	Carbono	5730 años
Co^{60}	Cobalto	$5'26$ años
I^{131}	Yodo	$8'07$ días
Cs^{137}	Cesio	$30'23$ años
Pu^{239}	Plutonio	24400 años
Ra^{226}	Radio	1602 años
U^{235}	Uranio	710000000 años

Datación por C^{14}*. El C^{14} se forma en las capas altas de la atmósfera por la acción de los rayos cósmicos. Para simplificar, podemos suponer que la proporción de C^{14} en el dióxido de carbono del aire se mantiene constante, aproximadamente $1'3 \times 10^{-12}$. Esta es la proporción que hay en las plantas y otros muchos seres vivos. Pero cuando un organismo muere, la proporción de C^{14} deja de renovarse, y disminuye exponencialmente. Viendo la proporción actual del isótopo es posible deducir la antigüedad del fósil. Por inventar este método, Willard F. Libby ganó el Nobel de Química en 1960.

2.3. Crecimiento exponencial

En el siguiente ejemplo usamos un modelo de *crecimiento exponencial*.

t	20	21	22	23	24	25	26	27	28
M	235	324	394	462	514	738	655	769	832

Cuadro 2.3: Días transcurridos y fallecidos por una epidemia

Ejemplo En la pandemia de coronavirus de 2020, el número de fallecidos diarios en España evolucionó de acuerdo a los siguientes datos:

Ajustar esos datos a una curva del tipo $M = M_0 e^{+kt}$ y hacer una previsión para el día $t = 31$. Estimar cada cuántos días se duplica el número diario de muertes.

SOL.: Como $\ln M = kt + \ln M_0$, tenemos con la calculadora

	$X = t$	$Y = \ln M$	X^2	XY	Y^2
sumas	216	56.0519	5244	1354.3803	350.5801

Cuadro 2.4: Disposición de los cálculos

Entonces $k = b = 0'15$ y $M_0 = e^a = e^{2'57} = 13'06$. Estimamos

$$M(31) = 13'06\, e^{0'15\times 31} = 1366.$$

El tiempo de duplicación es

$$t = \ln 2/k = 4'6.$$

2.4. Ajuste potencial (*regresión doble-log*)*

En ocasiones la curva que mejor se ajusta a los datos es un polinomio. El caso más sencillo es el ajuste potencial.

IMC La relación entre altura y peso no es lineal. El índice de masa corporal es el cociente Y/X^2 entre el peso Y (en kg) y el cuadrado de la altura X (en m). Se considera que este índice debe estar comprendido entre 20 y 25 para un adulto joven. Por debajo de 18 hay desnutrición y por encima de 25 se habla de sobrepeso. A partir de 30 es obesidad leve, y por encima de 40 obesidad mórbida.

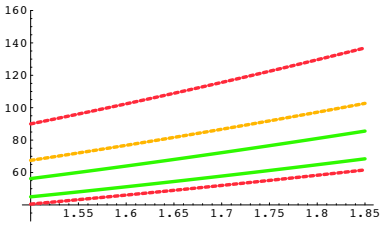


Figura 2.4: Índices de masa corporal para un adulto de 20 a 30 años: imc adecuado en verde (20–25). Para otras edades, añadir un punto al índice por cada diez años.

*Las secciones marcadas con asterisco no son materia de examen

Ejemplo El Cuadro 2.5 contiene alturas A (en m.) y pesos P (en kg.) de diez personas.

A	1.70	1.60	1.64	1.68	1.60	1.65	1.72	1.69	1.68	1.71
P	64	57	59	62	56	59	65	63	63	64

Cuadro 2.5: Pesos y alturas de diez personas

En este y otros casos nos interesa ajustar una curva del tipo

$$P = \alpha A^{\beta} \quad (\textit{ajuste potencial}).$$

La razón es que la relación entre alturas y pesos es *cuadrática* (una parábola) en vez de lineal. Se convierte en lineal tomando logaritmos,

$$\ln \widehat{P} = \beta \ln A + \ln \alpha.$$

Por tanto haremos un ajuste lineal con

$$X = \ln A, \quad Y = \ln P.$$

Los parámetros de la regresión serán $b = \beta$ y $a = \ln \alpha$, o sea $\alpha = e^a$.

Ejemplo Para las variables A altura (en m) y P peso (en kg) se han recopilado los datos del cuadro 2.5. Los cálculos aparecen en el cuadro 2.6. Al hacer la regresión lineal de los logaritmos de los pesos $Y = \ln P$ sobre los logaritmos de las alturas $X = \ln A$, obtenemos $b = 2,0064$ y $a = 3,0882$, luego $\beta = b \cong 2$ y $\alpha = e^{3,0882} \cong 22$.

A	P	$X = \ln A$	$Y = \ln P$	X^2	XY	Y^2
1.70	64	0.5306	4,1589	0,2816	2,2068	17,2963
1.60	57	0.4700	4,0431	...		...
1.64	59	0.4947	4,0775			
1.68	62	0.5188	4,1271			
1.60	56	0.4700	4,0254			
1.65	59	0.5008	4,0775			
1.72	65	0.5423	4,1744			
1.69	63	0.5247	4,1431			
1.68	63	0.5188	4,1431			
1.71	64	0.5365	4,1589			
		5.10724	41,1290	2.6144	21,0177	169,1848

Cuadro 2.6: Disposición de los cálculos para un ajuste potencial

Eso quiere decir que nuestros datos se ajustan a la curva $\widehat{P} = 22A^2$. Por ejemplo, para una persona de altura $A = 1'72$ m estimaríamos un peso de $\widehat{P} = 65$ kg («peso ideal»).

Si calculamos el coeficiente de determinación obtendremos $r^2 = 0'97$, es decir que el ajuste es muy bueno.

Capítulo 3

Expositivas 22-23

3.1. Contrastes de hipótesis para la regresión

Muchos problemas físicos, químicos y biológicos se describen mediante relaciones teóricas entre dos variables X, Y. Como esas relaciones no se pueden conocer con exactitud, en la práctica se obtienen a partir de un conjunto de observaciones que suelen contener un factor de error de carácter aleatorio.

El método de estimación más utilizado es el que hemos visto en la Sección 1.2.2. Se llama método de los mínimos cuadrados y consiste en minimizar la varianza s_e^2 de los errores. Fué inventado por K.F. Gauss entre 1795 y 1809, e independientemente por el matemático francés A.M. Legendre en 1805.

Para contrastar la hipótesis nula de que el modelo lineal es válido, se lleva a cabo un test *F de falta de ajuste*, cuyos resultados suelen presentarse en las denominadas tablas de análisis de la varianza (o tablas ANOVA). Se considera que los errores siguen una distribución normal.

(adaptado de L. Ramil y W. G. Manteiga, LA GACETA DE LA RSME, Vol. 10.2 (2007) 333–371).

3.1.1. Contraste de hipótesis para el coeficiente de correlación de Pearson

Cuando el coeficiente de correlación de Pearson r en el experimento es pequeño, podemos contrastar la hipótesis de que en la población sería realmente nulo:

$$H_0\colon \rho = 0, \quad H_1\colon \rho \neq 0,$$

lo que indicaría ausencia de relación lineal entre las variables.

El estimador de ρ (coeficiente de Pearson en la población) es r (coeficiente de Pearson en la muestra), y el estadístico que usaremos es

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = t_{n-2}. \tag{3.1}$$

Es un contraste bilateral.

Ejemplo Con $n = 25$ observaciones hemos obtenido $r = 0'18$. Entonces

$$t = \frac{0'18\sqrt{23}}{\sqrt{1 - 0'032}} = 0'88.$$

El área a la derecha de este valor en una t_{23} es $0'194$. Con un nivel de significación $\alpha = 0'01$ esa área es mayor que $\alpha/2$. Se escribe $p = 2 \times 0'194 > \alpha$. Por tanto aceptamos H_0, es decir r es tan bajo que es compatible con la hipótesis de que en la población sea $\rho = 0$.

```
DATO: DISTR.T(0,88;23;1)=0,193980935
```

Nota: En la práctica se considera que un coeficiente de correlación es nulo si se cumple que

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

está por debajo del valor t de la siguiente tabla (n es el número de pares de observaciones):

n	2	3	4	5	6	7	8	9	10
t	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17

Cuadro 3.1: Valores t_{n-2} para un nivel de confianza del 99 %

Nota Es posible plantear otros tipos de contrastes, por ejemplo $\rho = 0'80$ en la población, o para comparar dos poblaciones, $\rho_1 = \rho_2$, con muestras independientes o relacionadas, etc. También pueden darse intervalos de confianza para ρ. No los estudiaremos (ver Amón, J., Estadística para psicólogos, Ed. Pirámide, 1982, capítulo 13).

3.1.2. ANOVA

El contraste que explica el libro de Milton para la regresión (pág. 418) es completamente equivalente al que hemos estudiado en 3.1.1 para el coeficiente de correlación, salvo que usa una notación distinta y una prueba F (que en este caso concreto es el cuadrado de la t). Es interesante porque introduce una técnica nueva que se llama "análisis de varianza" (en inglés, *analysis of variance ANOVA*).

Para el ANOVA, en vez de "varianzas" usaremos "sumas de cuadrados". En vez de "covarianzas" usaremos "sumas de productos". La interpretación es la misma: son medidas de dispersión y de correlación.

Por ejemplo, en vez de la varianza

$$S_X^2 = \frac{\sum (X - \overline{X})^2}{n}$$

usaremos "sumas de cuadrados" (en inglés, *sum of squares*)

$$SC_X = \sum (X - \overline{X})^2;$$

en vez de la "covarianza"

$$S_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n}$$

usaremos "sumas de productos"

$$SP_{XY} = \sum (X - \overline{X})(Y - \overline{Y}),$$

y otras notaciones análogas.

Tenemos los siguientes términos (normalmente uno de ellos no lo calcularemos, sino que lo deduciremos de los otros dos):

- La variabilidad total SC_Y. Es la suma de cuadrados de las observaciones, es decir

  $$SC_Y = n \cdot S_Y^2. \tag{3.2}$$

  Es una χ^2 con $n-1$ grados de libertad.

- la variabilidad debida a la regresión SC_R. Es la suma de cuadrados de las estimaciones $\widehat{Y}$. Por tanto vale

  $$\sum (\widehat{Y} - \overline{Y})^2 = n \cdot S_{\widehat{Y}}^2 = n \cdot r^2 S_Y^2,$$

  porque r^2 era la proporción de varianza explicada. En resumen

  $$SC_R = r^2 SC_Y. \tag{3.3}$$

  Es una χ^2 con 1 grado de libertad.

- La variabilidad residual SC_E. Es la suma de los cuadrados de los errores, y por tanto vale

  $$SC_E = n \cdot S_e^2 = n \cdot (1 - r^2) S_Y^2. \tag{3.4}$$

  En resumen

  $$SC_E = (1 - r^2) SC_Y. \tag{3.5}$$

  Es una χ^2 con $n-2$ grados de libertad.

Podemos escribir la fórmula (1.10) como

$$SC_Y = SC_R + SC_E, \tag{3.6}$$

con lo que SC_E se puede deducir de los otros dos.

Se llaman *cuadrados medios* o *medias cuadráticas* los cocientes de las sumas de cuadrados por sus grados de libertad, es decir

$$MC_R = SC_R/1$$

y

$$MC_E = SC_E/(n-2).$$

Finalmente, el cociente

$$F = \frac{MC_R}{MC_E}$$

sigue una distribución F con 1 grado de libertad en el numerador y $n-2$ grados de libertad en el denominador.

Se trata de comparar el trozo "bueno"(MC_R) con el "malo"(MC_E) para decidir si el ajuste lineal es aceptable o no.

Suelen disponerse los cálculos de la siguiente manera:

Fuente de variabilidad	**g.l.**	**Suma de cuadrados**	**Media cuadrática**	**F**
Regresión	1	SC_R	MC_R	MC_R/MC_E
Error	$n-2$	SC_E	MC_E	
Total	$n-1$	SC_Y		

Ejemplo (11.1.3 pág. 392 y 11.4.2 pág 421)

$$H_0\colon \rho = 0, \quad H_1\colon \rho \neq 0.$$

Con $n = 28$ pares de observaciones hemos obtenido $s_{XY} = 6'52$, $s_X^2 = 3'27$, $s_Y^2 = 17'45$.

Calculamos por (3.2),

$$SC_Y = n \times S_Y^2 = 28 \times 17'45 = 488'6.$$

Además,

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{6'52}{\sqrt{3,27 \times 17,45}} = 0'86,$$

luego, por (3.3),

$$SC_R = r^2 SC_Y = 361'37$$

y anotamos los resultados en la tabla:

Fuente de variabilidad	**g.l.**	**Suma de cuadrados**	**Media cuadrática**	**F**
Regresión	1	$361'37$	$361'37$	$MC_R/MC_E = 73'90$
Error	26	$127'23$	$4'89$	
Total	27	$488'6$		

Ahora deducimos

$$SC_E = SC_Y - SC_R = 127'23$$

y calculamos

$$MC_R = 361'37/1 = 361'37$$

y

$$MC_E = 127'23/26 = 4'89.$$

Finalmente

$$F = 361'37/4'89 = 73'90.$$

Es un contraste unilateral derecho. En la distribución $F_{1,26}$, a un nivel de significación $\alpha = 1\,\%$ le corresponde un valor de $7'72$; como el valor muestral ($73'90$) es mucho mayor, rechazamos H_0, es decir ρ es significativamente diferente de cero.

Nota. No es difícil ver que

$$MC_R/MC_E = \frac{r^2(n-2)}{1-r^2}$$

con lo que este contraste F es exactamente el cuadrado de la t que estudiamos en 3.1.1.

Nota. La recta de regresión en la población adoptará la forma $\widehat{Y} = \beta X + \alpha$. Por la fórmula que relaciona r y b (y análogamente ρ y β), es lo mismo contrastar la hipótesis $\rho = 0$ que la hipótesis $\beta = 0$, donde ρ es el coeficiente de Pearson y β es la pendiente de la recta de regresión *en la población*.

Ejemplo Se aplicaron dos cuestionarios a 670 personas: uno medía el nivel de estrés al que habían estado sometidos (X) y el otro detectaba posibles trastornos de salud (Y). Al calcular el coeficiente de correlación de Pearson se obtuvo $r = 0'24$. ¿Es compatible este resultado con la hipótesis $\rho = 0$? (tomar $\alpha = 5\,\%$).

SOL.: Con estos datos no es posible completar el cuadro de un ANOVA, por lo que tenemos dos opciones:

1. Hacer el contraste t del apartado 3.1.1, es decir

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0'24\sqrt{668}}{\sqrt{1-0'24^2}} = \frac{6'20}{0'97} = 6'395.$$

   En la tabla de t con $n - 2 = 668$ grados de libertad, obtenemos $t_{\alpha/2} = t_{0'025} = 1'96$. Por tanto rechazamos H_0: $\rho = 0$, el coeficiente de correlación es significativamente distinto de cero (el ajuste no es malo).

2. O bien recordar que el valor $F = \frac{MC_R}{MC_E}$ de un ANOVA coincide con el cuadrado del contraste t anterior, es decir $F = 6'395^2 = 40'89$. En el cuadro del ANOVA sólo tendríamos

Fuente de variabilidad	**g.l.**	**Suma de cuadrados**	**Media cuadrática**	**F**
Regresión	1			40.89
Error	668			
Total	669			

En las tablas $F_{1,668}$ el valor $F_{0'05} = 3'841$, por tanto también rechazamos H_0.

3.1.3. Intervalos de estimación*

Pueden darse contrastes de hipótesis e intervalos de confianza para todos los parámetros de la regresión.

Para la pendiente β en la población se tiene el siguiente intervalo (pág. 425):

$$\beta = b \pm t_{\alpha/2} \cdot \frac{S}{S_X},$$

donde el libro llama $S = \sqrt{MC_E}$ a la raíz de la media cuadrática de los errores. La t es una distribución de Student con $n - 2$ grados de libertad.

*Las secciones marcadas con asterisco no son materia de examen

Nota. Recordemos que $MC_E = SC_E/(n-2) = (1-r^2)S_Y^2/(n-2)$.

Para la ordenada en el origen α en la población, el estimador es a, y tenemos el intervalo

$$\alpha = a \pm t_{\alpha/2} \cdot \frac{S}{S_X} \cdot \sqrt{\frac{\sum X^2}{n}},$$

donde de nuevo t tiene $n-2$ grados de libertad.

Por último, daremos un intervalo para la *respuesta media estimada* correspondiente a un valor concreto de X, digamos $X = x$. Cuando estimamos $\widehat{Y} = bx + a$ estamos dando una predicción del valor medio de Y en la población cuando la variable X vale x. Este valor medio se representa por $\mu_{Y|x}$. Se puede dar el intervalo

$$\mu_{Y|x} = \widehat{Y} \pm t_{\alpha/2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{S_X^2}}.$$

Apéndice

Lectura adicional*

En el Capítulo 13 del libro de J.S. Milton *Estadística para Biología y Ciencias de la salud*:

- Prueba de Lilliefors: una prueba de bondad de ajuste a la curva normal para muestras pequeñas.
- Prueba de los signos: contraste de hipótesis para la *mediana* de una variable continua de distribución desconocida.
- Prueba de los rangos de Wilcoxon: versión no paramétrica de la prueba t.
- Prueba de Kruskal-Wallis: compara simultáneamente la distribución de una variable continua en varias poblaciones. Es una versión no paramétrica del análisis de varianza.
- Test de Friedman: se comparan varias poblaciones pero agrupando los datos en bloques para eliminar el efecto de una variable auxiliar que podría afectar al experimento.
- Coeficiente de correlación de Spearman: una alternativa al de Pearson, fácil de calcular.
- Contraste de Bartlett para la homogeneidad de varianzas: para más de dos muestras simultáneamente.
- Contraste binomial para proporciones: cuando las muestras son pequeñas no se puede usar la aproximación normal de la binomial.

Se recomienda también el libro *Fundamentos de Estadística para las Ciencias de la vida* de M.L. Samuels *et al.*, capítulos del 6 al 12.

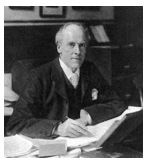
*Las secciones marcadas con asterisco no son materia de examen

Biografías

Karl F. Gauß (1777–1855) Gauss es uno de los mayores genios matemáticos de la historia. A los veinte años se hizo famoso por resolver un antiguo problema griego (la construcción con regla y compás del polígono regular de 17 lados). Trabajó en astronomía, geodesia, geometría, álgebra y física. Suyas son la "campana de Gauss" y la justificación del "método de los mínimos cuadrados", aplicadas a la estimación de órbitas de asteroides (fue director del Observatorio Astronómico de Göttingen).

K. F. Gauss

Karl Pearson (1857–1936) Pearson fué un matemático británico, fundador de la Bioestadística. Profesor en el University College de Londres, se interesó en el estudio de la herencia y la teoría de la evolución. En 1900 inventó la curva χ^2 que lleva su nombre.



K. Pearson

William Gosset (1876–1937) Gosset estudió Química y Matemáticas en Oxford. Obtuvo un puesto de químico en las destilerías Guinnes de Dublín y realizó trabajos de estadística con Pearson. Inventó la prueba t para llevar a cabo controles de calidad y publicó sus resultados en 1905 bajo el pseudómino de «Student».

W. Gosset

Ronald Fisher (1890 –1962) Fisher es el inventor de la Inferencia estadística. Estudió Matemáticas en Cambridge y se interesó en la Biología. En una novedosa Estación de Agricultura Experimental aplicó sus métodos de análisis de varianza y diseño de experimentos al estudio de la genética.

R. Fisher

George W. Snedecor (1881– 1974) Este matemático norteamericano estudió en Alabama. Profesor en Iowa, fundó el primer laboratorio de Estadística de los EE.UU. La distribución F de Snedecor fué llamada con esa letra en honor de Fisher.

G.W. Snedecor

Janet Susan Milton (1942–2024) La autora del libro que seguimos en nuestro curso. Estudió la licenciatura en ciencias en la Western Carolina University y el doctorado en el Virginia Polytechnic Institute. Fue profesora emérita de Estadística en la Radford University.

J.S. Milton