MANUAL DE
ESTADÍSTICA FARMACÉUTICA

Mariano J. Valderrama Bonnet

2021

Tema 1. MODELOS ALEATORIOS

1.1. ESQUEMA DEL PROCESO ESTADÍSTICO

Los estudios científicos persiguen conocer alguna característica de los individuos de una población con objeto de poder tomar decisiones y proceder a actuaciones futuras. Así, el primer concepto que surge es el de población, entendiendo como tal un conjunto de elementos (tangibles o no) que presentan alguna(s) característica(s) observable(s) en común, donde por observabilidad se entiende que dicha característica sea medible o susceptible de cuantificación. Se sobreentiende que la característica objeto de estudio ha de ser variable, ya que no tiene sentido considerar algo que permanece constante en todos los elementos de la población. Además, el valor o atributo de dicha variable se desconoce hasta que no se realiza la medición correspondiente, por lo que recibe el nombre de variable aleatoria. Por supuesto, no hay que confundir este concepto con el de azar, el cual no es más que un tipo particular de aleatoriedad, ya que si se considera como variable el peso de los individuos de una determinada población, éste no se asigna al azar a cada individuo, sino que se desconoce su valor hasta el momento de observarlo. Por tanto aleatoriedad es sinónimo de incertidumbre.

De cara a su estudio, las variables aleatorias se suelen clasificar de la forma siguiente:

Variables cualitativas. Son aquéllas que representan algún tipo de cualidad, modalidad o atributo que no puede expresarse numéricamente. Se subdividen en categóricas (como el color de pelo, el lugar de nacimiento, etc.) y ordinales (orden que ocupa por su calificación un opositor en el examen F.I.R., posición de las universidades según el ranking de Shangai, etc.).

Variables cuantitativas. Son aquéllas susceptibles de cuantificación numérica. Según los valores que toman sean números enteros o números reales en un intervalo se subdividen en discretas (número de hermanos, número de personas que son atendidas diariamente en una consulta, etc.) y continuas (peso de los habitantes de una localidad, nivel de glucemia basal, etc.).

Aunque en Ciencias de la Salud las poblaciones a las que se dirige un estudio usualmente son de seres humanos, los individuos que componen la población podrían ser, así mismo, animales o microbios, o pertenecientes al reino vegetal o, incluso, al mineral; pero también podrían ser elementos inmateriales como sucedería si se convoca un concurso de ideas para diseñar un logotipo, en cuyo caso son precisamente las ideas (normalmente plasmadas en un dibujo) quienes conforman la población.

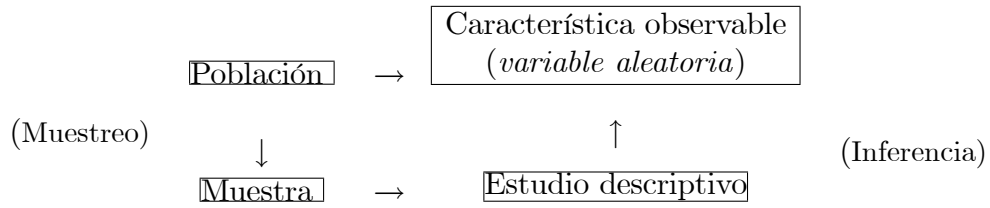
El número de elementos que integran la población se denomina tamaño poblacional, y se suele representar con la letra mayúscula N. Al estudiar alguna característica sobre seres humanos la población objeto de estudio es, casi siempre, finita, pues suelen existir censos o listados que identifican a todos los individuos conociéndose con exactitud el tamaño. Esto no ocurre cuando se trata de otro tipo de poblaciones en las que se desconoce el número de elementos que la componen y dado su gran tamaño se consideran infinitas.

Precisamente, cuando se estudian poblaciones de tamaño infinito o, incluso cuando siendo finito es muy elevado, la característica que se quiere estudiar no es posible observarla en todos y cada uno de los individuos de la población sino tan solo en un subconjunto o parte de la misma que se denomina muestra. El procedimiento mediante el que se selecciona la muestra se denomina muestreo y, como concepto general, ésta debe ser obtenida de forma aleatoria o no intencional, y ser representativa de la población de la que se extrae. El azar puede ocasionar que, en un estudio sobre hipertensión arterial, todos los individuos de la muestra tengan más de 80 años aunque el muestreo no haya seguido ninguna intencionalidad; o que el gordo de la lotería de Navidad recaiga en el número 00000.

Dado que, en consecuencia, nunca va a poder observarse la variable en todos y cada uno de los individuos de la población sino tan solo en aquéllos que componen la muestra, las conclusiones del análisis no podrán ser exactas sino afectadas de una incertidumbre que será necesario evaluar y es ahí donde entra en juego la Estadística. De tal forma, el proceso de extrapolación de

los resultados muestrales a toda la población, que se denomina inferencia, está sujeto a un erro de tipo aleatorio que es preciso modelizar.

En resumen, el esquema del proceso estadístico sería el siguiente:



1.2. DEFINICIÓN Y CLASIFICACIÓN DE VARIABLES ALEATORIAS CUANTITATIVAS

En el apartado anterior se ha introducido de manera intuitiva el concepto de variable aleatoria, en el sentido de representar una característica variable entre los individuos de una población. De manera algo más formal, si denotamos Ω al espacio muestral de un experimento aleatorio, es decir el conjunto de resultados individuales posible, $\mathbb{A}$ al álgebra de Boole asociada de sucesos y P a la probabilidad definida sobre $\mathbb{A}$, una variable aleatoria es una aplicación X: $\Omega \rightarrow$S, donde S es un conjunto numérico que, usualmente, es de los números reales (-∞,∞) tal que para todo elemento x de S, el subconjunto $\{w/X(w) \leq x\}$ de Ω pertenece al álgebra $\mathbb{A}$, es decir, es un suceso. En términos más sencillos, esta definición quiere decir que la variable aleatoria transforma sucesos en números reales.

La función F: (-∞,∞)$\rightarrow$[0,1] que transforma un número real x de la forma $F(x)$=Prob($\{w/X(w) \leq x\}$), que abreviadamente se denota $P(X \leq x)$, se denomina función de distribución, y verifica las siguientes propiedades:

1. Toma valores comprendidos entre 0 y 1, es decir: $0 \leq F(x) \leq 1$
2. Es monótona creciente, es decir si $x_1 < x_2$ entonces $F(x_1) \leq F(x_2)$
3. Es continua por la derecha en todo punto: $\lim_{x \to a^+} F(x) = F(a)$
4. $\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to \infty} F(x) = 1$

EJEMPLO

Se lanzan dos monedas al aire y se considera la variable aleatoria X que representa el número de caras obtenidas. Si denotamos cara(c) y cruz(x) entonces $\Omega = \{xx, xc, cx, cc\}$, por lo que X actúa de la forma siguiente: $X(xx) = 0 \;\; X(xc) = 1 \;\; X(cx) = 1 \;\; X(cc) = 2$. Así

$$Prob(X = 0) = \tfrac{1}{4} \;\; Prob(X = 1) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2} \;\; Prob(X = 2) = \tfrac{1}{4}$$

y la función de distribución sería:

$$\begin{array}{lcl} \text{Si } x < 0 & \rightarrow & F(x) = 0 \\ \text{Si } 0 \leq x < 1 & \rightarrow & F(x) = \frac{1}{4} \\ \text{Si } 1 \leq x < 2 & \rightarrow & F(x) = \frac{3}{4} \\ \text{Si } x \geq 2 & \rightarrow & F(x) = 1 \end{array}$$

Evidentemente las definiciones dadas de variable aleatoria y función de distribución pueden resultar excesivamente teóricas para un alumno de Farmacia, por lo que debe prevalecer el concepto intuitivo de que el objetivo es transformar un suceso aleatorio en un número real.

Dentro del análisis de las variables aleatorias nos centraremos fundamentalmente en las de tipo cuantitativo, es decir, las que toman valores numéricos. Para su estudio las dividiremos en dos tipos: discretas y continuas.

1.2.1. Variables discretas

Se dice que una v.a. es discreta si el conjunto de valores que puede tomar con probabilidad distinta de cero es finito o, a lo sumo, numerable, es decir:

$$P\{X = x_i\} = p_i \neq 0, \; i = 1, 2, ...; \quad P\{X \neq x_i\} = 0$$

Asociada a una v.a. discreta se define entonces la función de probabilidad, también denominada función de masa, de la forma:

$$f(x) = \begin{cases} p_i & x = x_i, \;\; i = 1, 2, ... \\ 0 & \text{en otro caso} \end{cases}$$

Claramente esta función es no negativa y verifica

$$\sum_i f(x_i) = 1$$

La función de distribución asociada a una v.a. discreta se expresa en términos de la función de probabilidad de la forma:

$$F(x) = \sum_{x_i \leq x} f(x_i)$$

EJEMPLOS

a) Sobre el mismo ejemplo anterior, la función de probabilidad asociada a la v.a. número de caras obtenidas al lanzar dos monedas es:

$$f(x) = \begin{cases} 1/4 & \text{si } x = 0 \\ 1/2 & \text{si } x = 1 \\ 1/4 & \text{si } x = 2 \\ 0 & \text{en otro caso} \end{cases}$$

b) Hallar el valor de la constante k de manera que $f(x) = \dfrac{k(x-1)}{n}$ sea la función de probabilidad de una v.a. discreta susceptible de tomar valores 1, 2,..., n.

Dado que, para k>0, se trata de una función no negativa, para que sea función de probabilidad debe verificar que $f(1) + f(2) + \cdots + f(n) = 1$. Así:

$$k(0 + 1 + \cdots + n - 1) = n \Rightarrow \frac{kn(n-1)}{2} = n \Rightarrow k = \frac{2}{n-1}$$

1.2.2. Variables continuas

Se dice que una v.a. es continua si el conjunto de posibles valores que puede tomar con probabilidad distinta de cero es infinito no numerable, es decir, dado que nos estamos limitando al estudio de v.a. reales, el conjunto de valores será un intervalo real o una unión de intervalos. La función de distribución asociada a la v.a. continua no tiene, por tanto, discontinuidades.

Dentro de las v.a. continuas existe un subtipo especialmente interesante y fácil de estudiar, que son las denominadas v.a. absolutamente continuas. En este caso existe una función no negativa $f(x)$, denominada función de densidad, tal que la función de distribución puede expresarse en términos de ella de la forma:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \text{ para todo } x \in \mathbb{R}$$

de manera que si $f(x)$ es continua en x, entonces, por el teorema fundamental del Cálculo, será $F'(x) = f(x)$. Además claramente es:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

En las v.a absolutamente continuas se verifica que $P\{X = x\} = 0$ para cualquier valor x, es decir, la probabilidad de que la variable tome cualquier valor aislado es siempre cero. De hecho la función de densidad no representa una probabilidad, sino que $f(x)\, dx$ se interpreta como la probabilidad infinitesimal de que la v.a. tome valores en el intervalo $[x, x + dx)$. Así se verifica que $P\{X < x\} = P\{X \leq x\}$. Por el contrario, la probabilidad de que una v.a. tome valores en un cierto subintervalo de valores no es nula, verificándose que:

$$P\{a < X \leq b\} = \int_{a}^{b} f(x)\, dx$$

EJEMPLOS

a) Hallar el valor de la constante k de manera que $f(x) = kx^2(1 - x)$ sea la función de densidad de una v.a. continua definida en $[0, 1]$. Calcular, asimismo, su moda y la probabilidad $P\{0,3 < X \leq 0,5\}$.

Claramente, para $k > 0$ se trata de una función no negativa en $[0, 1]$. Así, para que sea función de densidad, tendrá que cumplir:

$$k \int_{0}^{1} x^2(1 - x)\, dx = 1$$

por lo que $k/12 = 1$ y así, $k = 12$. La moda de X se obtendrá calculando el máximo de la densidad $f(x) = 12x^2(1 - x)$, para lo cual hay que resolver la

ecuación $f'(x) = 0$, es decir, $2x - 3x^2 = 0$. Ensayando sus soluciones, $x = 0$ y $x = 2/3$, en $f''(x) = 2 - 6x$ resulta que la moda es $x = 2/3$. Finalmente,

$$P\{0,3 < X \leq 0,5\} = 12 \int_{0,3}^{0,5} x^2 (1-x)\, dx = 0,228$$

b) La función de distribución de la v.a. que representa la duración en minutos de una llamada telefónica es:

$$F(x) = \begin{cases} 1 - \dfrac{2}{3}e^{-2x/3} - \dfrac{1}{3}e^{-x/3} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Hallar su función de densidad, así como la probabilidad de que una llamada dure entre 3 y 6 minutos.

La función densidad vendrá dada por:

$$f(x) = F'(x) = \begin{cases} \dfrac{4}{9}e^{-2x/3} + \dfrac{1}{9}e^{-x/3} & \text{si } x > 0 \\ \\ 0 & \text{si } x \leq 0 \end{cases}$$

Por otra parte,

$$P\{3 < X \leq 6\} = F(6) - F(3) = 0,1555$$

1.3. Esperanza y varianza

De igual forma que para variables estadísticas la media, como medida de posición, y la varianza, como medida de dispersión, son los parámetros más característicos, vamos a estudiar seguidamente su interpretación desde el punto de vista de las variables aleatorias.

6.3.1. Esperanza de una variable aleatoria

Sea X una v.a. discreta con función de probabilidad $f(x_i) = P\{X = x_i\}$, $i = 1, 2, \ldots, n$. Se define la *esperanza matemática* de X de la forma:

$$E[X] = \sum_{i=1}^{n} x_i f(x_i)$$

En caso de que el conjunto de posibles valores de la v.a. sea infinito numerable, la serie que define la esperanza ha de ser absolutamente convergente.

En el caso absolutamente continuo, si X tiene por función de densidad $f(x)$, la esperanza viene dada por la siguiente integral, que ha de existir:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

Si la v.a. representa el resultado de un juego de azar, la esperanza se interpreta como la ganancia o pérdida esperada del juego.

EJEMPLO

Se venden 5000 billetes para un sorteo a 100 € cada uno. Si el único premio del sorteo es de 300000 €, calcular el resultado que debe esperar una persona que compra 3 billetes.

La probabilidad de ganancia es de 3/5000, siendo en tal caso la ganancia de $300000 - 300 = 299700$ €; la probabilidad de pérdida es 4997/5000, siendo en tal caso la pérdida igual a la cantidad jugada, es decir, 300 €. Así, podemos considerar el juego como una v.a. con dos posibles valores y la esperanza es entonces:

$$E[X] = 299700\frac{3}{5000} - 300\frac{4997}{5000} = -120$$

es decir, cabe esperar en promedio una pérdida de 120 €.

La esperanza es un operador lineal en el sentido de que para cualesquiera constantes a y b verifica:

$$E[aX + bY] = aE[X] + bE[Y]$$

Además, si las v.a. X e Y son independientes, entonces:

$$E[XY] = E[X]\,E[Y]$$

6.3.2. Varianza de una variable aleatoria

Si X es una v.a., discreta o continua, con esperanza finita $E[X]$, se define la *varianza* de X como:

$$Var[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E^2[X]$$

probándose fácilmente la última igualdad. La raíz cuadrada positiva de la varianza se denomina *desviación típica*.

En general se verifica que $Var[aX] = a^2 Var[X]$, siendo a una constante cualquiera. Además si X e Y son v.a. independientes, entonces $Var[X+Y] = Var[X] + Var[Y]$, no siendo cierta esta igualdad en el caso general .

EJEMPLO

a) Una v.a. discreta toma los valores 0, 1, 2, 3 y 4 con función de probabilidad siguiente:

X	0	1	2	3	4
$f(x)$	0,3	0,25	0,25	0,1	0,1

Calcular su esperanza y varianza.

$$\begin{aligned} E[X] &= 0\times 0,3 + 1\times 0,25 + 2\times 0,25 + 3\times 0,1 + 4\times 0,1 = 1,45 \\ E\left[X^2\right] &= 0^2\times 0,3 + 1^2\times 0,25 + 2^2\times 0,25 + 3^2\times 0,1 + 4^2\times 0,1 = 3,75 \\ Var[X] &= 3,75 - 1,45^2 = 1,6475 \end{aligned}$$

b) Calcular la esperanza y varianza de la v.a. continua X definida en $(0,1)$ con función de densidad $f(x) = 12x^2(1-x)$.

$$\begin{aligned} E[X] &= 12\int_0^1 x^3(1-x)\,dx = 0,6 \\ E\left[X^2\right] &= 12\int_0^1 x^4(1-x) = 0,4 \\ Var[X] &= 0,4 - 0,36 = 0,04 \end{aligned}$$

Finalmente vamos a enunciar un resultado importante que permite calcular la probabilidad de que la desviación de una v.a. respecto de su esperanza esté acotada por una cierta cantidad.

Teorema de Tchebychev

Sea X una v.a. con esperanza y varianzas finitas. Entonces, dada una constante $k > 0$ se verifica:

$$P\{|X - E[X]| \geq k\} \leq \frac{Var[X]}{k^2}$$

1.4. Estudio de algunos modelos aleatorios discretos

A continuación vamos a estudiar algunas v.a. discretas de gran importancia práctica. La metodología que seguiremos para cada una consistirá en presentar qué tipo de fenómeno aleatorio modelizan, dando seguidamente su función de probabilidad y su esperanza y varianza.

1.4.1. Modelo binomial

Consideremos un experimento que puede dar lugar únicamente a dos resultados: A, denominado éxito, con probabilidad p, y $\overline{A}$, denominado fracaso, con probabilidad $q = 1 - p$. Este tipo de experimento se denomina prueba de Berouilli, y la v.a. que lo representa probabilísticamente, modelo de *Bernouilli*. Esta variable depende de un único parámetro p, y asigna al éxito el valor 1 y al fracaso el 0. Su función de probabilidad se reduce entonces a:

$$f(x) = p^x(1-p)^{1-x} = p^x q^{1-x},\ x = 0,1$$

Una sucesión de n pruebas de Bernouilli independientes da lugar al denominado modelo *binomial*, el cual depende de dos parámetros n y p, y se representa $B(n,p)$. Su función de probabilidad viene dada por:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, \ldots, n$$

ya que los x sucesos pueden presentarse de $\binom{n}{x}$ formas en el total de n pruebas. Esta función representa la probabilidad de obtener x éxitos en las n pruebas. Para probar que efectivamente $f(x)$ es una función de probabilidad, basta observar que:

$$\sum_{x=0}^{n} f(x) = (p+q)^n = 1$$

Además, operando convenientemente, se obtiene:

$$E[X] = np, \quad Var[X] = npq$$

EJEMPLOS

a) Un equipo de fútbol tiene 2/3 de probabilidad de ganar cuando juega en casa. Calcular la probabilidad de que gane más de dos partidos de un total de cuatro que disputa en casa.

La v.a. que representa esta situación es binomial de parámetros $n = 4$ y $p = 2/3$, por lo que su función de probabilidad particular vienen dada por:

$$f(x) = \binom{4}{x} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{4-x} = \binom{4}{x} \frac{2^x}{81}$$

Así, en nuestro caso concreto, será:

$$P\{X > 2\} = f(3) + f(4) = 4\frac{2^3}{81} + \frac{2^2}{81} = 0,592$$

b) Suponiendo que son equiprobables los sucesos tener hijo o hija, determinar el número esperado de varones en una familia de 8 hijos, así como la probabilidad de que efectivamente resulte el número esperado.

El fenómeno en cuestión puede modelizarse mediante una v.a. binomial de parámetros $n = 8$ y $p = 0,5$ de modo que su esperanza será $E[X] = 4$ hijos varones. Así:

$$P\{X = 4\} = \binom{8}{4} (0,5)^4 (0,5)^{8-4} = 0,273$$

Las probabilidades calculadas con el modelo binomial para valores de n desde 2 a 10 y determinados valores de p pueden consultarse en la *Tabla 1* del Anexo.

6.4.2. Modelo de Poisson

Es un modelo muy apropiado cuando se estudian fenómenos tales como: número de de individuos que entran en una oficina a lo largo de un día, número de unidades de un cierto medicamento vendidas durante el mes de marzo, número de accidentes de carretera ocurridos durante un fin de semana en un tramo de carretera, etc.

El modelo de *Poisson* depende de un único parámetro positivo λ, y se expresa $P(\lambda)$. Su función de probabilidad viene dada por:

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda},\ x = 0, 1, \ldots$$

Para comprobar que $f(x)$ está bien definida como función de probabilidad basta observar que $\sum_{i=1}^{n} \frac{\lambda^x}{x!}$ es el desarrollo en serie de McLaurin de la exponencial e^{λ}, de manera que:

$$f(x) = e^{-\lambda} e^{\lambda} = 1$$

además, operando convenientemente, se obtiene:

$$E[X] = \lambda, \quad Var[X] = \lambda$$

es decir, el parámetro λ del modelo coincide con su esperanza y con su varianza.

EJEMPLO

Entre las dos y las cuatro de la madrugada el promedio de personas que acuden a una cierta farmacia de guardia es 2,6. Hallar la probabilidad de que en una noche de guardia acuda en esta franja horaria sólo una persona a la farmacia.

Se trata de una variable de Poisson de parámetro $\lambda = 2,6$, de manera que:

$$f(1) = \frac{2,6}{1!}e^{-2,6} = 0,193$$

Las probabilidades calculadas con el modelo de Poisson para valores de λ desde 0,1 hasta 10 pueden consultarse en la *Tabla 2* del Anexo

Puede demostrarse que el modelo binomial $B(n,p)$ se aproxima al de Poisson cuando el número n de pruebas de Bernouilli es grande y la probabilidad de éxito p pequeña. En tal caso, el parámetro de la variable de Poisson resultante es $\lambda = np$.

EJEMPLOS

a) Se ha comprobado que el 2% de los medicamentos de un cierto lote están caducados. Hallar la probablidad de que en una muestra de 100 medicamentos de este lote haya tres caducados.

La variable que representa esta situación es binomial de parámetros $n = 100$ y $p = 0,02$. Podemos entonces aproximarla mediante una v.a. de Poisson de parámetro $\lambda = 2$, de manera que:

$$f(3) = \frac{2^3}{3!}e^{-2} = 0,181$$

b) La probabilidad de reacción adversa frente a una vacuna es de 0,001. Hallar la probabilidad de que al menos un individuo sufra reacción de un total de 2000 vacunados.

Este fenómeno se ajusta a un modelo binomial de parámetros $n = 2000$ y $p = 0,001$ pero al igual que en el ejemplo anterior, lo aproximamos por una variable de Poisson de parámetro $\lambda = 2$. Así, la probabilidad pedida sería:

$$P\{X \geq 1\} = 1 - f(0) = 1 - \frac{2^0}{0!}e^{-2} = 1 - 0,135 = 0,865$$

1.4.3. Modelo hipergeométrico

Este modelo representa fenómenos que responden genéricamente al esquema siguiente: supongamos que en una urna hay N bolas de las cuales N_1 son blancas y N_2 negras ($N_1 + N_2 = N$). Si se extraen de la urna n bolas sin reemplazamiento, la variable que representa el número de bolas blancas extraídas en la muestra se denomina v.a. *hipergeométrica*. Puede, asimismo, concebirse este esquema suponiendo que en una población de N individuos, éstos pueden clasificarse en dos bloques A y B, de los cuales N_1 pertenecen al A y N_2 al B.

La función de probabilidad del modelo hipergeométrico es:

$$f(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{x}},\ x = 0, 1, 2, \ldots, n$$

Denotando $p = \dfrac{N_1}{N}$ y $q = 1 - p$, su esperanza y varianza vienen dadas por:

$$E[X] = np, \quad Var[X] = npq\frac{N-n}{N-1}$$

1.4.4. Modelo geométrico o de Pascal

Se dice que una v.a. responde al modelo *geométrico* si representa el número de pruebas independientes de Bernouilli que hay que realizar hasta que se presenta el primer éxito. Su función de probabilidad es:

$$f(x) = pq^x,\ x = 0, 1, 2, \ldots$$

donde p es la probabilidad de éxito y $q = 1 - p$ la de fracaso. Su esperanza y varianza vienen dadas por:

$$E[X] = \frac{q}{p}, \quad Var[X] = \frac{q}{p^2}$$

1.4.5. Modelo binomial negativo

Se dice que una v.a. se ajusta al modelo *binomial negativo* si representa el número de pruebas independientes de Bernouilli a realizar de forma que aparezcan x fracasos antes del n-ésimo éxito. Así, denotando por p a la probablidad de éxito, la función de probabilidad de este modelo es:

$$f(x) = \binom{n+x-1}{x} p^n q^x, \ x = 0, 1, 2, \ldots, n+x-1$$

Su esperanza y varianza vienen dadas por:

$$E[X] = \frac{nq}{p}, \quad Var[X] = \frac{nq}{p^2}$$

1.5. Estudio de algunos modelos aleatorios continuos

En este apartado describiremos los modelos aleatorios continuos más importantes, siguiendo una metodología similar a la desarrollada en el caso discreto, es decir, presentaremos el tipo de fenómeno que puede modelizarse mediante cada v.a. y formularemos para cada uno su función de densidad, así como su esperanza y varianza.

1.5.1. Modelo normal o de Gauss

Es el más importante de todos los modelos aleatorios de tipo continuo debido al gran número de fenómenos biológicos, económicos, sociales, etc. que se aproximan a él. De hecho, a partir de un resultado fundamental en Cálculo de Probabilidades, denominado *Teorema Central del Límite*, se demuestra que otros muchos modelos, discretos o continuos, pueden aproximarse bajo ciertas condiciones al modelo normal.

También denominado modelo de *Gauss* o de Gauss-Laplace, tiene por función de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ -\infty < x < \infty$$

que depende de dos parámetros μ y $\sigma > 0$, y se representa $N(\mu, \sigma)$. Para comprobar que se trata de una verdadera densidad, basta realizar un cambio de variable en la integral de Gauss:

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$$

Su esperanza y varianza vienen dadas por:

$$E[X] = \mu, \quad Var[X] = \sigma^2$$

Gráficamente, su función de densidad tiene forma de campana, como puede apreciarse en la *Figura 1*; de ahí su apodo de campana de Gauss. Sus características principales son las siguientes:

1. Es simétrica respecto a $x = \mu$.
2. Alcanza su único máximo en $x = \mu$.
3. Es creciente para $x < \mu$ y decreciente para $x > \mu$
4. Sus puntos de inflexión son $x = \mu - \sigma$ y $x = \mu + \sigma$.
5. El eje de abscisas es asíntota horizontal

En la práctica, el cálculo directo de probabilidades con el modelo normal $N(\mu; \sigma)$ resulta prácticamente imposible debido a que la integral de su función de densidad no puede calcularse mediante métodos elementales. Para ello, lo que se hace es tipificar el modelo, operación que consiste en reducirlo a otro modelo normal de media 0 y varianza 1, es decir, $N(0; 1)$, para lo cual es necesario transformar la v.a. restándole la media y dividiendo por la desviación típica, es decir:

$$X \rightsquigarrow N(\mu; \sigma) \rightarrow Z = \frac{X - \mu}{\sigma} \rightsquigarrow N(0; 1)$$

Es fácil probar que efectivamente Z tiene media 0 y varianza 1, ya que:

$$E[Z] = \frac{1}{\sigma}E[X-\mu] = 0, \quad Var[Z] = \frac{1}{\sigma^2}E\left[(X-\mu)^2\right] = \frac{\sigma^2}{\sigma^2} = 1$$

Este modelo tipificado será simétrico respecto al eje de ordenadas, alcanza su único máximo en el origen y tiene inflexiones en los puntos -1 y +1. Una vez tipificado el modelo, las probabilidades o áreas bajo la curva normal $N(0;1)$ se buscan en la *Tabla 3* del Anexo.

EJEMPLOS

a) La temperatura T al mediodía en Granada durante el mes de Mayo se ajusta a un modelo normal de media 22^o y desviación típica 6^o. Hallar el porcentaje de días en que la temperatura está comprendida entre 16^o y 25^o.

Comenzaremos calculando la probabilidad de que la temperatura al mediodía esté comprendida en dicho intervalo un día elegido al azar del mes de Mayo. Dado que $T \rightsquigarrow N(\mu = 22; \sigma = 6)$, tipificando y usando la simetría del modelo resulta:

$$P\{16 < T < 25\} = P\left\{\frac{16-22}{6} < Z < \frac{25-22}{6}\right\} = P\{-1 < Z < 0,5\} =$$
$$= P\{Z \leq 0,5\} - P\{Z < -1\} = 0,6916 - 0,1587 = 0,5328$$

de manera que el 53,28% de los días la temperatura está comprendida entre 16^o y 25^o al mediodía.

b) El peso medio de los habitantes adultos de una población es $66kg$ y la desviación típica $5kg$. Si se elige un individuo al azar de dicha población, hallar la probabilidad de que pese más de 72 kg suponiendo que la distribución de pesos se ajusta a un modelo normal.

$$P\{X > 72\} = P\left\{Z > \frac{72-66}{5}\right\} = P\{Z > 1,2\} = 0,1151$$

Como indicábamos al comienzo, otros modelos de probabilidad pueden aproximarse al normal. Así, el teorema de *De Moivre* establece que un modelo binomial $B(n,p)$ con n grande y p tomando valores intermedios (algunos autores recomiendan que $npq > 5$), puede aproximarse a un modelo normal

de parámetros $\mu = np$ y $\sigma^2 = npq$. Asimismo, un modelo de Poisson $P(\lambda)$ con λ grande (digamos $\lambda > 5$), también puede aproximarse a uno normal de parámetros $\mu = \sigma^2 = \lambda$.

EJEMPLOS

a) En una cierta carrera universitaria el porcentaje de alumnas es del 42%. Si se eligen aleatoriamente 80 alumnos que cursen esa carrera, hallar la probabilidad de que la mitad sean mujeres.

La v.a. X que representa el número de alumnas es binomial con probabilidad de éxito $p = 0,42$. Como el número de pruebas de Bernouilli es elevado ($n = 80$), utilizaremos la aproximación normal de media $80 \times 0,42 = 33,6$ y varianza $80 \times 0,42 \times 0,58 = 19,488$, de manera que

$$P\{X > 40\} = P\left\{Z > \frac{40 - 33,6}{4,41}\right\} = P\{Z > 1,45\} = 0,0735.$$

b) En una carretera construida por la empresa *Vialis* se ha observado que cada 20 km en promedio aparece un defecto grave. Si la distribución de estos defectos se ajusta a un modelo de Poisson, calcular la probabilidad de que haya 6 defectos graves a lo largo de 200 km de carretera construida por esa empresa.

Al haber, en término medio, un defecto cada 20 km, a lo largo de 200 km habrá en promedio total 10 defectos. Por tanto, el modelo de Poisson tendrá parámetro $\lambda = 10$. Así:

$$f(6) = \frac{10^6}{6!}e^{-10} = 0,063$$

Si utilizamos la aproximación al modelo normal, sus parámetros serían: $\mu = 10$ y $\sigma^2 = 10$, por lo que escribiremos:

$$P\{5,5 < X < 6,5\} = P\left\{\frac{5,5 - 10}{3,16} < Z < \frac{6,5 - 10}{3,16}\right\} = 0,0571$$

1.5.2. Modelo exponencial

Es una especie de distribución de Poisson para el caso continuo, especialmente útil en la resolución de problemas de análisis de supervivencia, tales como el estudio de la duración de vida de un componente electrónico. Asimismo, este modelo da lugar a otros de gran interés práctico tales como el modelo de Weibull y el modelo de Gompertz. Su función de densidad viene dada por:

$$f(x) = \begin{cases} ae^{-ax} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

siendo $a > 0$ el parámetro de este modelo. Esta función es densidad para cualquier valor positivo de a, ya que

$$\int_{-\infty}^{\infty} f(x)\, dx = a \int_{0}^{\infty} e^{-ax} dx = 1$$

Integrando por partes se obtiene que:

$$E[X] = \frac{1}{a} \quad Var[X] = \frac{1}{a^2}$$

1.5.3. Modelo lognormal

Este modelo representa fenómenos tales como volumen de ventas anuales en una farmacia, PIB de los paises de la UE, etc. Su denominación se debe a que su logaritmo neperiano se ajusta a un modelo normal. Así, se trata igualmente de un modelo dependiente de dos parámetros μ y σ, siendo su función de densidad:

$$f(x) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} & \text{si } x > 0 \\ \\ 0 & \text{si } x \leq 0 \end{cases}$$

Su esperanza y varianza vienen dadas por:

$$E[X] = e^{\mu + \frac{\sigma^2}{2}}, \quad Var[X] = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

1.5.4. Modelo de Pareto

Este modelo es muy utilizado en Economía para representar el volumen de rentas familiares superiores a un cierto nivel x_0. Su función de densidad es:

$$f(x) = \begin{cases} \dfrac{\alpha x_0^{\alpha}}{x^{\alpha+1}} & \text{si } x \geq x_0 \\ \\ 0 & \text{si } x < x_0 \end{cases}$$

Este modelo depende de dos parámetros positivos x_0 y α, donde α se calcula generalmente a partir de la media muestral y toma valores próximos a 2. Su esperanza y varianza, son:

$$E[X] = \frac{\alpha x_0}{\alpha - 1}, \quad Var[X] = \frac{\alpha x_0^2}{(\alpha - 1)^2 (\alpha - 2)}$$

existiendo la varianza sólo para valores de α superiores a 2.

1.6. Distribuciones asociadas al muestreo

A partir de la distribución normal se obtienen otras tres de tipo continuo que son utilizadas a la hora de realizar inferencia a partir de muestras. Se trata de la distribución *chi-cuadrado* de Pearson, la t de Student y la F de Fisher-Snedecor.

1.6.1. Distribución chi-cuadrado de Pearson

Consideremos n variables normales $X_1, X_2, ..., X_n$ tipificadas, es decir $N(0,1)$, e independientes entre si. Se define la variable *chi-cuadrado* de Pearson con n grados de libertad, y se denota χ^2_n, como la suma de sus cuadrados: $X_1^2 + X_2^2 + ... + X_n^2$. Esta distribución es absolutamente continua y su función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{2^{-n/2}}{\Gamma\left(\frac{n}{2}\right)} e^{-x/2} x^{n/2-1} \text{ si } x > 0 \\ \\ 0 \text{ si } x \leq 0 \end{cases}$$

donde la función gamma está definida por:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \ , \quad x > 0$$

Si x es un número natural, es decir 1,2,3,..., entonces $\Gamma(x) = (x-1)!$

La gráfica de la función de densidad de la variable χ^2_n viene dada en la Figura 2.

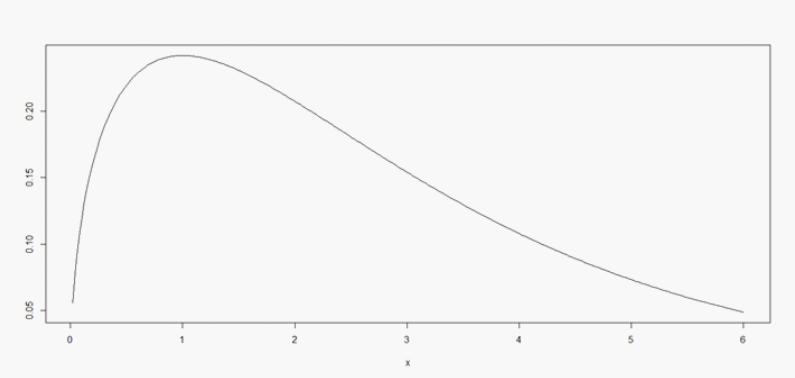


Figura 2. Distribución chi cuadrado de Pearson con 3 grados de libertad

Sus principales propiedades son las siguientes:

1. Su esperanza es $E[\chi_n^2] = n$ y su varianza $Var[\chi_n^2] = 2n$

2. Si χ_m^2 y χ_n^2 son dos variables independientes con distribuciones *chi-cuadrado* con m y n grados de libertad respectivamente, entonces la variable χ_{m+n}^2 se distribuye también según una *chi-cuadrado* con $m + n$ grados de libertad.

3. La distribución χ_2^2, es decir, la *chi-cuadrado* con 2 grados de libertad, coincide con la exponencial de parámetro $\alpha = \frac{1}{2}$.

4. Para un número elevado de grados de libertas, digamos $n > 30$, la variable $\sqrt{2\chi_n^2} - \sqrt{2n-1}$ se aproxima a una normal tipificada $N(0,1)$.

1.6.2. Distribución t de Student

Sea X una variable $N(0,1)$ y χ_n^2 una *chi-cuadrado* con n grados de libertad, ambas independientes. Se define la variable t de *Student* con n grados de libertad como la que se obtiene mediante el cociente:

$$t_n = \frac{X}{\sqrt{\chi_n^2/n}}$$

Esta distribución es absolutamente continua y su función de densidad viene dada por:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \ , \ -\infty < x < \infty$$

La gráfica de la función de densidad de la variable t_n viene dada en la Figura 3.

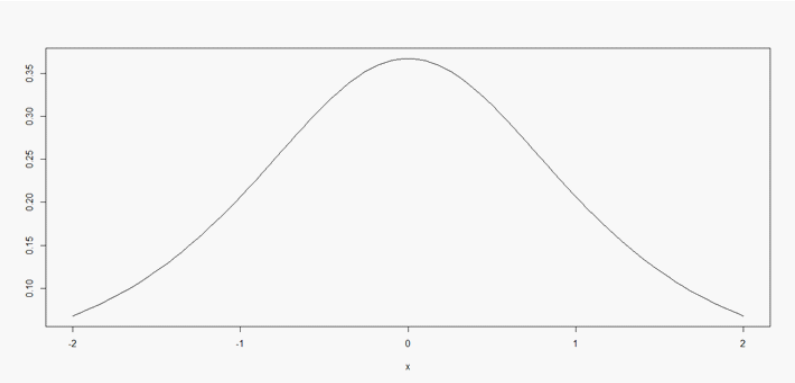


Figura 3. Distribución t de Student con 3 grados de libertad

Sus principales propiedades son las siguientes:

1. Su esperanza es $E[t_n] = 0$ (solo si $n \geq 2$) y su varianza $Var[t_n] = \frac{n}{n-2}$ (solo si $n \geq 3$)

2. La distribución límite de t_n cuando $n \to \infty$ es $N(0,1)$, es decir, la t_n es una versión platicúrtica de la $N(0,1)$.

Como anécdota cabe citar que la distibución t fue introducida en 1907 por Sir William Sealy Gosset (1876-1937) que, a la sazón, trabajaba en el Departamento de control de calidad de las destilerías Guinness en Dublín. Para los ensayos diponía de muestras pequeñas y el modelo de Gauss no resultaba útil debido al poco peso que le asignaba a los valores extremos, es decir, a las colas de la distribución, por lo que introdujo esta modificación aplastada de la ley normal. La empresa prohibía a sus empleados firmar ningún tipo de trabajo bajo su nombre propio debido a que otro investigador había publicado previamente secretos industriales sin autorización de la compañía, por lo que Gosset se vio obligado a utilizar el pseudónimo de *Student*.

1.6.3. Distribución F de Fisher-Snedecor

También conocida simplemente como distribución de *Snedecor*, para deducirla se consideran dos variables independientes, X con distribución χ^2_m e Y con χ^2_n. Se define entonces la distribución F de Fisher-Snedecor con (m, n) grados de libertad como el cociente:

$$F_{(m,n)} = \frac{\chi^2_m/m}{\chi^2_n/n}$$

Esta distribución es absolutamente continua y su función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{m^{m/2}n^{n/2}\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{x^{(m/2-1)}}{(mx+n)^{\frac{m+n}{2}}} \text{ si } x > 0 \\ \\ 0 \text{ si } x \leq 0 \end{cases}$$

La gráfica de la función de densidad de la variable $F_{(m,n)}$ viene dada en la Figura 4.

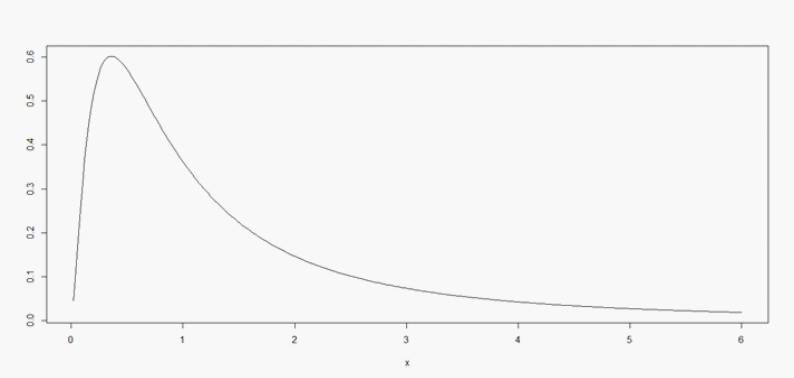


Figura 4. Distribución F de Fisher-Snedecor con (5,3) grados de libertad

Sus principales propiedades son las siguientes:

1. Su esperanza es $E[F_{(m,n)}] = \frac{n}{n-2}$ (solo si $n \geq 3$) y su varianza $Var[F_{(m,n)}] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ (solo si $n \geq 5$)

2. La distribución límite de $F_{(m,n)}$ cuando $n \to \infty$ es χ^2_m, es decir, $F_{(m,\infty)} \equiv \chi^2_m$.

3. Si $m \to \infty$ y $n \to \infty$ entonces $F_{(m,n)}$ converge en probabilidad a 1, es decir $F_{(\infty,\infty)} \equiv 1$

4. Si la variable X tiene una distribución $F_{(m,n)}$ entonces $\frac{1}{X}$ tiene la distribución $F_{(n,m)}$, por lo que $Prob(F_{(m,n)} \leq x) = Prob(F_{(n,m)} > \frac{1}{x})$

5. La variable $F_{(m,n)}$ tiene una moda en el punto $x = \frac{m-2}{m}\frac{n}{n+2}$ para $m \geq 3$.

1.7. Introducción a la inferencia estadística

La inferencia estadística es una parte fundamental del proceso estadístico que comprende un conjunto de métodos y técnicas que permiten extender o extrpolar, a partir de la información proporcionada por una muestra, el comportamiento de una determinada población. Dado que se inducen conclusiones sobre individuos no analizados, el proceso de inferencia lleva asociado un riesgo de error cuantificable en términos de probabilidad. Cuando la variable que representa la característica observable de la población se distribuye según un modelo dependiente de uno o varios parámetros, el procedimiento se reduce a extraer conclusiones sobre el valor de dichos parámetros y la inferencia se dice que es de tipo **paramétrico**. En caso contrario, o bien cuando la variable sea de tipo ordinal, será necesario recurrir a métodos de tipo **no paramétrico**.

Dentro de los métodos paramétricos cabe distinguir dos formas de realizar inferencia: mediante estimación, que puede ser puntual o por intervalo, y mediante contraste o test de hipótesis. A la inferencia por estimación se dedica el tema 2 y a los contrastes de hipótesis el tema 3.

RELACIÓN DE PROBLEMAS

1^o) Hallar la esperanza y varianza de la v.a. que representa la puntuación obtenida al lanzar un dado.

Solución: E[X]=3,5 V[X]=2,916

2^o) Dada una variable aleatoria discreta X con función de probabilidad:

X	0	1	2	3	4	5
p_i	0,1	0,3	0,4	0,1	0,05	0,05

calcular las probabilidades a) $P(X \leq 4,5)$, b) $P(X > 2)$, c) $P(2 < X \leq 4,5)$.

Solución: a) 0,95 b) 0,2 c) 0,15

3^o) Una v.a. continua se distribuye en el intevalo $[0, \infty)$ con función de densidad $f(x) = kxe^{-x^2}$. Calcular el valor de la constante k.

Solución: k=2

4^o) Una v.a. continua se distribuye en el intevalo (-∞, ∞) con función de densidad $f(x) = ke^{-x^2}$. Calcular a) el valor de la constante k; b) su esperanza y varianza

Solución: a) k=$\frac{\sqrt{\pi}}{\pi}$; b) E[X]=0 V[X]=$\frac{1}{2}$

5^o) Una v.a. continua se distribuye en el intevalo [1,3] con función de densidad $f(x) = \frac{k}{x^2}$. Calcular a) el valor de la constante k; b) $P(2 < X < 3)$; c) su esperanza

Solución: a) k=1,5; b) 0,25; c) 1,648

6^o) Una v.a. continua se distribuye en el intevalo [0,1] con función de densidad $f(x) = k\sqrt{x}$. Calcular a) el valor de la constante k; b) $P(X > 0,5)$; c) su esperanza

Solución: a) k=1,5; b) 0,65; c) 0,6

7^o) Una v.a. continua se distribuye en el intevalo [1,4] con función de densidad $f(x) = \frac{k}{\sqrt{x}}$. Calcular a) el valor de la constante k; b) $P(X \leqslant 2)$; c) su esperanza

Solución: a) k=1,5; b) 0,25; c) 1,648

8^o) Una v.a. continua se distribuye en el intevalo [1,3] con función de densidad $f(x) = kx^3$. Calcular a) el valor de la constante k; b) $P(X > 2)$; c) su esperanza y varianza

Solución: a) k=0,5; b) $\sqrt{2}-1 \approx 0,4142$; c) E[X]=$\frac{7}{3}$ V[X]=$\frac{34}{45}$

9^o) La información sobre el tiempo de vida (medido en horas) de un componente electrónico que funciona de forma ininterrumpida hasta su avería, viene dada por una función de densidad del tipo $f(x) = \frac{k}{x^3}$ en el intervalo $[100, \infty)$. Calcular: a) el valor de la constante k; b) probabilidad de que un componente funcione más de 2000 horas; c) vida esperada de un componente.

Solución: a) k=20000 b) 0,0025 c) E[X] =200 horas

10^o) La intensidad sensorial ante un estímulo eléctrico es una v.a. continua que tiene por función de densidad $f(x) = k \cos x$ siendo $0 \leq x \leq \frac{\pi}{2}$. Calcular a) el valor de la constante k; b) probabilidad de que la intensidad sensorial sea inferior a $\frac{\pi}{6}$; c) intensidad sensorial esperada.

Solución: a) k=1; b) 0,5; c) $\frac{\pi}{2}$-1

11^o) La variación de un aparto de medida es una v.a. con función de densidad $f(x) = ke^{-0,3x}$, para $x > 0$. Calcular a) el valor de la constante k; b) probabilidad de que la variaciónl sea inferior a 5 unidades; c) variación esperada.

Solución: a) k=0,3; b) 0,777; c) 0,333

12^o) El 15% de los alumnos matriculados en un curso logra superar todas las asignaturas en la primera convocatoria. Si se eligen al azar 7 alumnos de dicho curso, hallar la probabilidad de que 3 de ellos hayan superado el curso completo a la primera. Hallar también la probabilidad de que al menos dos de ellos lo consigan.

Solución: X$\rightsquigarrow$B(n=7;p=0,15)$\rightarrow$ P(X=3)=0,0617, P(X$\geq$2) =0,2834

13^o) Un examen tipo test consta de 6 preguntas con 4 respuestas posibles cada una, siendo válida sólo una de ellas. Si las respuestas incorectas no restan calificación y para aprobar es necesario acertar al menos 3 preguntas ¿qué probabilidad de aprobar tiene un alumno completamente pegado que responde al azar?

Solución: X$\rightsquigarrow$B(n=6;p=0,25)$\rightarrow$ P(X$\geq$3)=0,1694

14^o) Un libro de 500 páginas tiene un total de 300 erratas distribuidas aleatoriamente entre sus páginas. Hallar la probabilidad de que una página tomada al azar contenga al menos dos erratas.

Solución: X$\rightsquigarrow$P(λ=0,6)$\rightarrow$P(X$\geq$2)=0,122

15º) La demanda de un cierto medicamento en una farmacia se ajusta a un modelo de Poisson con promedio de demanda de 2,6 unidades diarias. Hallar la probabilidad de que durante una jornada cualquiera le demanden al menos dos unidades de dicho medicamento

Solución: $X \rightsquigarrow P(\lambda=2{,}6) \rightarrow P(X \geq 2)=0{,}7326$

16º) En el servicio de cardiología de un hospital fallece un promedio mensual de 2 pacientes. ¿Cuál es la probabilidad de que en un mes determinado se supere dicho promedio?.

Solución: $X \rightsquigarrow Poisson(\lambda=2) \rightarrow P(X>2)=0{,}3233$

21º) Un dado se lanza al aire 120 veces: a) ¿cuál es la esperanza de que aparezca un 4?; b) hallar la probabilidad de que aparezca entre 14 y 24 veces la puntuación 4.

Solución: a) 20 b) 0,7657

17º) Para el estudio de la actividad farmacológica de una sustancia se le suministró una dosis de $50mg$ a un grupo de 30 cobayas, observándose que en 24 de ellas la respuesta era positiva. Si se le administra dicha sustancia a otras 100 cobayas: a) ¿en cuántas se espera una respuesta positiva?; b) hallar la probabilidad de que entre 70 y 85 cobayas la respuesta sea positiva.

Solución: a) E[X]=80 b) Aplicando la aproximación de DeMoivre: P=0,8882

18º) Se han comprado 7 pinos procedentes de un área donde el 5% están afectados por una plaga. a) Calcular la probabilidad de que entre los pinos comprado al menos 3 estén afectados por la plaga; b) si se hubieran comprado 40 pinos ¿cuál sería la probabilidad de que más del 10% estuvieran afectados por la plaga?

Solución: a) $X \rightsquigarrow B(n=7;p=0{,}05) \rightarrow P(X \geq 3)=0{,}0038$;
b) $X \approx Poison(\lambda=2)$: $P(X>4)=0{,}0527$

19º) Un tratamiento anticelulítico es eficaz en el 45% de los casos ensayados. a) Si se aplica a un grupo de 5 personas calcular la probabilidad de que sea eficaz en al menos 4 de ellos; b) número esperado de personas en las que el anticelulítico es eficaz, y varianza asociada; c) si se aplica a un grupo de 50 personas calcular la probabilidad de que sea eficaz entre 20 y 30.

Solución: a) $X \rightsquigarrow B(n=5;p=0{,}45) \rightarrow P(X \geq 4)=0{,}1313$; b) E[X]=2,25 V[X]=1,24
c) Aplicando la aproximación de DeMoivre: P=0,7445

20^o) Un almacén farmacéutico recibió un lote de 1000 botes de suero fisiológico. Sabiendo que, en promedio, 3 de cada 1000 botes se deterioran en el transporte, hallar la probabilidad de que se rompa al menos un bote y la de que se rompan menos de dos.

Solución: $X \rightsquigarrow \text{Binomial}(n=1000;p=0{,}003) \approx \text{Poison}(\lambda=3)$
$\rightarrow P(X \geq 1)=0{,}9502$; $P(X<2)=0{,}1992$

21^o) Se ha observado que el 2 por 1000 de los individuos presentan reacción adversa ante una cierta vacuna. Hallar la probabilidad de que en un colectivo de 2000 personas vacunadas, al menos 3 de ellas tengan reacción adversa.

Solución: 0,7619

22^o) El 15% de los escolares de cierto nivel educativo presentan sobrepeso. Calcular la probabilidad de que: a) si se eligen 8 escolares al azar al menos dos de ellos presenten sobrepeso; b) si se eligen 80 escolares al azar haya entre 10 y 15 con sobrepeso.

Solución: a) $X \rightsquigarrow B(n=8;p=0{,}15) \rightarrow P(X \geq 2)=0{,}3428$;
b) $X \rightsquigarrow B(n=80;p=0{,}15) \approx N(\mu=12;\sigma=3{,}194) \rightarrow P(10 \leq X \leq 15)=0{,}5621$

23º) Una cierta vacuna produce reacción en el 10% de los casos. Si se vacunan 1000 personas, hallar el número esperado de casos con reacción, así como la probabilidad de obtener 80 ó más casos con reacción.

Solución: a) E[X]=100 casos b) p=0,9826

24º) Las calificaciones de un examen se distribuyen según una ley normal de media 6 y desviación típica 1,44. Hallar la probabilidad de que un alumno tomado al azar obtenga calificación: a) mayor que 5; b) entre 5,5 y 7.

Solución: a) 0,7549 b) 0,392

25º) El contenido de riboflavina de un determinado tipo de alga sigue una distribución normal de media 5,2 y desviación típica 3,7. Calcular el porcentaje de algas con contenido de riboflavina comprendido entre 1,5 y 12,6.

Solución: 81,86%

26º) El nivel de colesterol total en una población adulta es una variable aleatoria que se distribuye según un modelo normal con media $192 mg/dL$ y desviación típica $25 mg/dL$. a) Calcular la probabilidad de que un individuo tomado al azar de dicha población tenga nivel de colesterol total superior a $250 mg/dL$; b) hallar el porcentaje de individuos con nivel de colesterol total comprendido entre $180 mg/dL$ y $200 mg/dL$.

Solución: a) 0,0102 b) 30,99%

27º) En un ensayo farmacológico sobre un nuevo principio activo se ha observado que era ineficaz en solo un 10% de los casos estudiados. Calcular la probabilidad de que: a) si se ensaya sobre 8 individuos sea ineficaz en menos de 2; b) si se ensaya sobre 60 pacientes sea ineficaz en menos de 5.

Solución: a) $X \rightsquigarrow B(n=8;p=0,1) \rightarrow P(X<2)=0,8131$

b) $X \approx N(\mu=6;\sigma=2,32) \rightarrow P(X<5)=0,3336$

28º) Se ha comprobado que el 15% de los vacunados de gripe común presentan algún tipo de reacción cutánea. a) Si en un centro de salud se vacunan 9 personas, calcular la probabilidad de que 4 de ellas tengan reacción y de que, al menos, 7 no tengan reacción; b) si en otro centro de salud se vacunan 500 personas, calcular la probabilidad de que, al menos, 60 de ellas tengan reacción; c) si en un tercer centro de salud vacunan a otras 500 personas con otro tipo de vacuna que da reacción sólo en el 2% de los casos, calcular la probabilidad de que, al menos, 2 presenten reacción.

Solución: a) $X \rightsquigarrow B(n=9;p=0,15) \rightarrow P(X=4)=0,0283$ $P(X \leqslant 2)=0,8592$

b) $X \rightsquigarrow B(n=500;p=0,15) \approx N(\mu=75;\sigma=7,98) \rightarrow P(X \geq 60)=0,97$
c) $X \rightsquigarrow B(n=500;p=0,02) \approx \text{Poisson}(\lambda=10) \rightarrow P(X \geq 2)=0,9995$

29^o) En un estudio reciente publicado en España se estima que el 80% de las personas que practican una dieta fracasan. Para estudiar las causas de dicho fracaso se lleva a cabo un estudio piloto donde se toman al azar 8 personas que practican dicha dieta. Calcular: a) probabilidad de que 3 individuos que siguen la dieta no fracasen; b) probabilidad de que no fracasen más de 5; c) número esperado y desviación típica de fracasados en la dieta; d) si la muestra de seguidores de la dieta se ampliara a 100 personas, probabilidad de que no fracase un número comprendido entre 15 y 25.

Solución: a) $X \rightsquigarrow B(n=8;p=0,2) \rightarrow P(X=3)=0,1468$; b) $P(X>5)=0,0012$
c) $E[X]=6,4$ $D[X]=1,13$; d) $X \approx N(\mu=20;\sigma=4) \rightarrow P(15<X<25)=0,7888$

30^o) El 15% de los individuos que acuden a una consulta psiquiátrica padecen insomnio. Si se eligen 6 pacientes al azar, calcular: a) probabilidad de que al menos uno padezca insomnio; b) número esperado y desviación típica de pacientes con insomnio; c) si se amplía la muestra a 60 pacientes, probabilidad de que lo padezcan entre 5 y 10.

Solución: a) $X \rightsquigarrow B(n=6;p=0,15) \rightarrow P(X \geq 1)=0,6229$; b) $E[X]=0,9$ $D[X]=0,8746$
c) $X \approx N(\mu=9;\sigma=2,766) \rightarrow P(5<X<10)=0,5657$

Tema 2. ESTIMACIÓN

2.1. CONCEPTO Y PROPIEDADES DE UN ESTIMADOR

En el tema 1 hemos comentado que el objeto de un estudio científico es conocer el comportamiento de una población desde el punto de vista de una magnitud variable la cual, en numerosas ocasiones, se ajusta a un cierto modelo matemático conocido dependiente de uno o varios parámetros, tales como la media poblacional, su desviación típica, ... En tal caso el objetivo es conocer el valor de dichos parámetros para cada situación particular estudiada. El problema surge cuando las poblaciones son muy grandes (incluso infinitas) o el sistema de medida de la variable es complejo y costoso, de forma que no puede medirse la variable sobre todos y cada uno de los individuos que componen la población, siendo necesario recurrir a una muestra de la misma sobre la que realizar las observaciones. Por esta razón, el posterior proceso de inferencia está afectado de incertidumbre en cuanto que no podremos conocer el valor exacto de los parámetros al no disponer de valores de la variable en todos los individuos de la población.

Así, consideremos el caso más simple en el que una variable aleatoria cuantitativa X se ajusta a un modelo de probabilidad dependiente de un parámetro θ, cuya función de probabilidad o de densidad denotaremos $f(x/\theta)$. En el caso de depender de varios parámetros $\theta_1, \theta_2, ..., \theta_k$, el esquema sería similar y la función de probabilidad o de densidad se denotaría $f(x/\theta_1, \theta_2, ..., \theta_k)$. Consideremos una muestra aleatoria genérica de tamaño n de la variable X: $x_1, x_2, ..., x_n$. Se denomina **estimador** del parámetro θ a una función de la muestra $\varphi(x_1, x_2, ..., x_n)$ tal que para cada conjunto particular de valores muestrales proporciona un valor aproximado de θ que se denomina **estimación** de θ y se denota $\hat{\theta}_n$, es decir, si $x_1 = a_1, x_2 = a_2, ..., x_n = a_n$,

siendo $a_1, a_2, ..., a_n$ valores numéricos, entonces $\varphi(a_1, a_2, ..., a_n) = \hat{\theta}_n \approx \theta$. Observemos que el estimador es también una variable aleatoria pues para cada conjunto de datos muestrales toma un valor numérico diferente.

EJEMPLO

Consideremos la variable X que representa el número de ansiolíticos de un cierto tipo dispensados diariamente en las farmacias de Andalucía, la cual se distribuye según un modelo de Poisson de parámetro λ que, como es sabido, representa el promedio de ventas diarias de ese tipo de medicamento en la región. El valor de λ es desconocido, pues para conocerlo con exactitud necesitaríamos conocer las ventas en todas las farmacias andaluzas, de forma que para obtener un valor aproximado consideraremos el estimador media muestral obtenido a partir de los datos proporcionados por n farmacias de Andalucía:

$$\varphi\left(x_1, x_2, ..., x_n\right) = \frac{x_1 + x_2, ... + x_n}{n}$$

Supongamos que el Consejo Andaluz de Colegios Farmacéuticos recoge diariamente información de 4 farmacias diferentes y que los datos son los siguientes:

Muestra	Datos				Media
Día 1	4	6	1	3	3,50
Día 2	2	7	3	1	3,25
Día 3	3	7	5	2	4,25
...		...			...

Cada uno de los valores medios obtenidos, 3,50 3,25 4,25 ..., es una estimación del número medio de ansiolíticos del tipo estudiado que dispensan las farmacias andaluzas y todos son igual de válidos, es decir, tan correcto es considerar que el promedio de unidades dispensadas diariamente en cada farmacia es 3,50 como decir 3,25 o 4,25.

Pero también sería un estimador del parámetro λ la media geométrica: $\psi(x_1, x_2, ..., x_n) = \sqrt[n]{x_1.x_2 \cdots x_n}$, que para cada uno de los días anteriores tomaría respectivamente los valores 2,91 2,55 3,81 ... Cabe entonces plantearse si cualquier función de la muestra puede considerarse un estimador aceptable.

Pues bien, para que un estimador sea estadísticamente óptimo ha de verificar las siguientes propiedades:

1. Insesgado. Un estimador es insesgado si su esperanza coincide con el verdadero valor (desconocido) del parámtero, es decir: $E\left[\varphi(x_1, x_2, ..., x_n)\right] = \theta$. En caso de ser $E[\varphi(x_1, x_2, ..., x_n)] = \vartheta$, la diferencia $\theta - \vartheta$ se denomina **sesgo** de estimación.

2. Eficiente. Dados dos estimadores de un parámetro θ, el más eficiente será el que tenga menor varianza. La mínima varianza que puede tener un estimador insesgado θ se denomina *cota de Frechet-Cramer-Rao* y viene dada por $1/I_n(\theta)$ donde:

$$I_n\left(\theta\right) = n \cdot E\left[\left(\frac{\partial \ln f(x/\theta)}{\partial \theta}\right)^2\right]$$

se denomina *cantidad de información de Fisher*.

3. Consistente. Un estimador es consistente si converge en probabilidad hacia el parámetro, es decir: $Prob(\mid \varphi(x_1, x_2, ..., x_n) - \theta \mid < \epsilon) \underset{n\to\infty}{\rightarrow} 1$, lo que significa que a medida que aumenta el tamaño muestral, el estimador proporciona valores más próximos al parámetro.

4. Suficiente. En términos sencillos, un estimador es suficiente si resume toda la información relevante que contiene la muestra. Se caracterizan por un resultado conocido como *teorema de Neyman-Fisher*.

EJEMPLO 1

Se quiere estimar el nivel de colesterol LDL en una población cuyo nivel medio (desconocido para nosotros) es de 95 mg/dL, para lo cual se utilizan 3 técnicas de determinación que aplicadas a sendas muestras de 5 individuos cada una proporcionan los siguientes resultados:

Técnica	Determinaciones	Valor medio	Desviación
1	86 109 102 87 91	95	9,01
2	85 89 82 87 82	85	2,76
3	92 92 99 95 97	95	2,76

A la vista de los resultados anteriores observamos que las estimaciones proporcionadas por las técnicas 2 y 3 son más eficientes que la de la técnica 1; sin embargo entre la 2 y la 3, la última da una estimación sin sesgo de la media poblacional, por lo que se concluye que la técnica 3 es la óptima.

EJEMPLO 2

Supongamos que una variable aleatoria X se distribuye según un modelo normal con función de densidad: $f(x/\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$. Para obtener la cota de Frechet-Cramer-Rao del estimador óptimo de la media μ se procede de la forma siguiente:

$$\ln f(x/\mu,\sigma) = -\ln\left(\sigma\sqrt{2\pi}\right) - \frac{(x-\mu)^2}{2\sigma^2} \rightarrow \frac{\partial \ln f(x/\mu,\sigma)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

Así:

$$E\left[\left(\frac{\partial \ln f(x/\mu,\sigma)}{\partial \mu}\right)^2\right] = \frac{E\left[(x-\mu)^2\right]}{\sigma^4}$$

Como $Var\left[X\right] = E\left[(X-\mu)^2\right] = \sigma^2$ resulta que $I_n\left(\mu\right) = n\frac{\sigma^2}{\sigma^4} = \frac{n}{\sigma^2}$, por lo que la mínima varianza dada por la cota de Frechet-Cramer-Rao es $\frac{\sigma^2}{n}$.

EJEMPLO 3

El estimador media aritmética $\bar{x_n}$ es insesgado para el parámetro μ del modelo de Gauss pues:

$$E[\bar{x_n}] = E\left[\tfrac{x_1+x_2,\ldots+x_n}{n}\right] = \tfrac{E[x_1]+E[x_2],\ldots+E[x_n]}{n} = \tfrac{\mu+\mu+\cdots+\mu}{n} = \tfrac{n\mu}{n} = \mu$$

Sin embargo la varianza muestral s_n^2 no lo es porque operando se tiene que:

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}\left(x_i-\bar{x_n}+\bar{x_n}-\mu\right)^2 =$$
$$= \sum_{i=1}^{n}\left(x_i-\bar{x_n}\right)^2 + \sum_{i=1}^{n}\left(\bar{x_n}-\mu\right)^2 + 2(\bar{x_n}-\mu)\sum_{i=1}^{n}\left(x_i-\bar{x_n}\right)$$

Dado que

$$\sum_{i=1}^{n}\left(\bar{x_n}-\mu\right)^2 = n\left(\bar{x_n}-\mu\right)^2$$

y que

$$\sum_{i=1}^{n}\left(x_i-\bar{x_n}\right) = \sum_{i=1}^{n}x_i - n\cdot\bar{x}_n = 0,$$

resulta:

$$\tfrac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \tfrac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x_n}\right)^2 + \left(\bar{x_n}-\mu\right)^2 = s_n^2 + \left(\bar{x_n}-\mu\right)^2 \tag{2.1}$$

y tomando esperanzas:

$$\tfrac{1}{n}\sum_{i=1}^{n}E[(x_i-\mu)^2] = E[s_n^2] + E\left[\left(\bar{x_n}-\mu\right)^2\right]$$

Como

$$E[(x_i-\mu)^2] = \sigma^2 \text{ y } E\left[\left(\bar{x_n}-\mu\right)^2\right] = Var\left[\bar{x_n}\right] = \sigma^2/n,$$

al ser la cota de Frechet-Cramer-Rao, se concluye que:

$$\tfrac{1}{n}n\sigma^2 = E[s_n^2] + \tfrac{\sigma^2}{n} \rightarrow E[s_n^2] = \sigma^2 - \tfrac{\sigma^2}{n} = \tfrac{n-1}{n}\sigma^2$$

por lo que presenta un sesgo de σ^2/n. Para corregirlo se considera el estimador cuasivarianza muestral:

$$s_{n-1}^2 = \tfrac{n}{n-1}s_n^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x_n}\right)^2$$

que ya sí es insesgado pues:

$$E\left[s_{n-1}^2\right] = \tfrac{n}{n-1}E\left[s_n^2\right] = \tfrac{n}{n-1}\tfrac{n-1}{n}\sigma^2 = \sigma^2$$

2.2. ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

Existen diversos métodos estadísticos para obtener estimadores con propiedades relativamente buenas, tales como el método por analogía, el de los momentos, el de mínimos cuadrados, el de de Bayes, etc. En esta sección describiremos el denominado método de máxima verosimilitud pues, además de ser muy intuitivo su cálculo proporciona estimadores asintóticamente normales, insesgados y eficientes. El procedimiento consiste en lo siguiente: Sea X una variable aleatoria con función de probabilidad o de densidad $f(x/\theta)$ y consideremos una muestra aleatoria suya: $x_1, x_2, \ldots x_n$. Se define la verosimilitud de la muestra a la función:

$$L(x_1, x_2, \ldots x_n/\theta) = f(x_1/\theta) \cdot f(x_2/\theta) \cdot \cdots \cdot f(x_n/\theta)$$

(se denota con la letra L porque en inglés verosimilitud se dice *likelihood*). El método consiste en elegir como estimador $\hat{\theta}_n$aquel valor de θ que maximiza la función L. Como, en general, trabajar con la función L puede ser complicado, el procedimiento consiste en maximizar una función monótona creciente de L como es su logaritmo neperiano: $\ln L$:

$$\ln L = f(x_1/\theta) + f(x_2/\theta) + \cdots + f(x_n/\theta) = \sum_{i=1}^{n} f(x_i/\theta)$$

Para ello hay que resolver la denominada ecuación de verosimilitud y comprobar que verifica la condición de máximo:

$$\frac{\partial \ln L}{\partial \theta} = 0 \quad \frac{\partial^2 \ln L}{\partial \theta^2} < 0$$

EJEMPLO 1

El estimador máximo versosímil del parámetro media μ de una ley normal se obtiene de la forma siguiente:

$$\ln f(x/\mu, \sigma) = -\ln\left(\sigma\sqrt{2\pi}\right) - \frac{(x-\mu)^2}{2\sigma^2} \rightarrow \ln L = -n \ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Derivando respecto a μ e igualando a 0 se obtiene la siguiente ecuación de verosimilitud:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \rightarrow \sum_{i=1}^{n} x_i - n\mu = 0 \rightarrow \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$$

Derivando por segunda vez respecto de μ resulta:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$$

lo que permite concluir que la solución de máximo es la media aritmética muestral $\bar{x}_n$.

EJEMPLO 2

El estimador máximo versosímil del parámetro media σ de una ley normal se obtiene de la forma siguiente:

$$\ln L = -n \ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \rightarrow \frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} =$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

Igualando a 0 y sustituyendo μ por su estimador máximo verosímil se obtiene:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right)^2 = s_n^2$$

es decir, resulta la varianza muestral. Sin embargo, como se vio en la sección 2.1, este estimador no es insesgado, aunque asintóticamente sí, por lo que se sustituye por la cuasivarianza muestral:

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right)^2$$

2.3. ESTIMACIÓN SOBRE EL MODELO DE GAUSS

Aplicando el método de máxima verosimilitud y las propiedades de los estimadores, hemos visto que los estimadores óptimos de los parámetros de un modelo de Gauss $N(\mu, \sigma)$ son respectivamente la media muestral $\bar{x}_n$y la cuasivarianza muestral s^2_{n-1}. Para poder hace posterior inferencia con ellos es necesario conocer su distribución de probabilidad como variables aleatorias que son. El resultado fundamental es el siguiente:

Teorema de Fisher

Sea X una variable normal con parámetros μ y σ y $x_1, x_2, \ldots x_n$ una muestra aleatoria suya. Entonces:

a) La media muestral $\bar{x_n}$ se distribuye según un modelo normal de media μ y desviación típica $\frac{\sigma}{\sqrt{n}}$.

b) La cuasivarianza muestral verifica que $\frac{(n-1)s^2_{n-1}}{\sigma^2}$ se distribuye según un modelo *chi-cuadrado* con n-1 grados de libertad χ^2_{n-1}

c) Los estimadores $\bar{x_n}$ y s^2_{n-1} son variables aleatorias independientes.

Demostración

Para demostrar el apartado a) basta tener en cuenta que la suma de variables normales independientes es también una variable normal, y que $\bar{x_n}$ es un estimador insesgado de μ como se demostró en el apartado 2.1. Además, se verifica que:

$$Var\left[\bar{x_n}\right] = Var\left[\tfrac{1}{n}\sum_{i=1}^{n} x_i\right] = \tfrac{1}{n^2}\sum_{i=1}^{n} Var[x_i] = \tfrac{1}{n^2} n\sigma^2 = \tfrac{\sigma^2}{n}$$

por lo que la desviación típica del estimador $\bar{x_n}$ es $\frac{\sigma}{\sqrt{n}}$.

En cuanto al apartado b), operando sobre la expresión (2.1) resulta:

$$\sum_{i=1}^{n} \left(\tfrac{x_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\tfrac{x_i-\bar{x_n}}{\sigma}\right)^2 + \left(\tfrac{\bar{x_n}-\mu}{\sigma/\sqrt{n}}\right)^2$$

Dado que $x_i \rightsquigarrow N(\mu, \sigma)$ y que $\bar{x_n} \rightsquigarrow N(\mu, \sigma/\sqrt{n})$ se tiene que:

$$\tfrac{x_i-\mu}{\sigma} \rightsquigarrow N(0,1) \rightarrow \sum_{i=1}^{n} \left(\tfrac{x_i-\mu}{\sigma}\right)^2 \rightsquigarrow \chi^2_n \quad \text{y} \quad \tfrac{\bar{x_n}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1) \rightarrow \left(\tfrac{\bar{x_n}-\mu}{\sigma/\sqrt{n}}\right)^2 \rightsquigarrow \chi^2_1$$

lo que implica que $\sum_{i=1}^{n}\left(\frac{x_i-\bar{x_n}}{\sigma}\right)^2$ se distribuye según un modelo χ^2_{n-1} y como:

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

se concluye.

Para la demostración del apartado c) es necesario probar que la función de distribución conjunta de $\bar{x_n}$ y s^2_{n-1} puede factorizarse como producto de las funciones de distribución de cada uno de los dos estimadores, lo que queda fuera del alcance de los objetivos de este curso.

La interpretación del apartado a) es muy intuitiva, pues indica que si una variable tiene una dispersión dada por σ, el resultado de calcular medias de dicha variable con muestras de tamaño n hace reducir la dispersión a $\sigma/\sqrt{n}$. Éste es el fundamento de la técnica denominada *medias móviles* en la cual se sustituye la serie original de datos por otra formada por medias de tamaño 3,4,5,... la cual es más suave (con menos picos) que la original.

Corolario

Consideremos una variable con distribución binomial de parámetro p (proporción) y otra con distribución de Poisson con parámetro λ (promedio de ocurrencias). Si el número de muestras de las variables es grande (en ocasiones se considera que basta con que sea n>30), entonces:

a) El estimador óptimo de p es la frecuencia relativa $\hat{p}_n$ y su distribución es aproximadamente $N(p;\sqrt{\frac{p(1-p)}{n}})$

b) El estimador óptimo de λ es la media aritmética de la muestra $\bar{x}_n$ y su distribución es aproximadamente $N(\lambda;\sqrt{\frac{\lambda}{n}})$.

Demostración

Es consecuencia inmediata de la aplicación de las aproximaciones de los modelos binomial y Poisson a la ley normal cunado el tamaño muestral es grande.

2.4. ESTIMACIÓN POR INTERVALO

La estimación puntual permite obtener, para una muestra dada, un valor aproximado del parámetro de la distribución que se quier estimar, pero no proporciona información alguna sobre la precisión del mismo. Así, dar como resultado que una estimación del nivel de colesterol HDL en una población es de 47,3 mg/dL no es de gran utilidad si al realizar un estudio experimental con una muestra de igual tamaño de esa población da como resultado 53,2 mg/dL. Con objeto de solventar este inconveniente de la estimación puntual se introduce la estimación por intervalo que consiste en que dada una variable aleatoria X con función de probabilidad o de densidad $f(x/\theta)$, y una muestra aleatoria suya de tamaño n: $x_1, x_2, ..., x_n$, se construyen dos estimadores, que denotaremos $\varphi_1(x_1, x_2, ..., x_n)$ y $\varphi_2(x_1, x_2, ..., x_n)$, de forma que para toda muestra particular de tamaño n de la variable: $x_1 = a_1, x_2 = a_2, ..., x_n = a_n$, si denotamos

$$\overset{\wedge \inf}{\theta}_n = \varphi_1(a_1, a_2, ..., a_n) \quad , \quad \overset{\wedge \sup}{\theta}_n = \varphi_2(a_1, a_2, ..., a_n),$$

se verifica que $\overset{\wedge \inf}{\theta}_n < \overset{\wedge \sup}{\theta}_n$ y, además, la probabilidad de que el verdadero valor (desconocido) del parámetro θ se encuentre entre ambas estimaciores es conocida y prefijada por el investigador, y se denomina nivel de confianza:

$$\Pr ob(\overset{\wedge \inf}{\theta}_n \leq \theta \leq \overset{\wedge \sup}{\theta}_n) = 1 - \alpha.$$

El intervalo formado por ambos estimadores para cada muestra particular $\left[\overset{\wedge \inf}{\theta}_n , \overset{\wedge \sup}{\theta}_n\right]$ se denomina *intervalo de confianza* de nivel $(1 - \alpha)100\%$.

En Bioestadística es usual tomar como nivel de confianza estándar el 95%, que corresponde a un $\alpha = 0,05$, aunque también se consideran niveles del 90% ($\alpha = 0,10$) y del 99% ($\alpha = 0,01$). Es importante tener en cuenta que los límites de un intervalo de confianza son estimadores y, por tanto, para cada muestra proporcionan valores diferentes.

Mientras mayor es el nivel de confianza mayor será la amplitud del intervalo. Por el contrario, mientras mayor sea el tamaño de la muestra menor será su amplitud.

Existen diversos métodos de obtención de estimadores y a continuación veremos la forma de deducirlos sobre el modelo normal.

2.4.1. Intervalo para la media μ de una variable normal con σ conocida

Aunque esta situación es bastante irreal, dado que en una población cuya media es desconocida también lo será su desviación típica, el desarrollo de cálculo es sencillo e ilustrativo. Para ello nos basaremos en el teorema de Fisher que establece en sun apartado a) que si $X \rightsquigarrow N(\mu, \sigma)$ entonces la media muestral se distribuye de la forma $\bar{x}_n \rightsquigarrow N(\mu, \sigma/\sqrt{n})$, por lo que

$$\frac{\bar{x}_n - \mu}{\sigma}\sqrt{n} \rightsquigarrow N(0,1) \tag{2.2}$$

En consecuencia, si denotamos $z_{\alpha/2}$ al valor de la variable normal tipificada que deja a su derecha un área igual a $\alpha/2$, se verificará:

$$\Pr ob(-z_{\alpha/2} \leq \frac{\bar{x}_n - \mu}{\sigma}\sqrt{n} \leq z_{\alpha/2}) = 1 - \alpha$$

lo que equivale a escribir, tras operar en las desigualdades:

$$\Pr ob(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

por lo que el intervalo de confianza para μ es:

$$I_\alpha(\mu) = \left[\bar{x}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Los valores de $z_{\alpha/2}$ para los niveles de confianza más usuales son los siguientes:

$Nivel$	α	$z_{\alpha/2}$
90%	0,10	1,645
95%	0,05	1,960
99%	0,01	5,576

EJEMPLO

Estimar mediante intervalos de confianza del 90%, 95% y 99% la media del índice de masa corporal de una población geriátrica suponiendo que el valor de la desviación típica poblkacional es $\sigma = 2,83 kg/m^2$, si para una muestra de 50 individuos de dicha población se ha obtenido un valor medio de $17,2 kg/m^2$.

La fórmula del intervalo para cualquier nivel de confianza será en este caso

$$I_\alpha(\mu) = \left[17,2 \pm z_{\alpha/2}\frac{2,83}{\sqrt{50}}\right]$$

que, para cada uno de los niveles del enunciado quedará de la forma:

$$\text{Nivel: } 90\% \rightarrow I_{0,10}(\mu) = \left[17,2 \pm 1,645\frac{2,83}{\sqrt{50}}\right] = [17,2 \pm 0,66] = [16,54; 17,86]$$
$$\text{Nivel: } 95\% \rightarrow I_{0,05}(\mu) = \left[17,2 \pm 1,960\frac{2,83}{\sqrt{50}}\right] = [17,2 \pm 0,78] = [16,42; 17,98]$$
$$\text{Nivel: } 99\% \rightarrow I_{0,01}(\mu) = \left[17,2 \pm 2,576\frac{2,83}{\sqrt{50}}\right] = [17,2 \pm 1,03] = [16,17; 18,23]$$

2.4.2. Intervalo para la media μ de una variable normal con σ desconocida

Nos encontramos ante la situción más usual en los estudios prácticos, que consiste en analizar una variable normal $N(\mu, \sigma)$ cuyos parámetros μ y σ son desconocidos. En esta situación la fórmula del intervalo para μ se modifica ligeramente como a continuación se indica. Por el apartado a) del teorema de Fisher tenemos en primer lugar la expresión (2.2) antes citada:

$$\frac{\bar{x}_n - \mu}{\sigma}\sqrt{n} \rightsquigarrow N(0,1).$$

Por otra parte, el apartado b) de dicho teorema establece que:

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

En consecuencia, por definición de la distribución t de *Student* se verifica que:

$$\frac{\frac{\bar{x}_n-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{\frac{(n-1)s^2_{n-1}}{\sigma^2}}{n-1}}} = \frac{\bar{x}_n-\mu}{s_{n-1}}\sqrt{n} \rightsquigarrow t_{n-1},$$

por lo que procediendo de forma análoga al subapartado 2.4.1 se obtiene:

$$\Pr ob(\bar{x}_n - t^{\alpha/2}_{n-1}\frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{x}_n + t^{\alpha/2}_{n-1}\frac{s_{n-1}}{\sqrt{n}}) = 1-\alpha$$

siendo $t^{\alpha/2}_{n-1}$ el valor de la variable t de *Student* con $n-1$ grados de libertad que deja a su derecha un área igual a $\alpha/2$, y así el intervalo de confianza para μ es:

$$I_\alpha(\mu) = \left[\bar{x}_n \pm t^{\alpha/2}_{n-1}\frac{s_{n-1}}{\sqrt{n}}\right]$$

EJEMPLO

Para probar la eficacia de un cierto antipirético infantil se seleccionó de forma aleatoria un grupo de 10 niños de edad similar que padecían gripe, y se les midió la temperatura orporal en estado basal y dos horas después de administrarles el medicamento observándose las siguientes reducciones de temperatura: 1,2 1,0 1,7 2,0 1,6 1,7 1,0 2,6 1,0 2,0. Suponiendo que la variable que representa el descenso de temperatura en esas dos horas se distribuye según un modelo normal, vamos a estimar la reducción media de temperatura mediante un intervalo de confianza del 99%.

A partir de los datos muestrales se obtiene que: $\bar{x}_{10} = 1,58$ y $s_9 = 0,535$, por lo que el intervalo será:

$$I_{0,01}(\mu) = \left[1,58 \pm 3,25\frac{0,535}{\sqrt{10}}\right] = [1,58 \pm 0,55] = [1,03; 2,13].$$

Es decir, existe una probabilidad de 0,99 de que la reducción media de temperatura se encuentre entre esos límites es. De tal forma, si el laboratorioi que produce el medicamento estableciera en su prospecto que la reducción media de temperatura corporal tras dos horas desde que se consume es de 2^oC, estaría realizando una afirmación correcta con una probabilidad de error del 1%.

2.4.3. Intervalo para la desviación típica σ de una variable normal

Para deducir la expresión de este intervalo aplicaremos nuevamente el apartado b) del teorema de Fisher de manera que, si denotamos $\chi^{2(\text{inf})}_{n-1}$ y $\chi^{2(\text{sup})}_{n-1}$ a los valores de la variable *chi-cuadrado* con $n-1$ grados de libertad que dejan respectivamentea su izquierda y derecha sendas *colas* de área $\alpha/2$, entonces:

$$\Pr ob(\chi^{2(\text{inf})}_{n-1} \leq \tfrac{(n-1)s^2_{n-1}}{\sigma^2} \leq \chi^{2(\text{sup})}_{n-1}) = 1-\alpha$$

y operando en la cadena de desigualdades resulta:

$$\Pr ob(\tfrac{(n-1)s^2_{n-1}}{\chi^{2(\text{sup})}_{n-1}} \leq \sigma^2 \leq \tfrac{(n-1)s^2_{n-1}}{\chi^{2(\text{inf})}_{n-1}}) = 1-\alpha$$

por lo que el intervalo de confianza para la varianza σ^2 es:

$$I_\alpha(\sigma^2) = \left[\tfrac{(n-1)s^2_{n-1}}{\chi^{2(\text{sup})}_{n-1}}; \tfrac{(n-1)s^2_{n-1}}{\chi^{2(\text{inf})}_{n-1}}\right]$$

y, consiguientemente, el de la desviación σ típica es:

$$I_\alpha(\sigma) = \left[\sqrt{\tfrac{(n-1)s^2_{n-1}}{\chi^{2(\text{sup})}_{n-1}}}; \sqrt{\tfrac{(n-1)s^2_{n-1}}{\chi^{2(\text{inf})}_{n-1}}}\right] = \left[s_{n-1}\sqrt{\tfrac{(n-1)}{\chi^{2(\text{sup})}_{n-1}}}; s_{n-1}\sqrt{\tfrac{(n-1)}{\chi^{2(\text{inf})}_{n-1}}}\right]$$

EJEMPLO

Sobre el mismo enunciado del ejemplo anterior (subapartado 2.4.2) la estimación de la desviación típica σ de la variable reducción de temperatura mediante un intervalo de confianza del 99% será:

$$I_{0,01}(\sigma) = \left[0,535\sqrt{\tfrac{9}{23,59}}; 0,535\sqrt{\tfrac{9}{1,735}}\right] = [0,33; 1,22]$$

2.4.4. Intervalo para el parámetro p de una variable binomial y λ de una Poisson con muestras grandes

Aplicando el corolario del teorema de Fisher se obtienen como intervalos de confianza para ambos parámetros en el caso de muestras grandes los siguientes:

$$I_\alpha(p) = \left[\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right] \quad ; \quad I_\alpha(\lambda) = \left[\bar{x}_n \pm z_{\alpha/2}\sqrt{\frac{\bar{x}_n}{n}}\right]$$

siendo $z_{\alpha/2}$ el valor de la variable normal tipificada que deja a su derecha un área igual a $\alpha/2$

EJEMPLO 1

En una farmacia se ha observado que de un total de 220 productos dispensados durante el último mes 77 eran de parafarmacia y se quiere estimar mediante un intervalo de confianza del 95% la proporción de productos ded parafarmacia vendidos.

$$\hat{p}_{220} = \frac{77}{220} = 0,35 \rightarrow I_{0,05}(p) = \left[0,35 \pm 1,96\sqrt{\frac{0,35 \cdot 0,65}{220}}\right] = [0,287; 0,413]\,.$$

Es decir, con una probabilidad de 0,95 se venden mensualmente entre un 28,7% y un 41,3% de productos de parafarmacia.

EJEMPLO 2

En una farmacia se ha realizado un seguimiento anual de las ventas de un nuevo analgésico, observándose un promedio de ventas semanales de 37,6 unidades, y el farmacéutico se pregunta si puede admitirse a con un nivel de confianza del 99% si el promedio de ventas semanales alcanza las 40 unidades. Para ello, considrando 52 semanas al año, se construye el intervalo al 99%:

$$I_{0,01}(\lambda) = \left[37,6 \pm 2,576\sqrt{\frac{37,6}{52}}\right] = [37,6 \pm 2,2] = [35,4; 39,8]$$

y se concluye que no puede aceptarse a dicho nivel que el promedio de ventas semanales alcance los 40 ya que este valor está fuera del intervalo de confianza.

2.5. CÁLCULO DEL TAMAÑO MUESTRAL

Una de las cuestiones de mayor interés a la hora de realizar un estudio experimental consiste en determinr cuántas unidades muestrales es necesarion tomar para que las conclusiones sean fiables, concretamente para que la estimación $\hat{\theta}_n$del parámetro θ de la variable en cuestión sea suficientemente válido. Para ello es necesario fijar de antemano dos valores:

1. El riesgo de error aleatorio α del estudio o, equivalentemente, el nivel de confianza $1-\alpha$.
2. La precisión del estudio o error máximo admisible, que puede expresarse en términos absolutos $E = \left|\theta - \hat{\theta}_n\right|$ o relativos $E = \theta/\hat{\theta}_n$.

En este apartado nos limitaremos a deucir las expresiones del tamaño muestral para los parámetros de un modelo normal. Así, a partir de las fórmulas de los intervalos de confianza, en el caso de la media μ consideraremos $E = \left|\mu - \bar{x}_n\right|$ con lo cual:

$$E = t_{n-1}^{\alpha/2}\frac{s_{n-1}}{\sqrt{n}} \rightarrow n = \left(t_{n-1}^{\alpha/2}\frac{s_{n-1}}{E}\right)^2$$

donde n se aproxima por exceso al número entero inmediatamente superior.

En el caso de estimar σ consideraremos $E = \sigma^2/s_{n-1}^2$, con lo cual:

$$\frac{(n-1)}{\chi_{n-1}^{2(\text{sup})}} \leq E \leq \frac{(n-1)}{\chi_{n-1}^{2(\text{inf})}} \rightarrow n = \chi_{n-1}^{2(\text{sup})} \cdot E + 1$$

donde n se aproxima igualmente por exceso al número entero inmediatamente superior.

En la práctica lo recomendable a la hora de calcular el tamaño muestral es tomar una muestra piloto pequeña y, a partir de ella, calcular la cuasivarianza y seguir tomando muestras hasta alcanzar el número necesario según la fórmula. Es conveniente, no obstante, sobre todo si dicho número es muy alto, realizar cálculos intermedios a medida que se vayan tomando nuevas muestras.

EJEMPLO

Se quiere calcular el tamaño muestral necesario para estimar al 95% el nivel medio de colesterol total en los individuos de una población con un error absoluto máximo de $10mg/dL$, para lo cual se toma una primera muestra de 10 individuos de la misma resultando una cuasidesviación típica de $35,7mg/dL$. En la práctica es usual considerar como aproximación al nivel del 95%: $t_{n:-1}^{\alpha/2} \simeq 2$, y así:

$$n = (2\frac{35,7}{10})^2 = 50,98 \simeq 51$$

es decir, sería necesario considerar 51 individuos a lo que tomarles una muestra de sangre para medir el colesterol.

RELACIÓN DE PROBLEMAS

1^o) Para estudiar la permanencia de enfermos en un gran hospital se elige aleatoriamente una muestra de 300 enfermos encamados y se anota el número de días de hospitalización, obteniendo una estancia media muestral de 8,25 días y una cuasi desviación típica muestral de 5,7 días. Sabiendo que la variable aleatoria que representa el número de días de hospitallización se ajusta a un modelo de Gauss, y considerando un nivel de confianza del 95%:
a) Construir un intervalo de confianza para la permanencia media.
b) ¿Puede aceptarse que la estancia media de un paciente en el hospital es de 7 días?

Solución: a) $I_{0,05}(\mu)$=[7,6; 8,9] b) No puede aceptarse al 95%

2^o) Una marca de alimentos infantiles especifica que un cierto producto de la marca contiene $42g$ de proteinas por cada $100g$ de producto. Con objeto de comprobar esta especificación se analizaron 10 muestras al azar de dicho producto, resultando que el contenido medio de proteinas era de $40g$ y la correspondiente cuasi desviación típica $3,5g$ ambos referidos a $100g$ de producto. Si suponemos que la variable objeto de estudio se distribuye según una ley de Gauss:
a) Construir intervalos de confianza del 95% para el peso medio y para la desviación típica.
b) ¿Puede aceptarse a dicho nivel una desviación típica de $6g$?

Solución: a) $I_{0,05}(\mu)$=[37,36; 42,64] b) $I_{0,05}(\sigma)$=[2,54; 6,74]

3^o) Se ha comprobado que un nuevo fármaco es eficaz en 80 pacientes que presentan una cierta patología de una muestra de 200. Estimar un intervalo de confianza del 99% para la proporción de pacientes de una población afectada de dicha patología para los que el fármaco es eficaz.

Solución: $I_{0,01}(p)$=[0,311; 0,489]

4^o) El número medio de pacientes que acuden a la consulta de un especialista durante los 225 días de trabajo anual ha sido de 8,24. Sabiendo que la variable aleatoria que representa el número de pacientes que van diariamente a las consulta se ajusta a un modelo de Poisson de parámetro λ, obtener un intervalo de confianza del 90% para λ.

Solución: $I_{0,10}(\lambda)$=[7,925; 8,555]

5^o) Para determinar el contenido de principio activo en un antibiótico que se comercializa en forma de suspensión se elige aleatoriamente una muestra de 25 sobres y se mide dicho contenido, resultando un valor medio de $490mg$ y una cuasi-desviación típica muestral (raíz cuadrada de la cuasivarianza) de $18mg$. Suponiendo que se la variable que representa el contenido en mg de principio activo en un sobre se distribuye de forma normal:
a) Calcular los intervalos de confianza al 95% para la media y la desviación típica de dicha variable
b) Calcular el tamaño muestral necesario para estimar el nivel medio de la variable con una tolerancia máxima de $2mg$ y con una de $1,5mg$.
c) En el etiquetado de dicho medicamento figura un contenido de $500mg$ de antibiótico por sobre ¿puede aceptarse ese valor como promedio al nivel de confianza establecido? ¿Qué se podría hacer para que fuese aceptable?

Solución: a) $I_{0,05}(\mu)$=[482,57;497,43] Solución: $I_{0,05}(\sigma)$=[14,06; 25,04]

b) Para E=$2mg \rightarrow n = 324$ y para E=$1,5mg \rightarrow n = 576$

c) No sería aceptable. Sería necesario aumentar el nivel de confianza, por ejemplo al 99%: $I_{0,01}(\mu)$=[479,93;500,07]

6^o) La glucemia basal (expresada en mg/dL) de una muestra de 8 pacientes diabéticos del tipo 2 es la siguiente: 114 130 95 184 100 115 145 122. Suponiendo que dicha variable aleatoria sigue una distribución normal, obtener:
a) Intervalos de confianza del 95% para la media y la desviación típica de la variable.
b) Determinar el tamaño muestral necesario para estimar la media con un nivel máximo de tolerancia de $5mg/dL$.

Solución: a) $I_{0,05}(\mu)$=[101,85;149,40] $I_{0,05}(\sigma)$=[18,80; 57,86] b) $n = 181$

Tema 3. CONTRASTE DE HIPÓTESIS

3.1. PLANTEAMIENTO DE UN CONTRASTE

En lenguaje común, una hipótesis es sinónimo de suposición o conjetura; de tal forma, una hipótesis estadística será una suposición sobre el valor de un parámetro de una variable aleatoria o, incluso, sobre la propia distribución de probabilidad de dicha variable. Por supuesto, la hipótesis que se plantea puede ser verdadera o falsa; así, un contraste o test de hipótesis es un procedimiento para concluir si la hipótesis debe aceptarse o rechazarse. En consecuencia, puede ocurrir que el contraste lleve a la conclusión de aceptar o rechazar una hipótesis que es verdadera, e igualmente si es falsa, por lo que se pueden presentar una de estas cuatro situaciones:

Decisión	Aceptar	Rechazar
Hipótesis verdadera	Decisión correcta	Error tipo I
Hipótesis falsa	Error tipo II	Decisión correcta

La probabilidad de cometer un erro tipo I se denota α y se denomina *nivel de significación*, y la probabilidad de cometer un error tipo II se denota β y su complementaria $1-\beta$ se denomina *potencia* del contraste; es decir, la potencia es la probabilidad de rechazar la hipótesis nula cuando es falsa, por lo que un buen contraste deberá ser potente.

Observemos que en la vida, cuando uno es joven rechaza muchas cosas, como tener una pareja estable e hijos, para poder disfrutarla a tope e irse de guateque; comete de esta forma un error tipo I. Por el contrario, cuando se es viejo, se acepta por temor cualquier cosa, incluso algunas que deberían ser inaceptables, cometiendo un error tipo II.

A la hora de formular correctamente un contraste no basta con enunciar la hipótesis sobre la que se quiere decidir su aceptación o rechazo, que se denomina *hipótesis nula* y se representa H_0, sino que es necesario enunciar también una alternativa H_1. Así, si consideramos una variable X con función de probabilidad o de densidad $f(x/\theta)$, sobre su parámetro θ podrían formularse los siguientes contrastes:

$$\begin{array}{cccc} H_0 : \theta = \theta_0 & H_0 : \theta \leq \theta_0 & H_0 : \theta \geq \theta_0 & H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 & H_1 : \theta > \theta_0 & H_1 : \theta < \theta_0 & H_1 : \theta = \theta_1 \end{array}$$

El último de los cuatro contrastes anteriores se dice que es de hipótesis nula simple frente a alternativa simple, y su resolución queda fuera del alcance de este curso. Los tres primeros son de hipótesis nula simple frente a alternativa compuesta, pues incluso en el caso segundo y tercero, la hipótesis nula incluye la igualdad, por lo que es equivalente a enunciar $H_0 : \theta = \theta_0$. A su vez el primero es de tipo bilateral, pues la alternativa a la igualdad es que el parámetro sea tanto mayor como menor al valor supuesto, mientras que el segundo y tercero son de tipo unilateral.

Una vez formuladas las hipótesis nula y alternativa, para resolver el contraste, el procedimiento consiste en construir una función de una muestra genérica de la variable, que se denominará *estadístico del contraste*, y que es, igual que los estimadores, una variable aleatoria con una cierta distribución de probabilidad, de manera que para cada muestra particular tomará un valor numérico M. Entonces, a partir de la distribución del estadístico del contraste, se obtienen uno o dos valores que dividen la recta real en dos partes, una se denomina *región de aceptación*, pues si el valor M está dentro de ella se aceptará la hipótesis nula, y el dominio complementario se denomina *región de rechazo*, pues representa el conjunto de valores M para los cuales se rechaza la hipótesis nula.

Es importante tener presente que lo que se somete a contraste es la hipótesis nula y, por tanto, la conclusión debe ser si se acepta o rechaza. Por supuesto el rechazo conlleva la aceptación implícita de la hipótesis alternativa y, en ocasiones, se dice por abuso de lenguaje que aceptamos H_1. Un ejemplo de buen entendimiento de la metodología de contraste de hipótesis es la que sigue el sistema judicial de los EE.UU., pues cuando un acusado es sometido a juicio es porque el juez ha considerado que hay indicios de culpabilidad y ésta es precisamente la hipótesis nula H_0 : *culpable*. A lo largo

del proceso el jurado debe decidir su veredicto, y si no encuentra suficientes argumentos para probar su culpabilidad concluye H_1 : *no culpable*, lo que no significa que el acusado sea inocente, sino tan solo que no puede probar que no lo sea.

También hay que tener presente que, como su propio nombre indica, la hipótesis nula, tanto de los contrastes unilaterales como bilaterales, incluye la igualdad, es decir, su significado es que no hay diferencias entre el valor del parámetro y el que se conjetura. Por tanto, si se intenta probar la existencia de una diferencia significativa, ésta hay que enunciarla dentro de la hipótesis alternativa que es la que recoge las diferencias (distinto, mayor o menor).

A los contrastes también se les denomina *test* de hipótesis o *prueba* de hipótesis, incluso en algunos textos clásicos, normalmente traducciones realizadas en Hispanoamérica, se les denominaban *docimasia* de las verosimilitudes.

3.2. CONTRASTES DE NORMALIDAD

El primer paso, por tanto, para proceder al análisis paramétrico que tratamos en este tema consiste en contrastar la hipótesis de normalidad de las variables implicadas::

$$H_0 : X \rightsquigarrow N(\mu; \sigma)$$
$$H_1 : X \not\rightsquigarrow N(\mu; \sigma)$$

Para ello existen diversos contrastes de tipo no paramétrico, entre las que cabe citar la prueba de Shapiro-Wlks o la de Anderson-Darling, además del test de Kolmogorov-Smirnov que se desarrollará enn el Tema 6. Uno de los más fáciles de implementar es la prueba de **D'Agostino** que describiremos a continuación .

Consideremos una muestra ordenada: $x_1 < x_2 < \cdots < x_n$ cuya media es $\bar{x}$ y la desviación típica s_n. El estadístico de D'Agostino se construye de la forma siguiente:

$$D = \frac{\sum_{i=1}^{n} i x_i - \frac{n(n+1)}{2}\bar{x}}{n^2 s_n}$$

Si su valor está dentro del intervalo de aceptación de la Tabla *** del apéndice se aceptará la hipótesis H_0 de normalidad de la variable.

EJEMPLO

3.3. CONTRASTE SOBRE UNA VARIABLE NORMAL

Sea X una variable normal $N(\mu, \sigma)$ y a partir de una muestra aleatoria suya denotemos $\bar{x}_n$ a su media y s_{n-1} a su cuasi-desviación típica. Vamos a ver cómo se plantean y resuelven los contrastes sobre μ y σ respectivamente.

3.3.1. Contrastes sobre μ

En este caso pueden plantearse los tres siguientes:

$$\begin{array}{lll} H_0 : \mu = \mu_0 & H_0 : \mu \leq \mu_0 & H_0 : \mu \geq \mu_0 \\ H_1 : \mu \neq \mu_0 & H_1 : \mu > \mu_0 & H_1 : \mu < \mu_0 \end{array}$$

El estadístico del contraste viene dado por:

$$t_{\text{exp}} = \frac{\bar{x}_n - \mu_0}{s_{n-1}} \sqrt{n}$$

que se distribuye según una ley t de Student con $n-1$ grados de libertad. Así, las regiones de aceptación para los tres contrastes fomulados son, respectivamente:

$$\left[-t_{n-1}^{\alpha/2}; t_{n-1}^{\alpha/2}\right] \quad (-\infty; t_{n-1}^{\alpha}] \quad [-t_{n-1}^{\alpha}; \infty)$$

3.3.2. Contrastes sobre σ

En este caso pueden plantearse los tres siguientes:

$$\begin{array}{lll} H_0 : \sigma = \sigma_0 & H_0 : \sigma \leq \sigma_0 & H_0 : \sigma \geq \sigma_0 \\ H_1 : \sigma \neq \sigma_0 & H_1 : \sigma > \sigma_0 & H_1 : \sigma < \sigma_0 \end{array}$$

El estadístico del contraste viene dado por:

$$\chi^2_{\text{exp}} = \frac{(n-1)s^2_{n-1}}{\sigma_0^2}$$

que se distribuye según una ley χ^2 de Student con $n-1$ grados de libertad. Así, las regiones de aceptación para los tres contrastes fomulados son, respectivamente:

$$\left[\chi^{2\,(1-\alpha/2)}_{n-1};\chi^{2\,(\alpha/2)}_{n-1}\right] \quad [0;\chi^{2\,(\alpha)}_{n-1}] \quad [\chi^{2\,(1-\alpha)}_{n-1};\infty)$$

EJEMPLO

Una tabacalera que fabrica cigarrillos con un contenido de alquitrán de 15 mg por cigarrillo, sobrepasando el límite de 10 mg que establece la legislación. Para adaptarse a la normativa introduce una modificación en el proceso productivo que reduce dicho contenido de alquitrán. Así, para comprobar que el procedimiento ha sido efectivo, toma al azar una muestra de 20 cigarrillos fabricados con el método modificado, a los que le mide el contenido de alquitrán, resultando en promedio 9,5 mg con una cuasi-desviación típica de 1,8 mg. Suponiendo que la variable que representa el contenido de alquitrán por cigarrillo se distribuye según una ley normal y considerando el nivel de significación estándar (α=0,05) vamos a realizar los siguientes análisis:

a) Contrastar si el procedimiento es efectivo, es decir, si la media de contenido de alquitrán con el nuevo procedimiento es inferior a 10 mg por cigarrillo

El contraste se formula de la forma siguiente:

$$H_0 : \mu \geq 10$$
$$H_1 : \mu < 10$$

El valor del estadístico de contraste es:

$$t_{\text{exp}} = \frac{9{,}5-10}{1{,}8}\sqrt{20} = -1,242$$

Como el valor crítico de la distribución t con 19 grados de libertad es $1,729$ la región de aceptación es $[-1,729;\infty)$, por lo que se acepta la hipótesis nula al estar el valor de dentro de dicho intervalo y concluimos que, al nivel estándar, no podemos considerar efectiva la modificación introducida en el método de producción de cigarrillos.

Es interesante saber interpretar este resultado, pues puede parecer contradictorio que si el valor medio de alquitrán en la muestra de cigarrillos es de 9,5 mg, no podamos considerar que la media global es inferior a 10 mg. Esto es debido a una de dos, o bien que la muestra no es lo suficientemente grande como para poder rechazar con una probabilidad del 95% la hipótesis nula (no hay suficiente evidencia para ello), o bien que la cuasidesviación típica es demasiado alta como para no incluir la posibilidad de que sea $\mu \geq 10$.

b) Si se obtuviera la misma media y cuasi-desviación muestral con una muestra de 50 cigarrillos ¿qué conclusión se obtendría?

En este caso el valor del estadístico de contraste sería:

$$t_{\text{exp}} = \frac{9{,}5-10}{1{,}8}\sqrt{50} = -1,964$$

y la región de aceptación $[-1,673;\infty)$, ya que $t_{49}^{(0,05)} = 1,673$, por lo que quedaría fuera del intervalo y, en este caso, sí se rechazaría la hipótesis nula concluyendo que $\mu < 10$ y, por tanto, la modificacióndel proceso productivo sí sería eficaz.

c) Si con una muestra de tamaño 20, como la inicial, se obtuviera un valor medio de 9,5 mg y una cuasi-desviación de 1 mg ¿qué conclusión se obtendría en este caso?

Ahora el valor del estadístico de contraste sería:

$$t_{\text{exp}} = \frac{9{,}5-10}{1}\sqrt{20} = -2,23$$

que también quedaría fuera del intervalo de aceptación, rechazándose la hipótesis nula y concluyendo, asimismo, que $\mu < 10$.

d) Contrastar si con los datos iniciales la desviación típica muestral es inferior a 2 mg

En este caso el contraste a plantear sería:

$$H_0 : \sigma \geq 2$$
$$H_1 : \sigma < 2$$

y el estadístico de contraste:

$$\chi^2_{\text{exp}} = \frac{(20-1)\cdot 1{,}8^2}{2^2} = 15,39$$

Como la región de aceptación es $[10,117;\infty)$, ya que el valor crítico de la distribución *chi-cuadrado* con 19 grados de libertad es $10,117$, concluimos que hay que aceptar H_0 y no podemos considerar una desviación típica poblacional inferior a 2 mg.

3.4. COMPARACIÓN DE DOS VARIABLES NORMALES

A la hora de comparar dos varibles normales, pueden presentarse las dos situaciones que a continuación vamos a desarrollar, en función de que sean pareadas o independientes

3.4.1. Caso de dos variables normales pareadas

Al hablar de variables pareadas queremos significar que representan sendas observaciones sobre los mismo individuos como, por ejemplo, temperatura corporal en estado basal y transcurridas dos horas desde la administración de un antipirético, o número de dioptrías del ojo derecho e izquierdo, pues ambas se miden sobre los mismos pacientes. En este caso, si $X_1 \rightsquigarrow N(\mu_1;\sigma_1)$ y $X_2 \rightsquigarrow N(\mu_2;\sigma_2)$, entonces la variable diferencia $X_D = X_1 - X_2$ se distribuye también según un modelo normal $N(\mu_D;\sigma_D)$ y el contraste de comparación entre las medias quedaría reducido a uno de los siguientes:

$$\begin{array}{lll} H_0 : \mu_D = 0 & H_0 : \mu_D \leq 0 & H_0 : \mu_D \geq 0 \\ H_1 : \mu_D \neq 0 & H_1 : \mu_D > 0 & H_1 : \mu_D < 0 \end{array}$$

3.4.2. Caso de dos variables normales independientes

Vamos ahora a considerar el caso de dos variables independientes distribuidas según el modelo de Gauss: $X_1 \rightsquigarrow N(\mu_1;\sigma_1)$ y $X_2 \rightsquigarrow N(\mu_2;\sigma_2)$. El primer paso consiste en tomar muestras de ambas variables y calcular sus medias y cuasi-desviaciones, que denotaremos

Muestra de X_1 de tamaño $n_1 : x_{11}\ x_{12}\ ...\ x_{1n_1} \rightarrow \bar{x}_1 \quad s_1$

Muestra de X_2 de tamaño $n_2 : x_{21}\ x_{22}\ ...\ x_{2n_2} \rightarrow \bar{x}_2 \quad s_2$

El procedimiento a seguir es el siguiente:

Paso 1^o. Contrastar la igualdad de desviaciones:

$$\begin{array}{l} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{array}$$

Para ello es necesario evaluar el estadístico de contraste:

$$F_{\text{exp}} = \frac{s_1^2}{s_2^2}$$

suponiendo que es $s_1^2 \geq s_2^2$, el cual se distribuye según una ley F de *Fisher-Snedecor* con $(n_1 - 1, n_2 - 1)$ grados de libertad. Por tanto, la región de aceptación viene dada por el intervalo:

$$\left[F^{(1-\alpha/2)}_{(n_1-1,n_2-1)}; F^{(\alpha/2)}_{(n_1-1,n_2-1)}\right]$$

donde $F^{(1-\alpha/2)}_{(n_1-1,n_2-1)}$ y $F^{(\alpha/2)}_{(n_1-1,n_2-1)}$ representan los valores de la distribución F que dejan a su izquiera y derecha respectivamente recintos de área $\alpha/2$. Utilizando la propiedad 4 de dicha distribución se tiene que

$$F^{(1-\alpha/2)}_{(n_1-1,n_2-1)} = \frac{1}{F^{(\alpha/2)}_{(n_2-1,n_1-1)}}$$

En función del resultado del contraste de igualdad de desviaciones, se procede al contraste sobre las medias.

Paso 2^o. Pueden plantearse los tres contrastes siguientes:

$$\begin{array}{lll} H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 \leq \mu_2 & H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 \neq \mu_2 & H_1 : \mu_1 > \mu_2 & H_1 : \mu_1 < \mu_2 \end{array}$$

cuya resolución depende del resultado del Paso 1^o. Así cabe distinguir:

a) Si las desviaciones son iguales ($\sigma_1 = \sigma_2$) el estadístico de contraste es:

$$t_{\text{exp}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde s_p^2 es una media ponderada de las cuasivarianzas:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

distribuyéndose el estadístico t_{exp} según una t con $n_1 + n_2 - 2$ grados de libertad, de forma que las regiones de aceptación para los tres contrastes formulados son respectivamente:

$$\left[-t^{(\alpha/2)}_{n_1+n_2-2}; t^{(\alpha/2)}_{n_1+n_2-2}\right] \qquad (-\infty; t^{(\alpha)}_{n_1+n_2-2}] \qquad [-t^{(\alpha)}_{n_1+n_2-2}; \infty)$$

b) Si las desviaciones no son iguales ($\sigma_1 \neq \sigma_2$) el estadístico de contraste es ahora:

$$t_{\exp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

el cual se distribuye según una t con f grados de libertad, donde f se conoce como *aproximación de Welch* y es el entero más próximo a:

$$f \simeq \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n2-1}}$$

Las regiones de aceptación son, entonces respectivamente:

$$\left[-t_f^{(\alpha/2)}; t_f^{(\alpha/2)}\right] \qquad (-\infty; t_f^{(\alpha)}] \qquad [-t_f^{(\alpha)}; \infty)$$

Respecto al Paso 1^o cabe indicar que pueden formularse de manera similar contrastes unilaterales sobre las desviaciones.

3.4.3. Contrastes asintóticos sobre los parámetros de los modelos binomial y Poisson

Aplicando el corolario del Teorema de Fisher enunciado en la sección 2.3, cuando el tamaño muestral n es grande, el estimador $\hat{p}_n$= frecuencia relativa, del parámetro p de un modelo binomial, y el estimador $\hat{\lambda}_n$= media muestral del parámetro λ de un modelo de Poisson se disribuyen aproximadamente según sendos modelos normales:

$$N(p; \sqrt{\tfrac{p(1-p)}{n}}) \quad \text{y} \quad N(\lambda; \sqrt{\tfrac{\lambda}{n}})$$

respectivamente. Por tanto, la inferencia será en estos casos similar a la desarrollada en las secciones anteriores utilizando las aproximaciones de normalidad.

3.5. SIGNIFICACIÓN DE UN CONTRASTE: EL VALOR P

Cuando el procedimiento de resolución de un contraste concluye la decisión de rechazar la hipótesis nula, se dice que el contraste es significativo. Esto ocurre, por ejemplo, cuando tras extraer una muestra de tamaño $n = 20$ de una variable normal se obtiene una media muestral $\bar{x}_{20}$=7,6 y una cuasidesviación s_{19}=1,2 y se formula el contraste:

$$H_0 : \mu \leq 7$$
$$H_1 : \mu > 7$$

El valor del estadístico de contraste es entonces $t_{\text{exp}} = 2,236$ y la región de aceptación $(-\infty; 1,729]$, por lo que se rechaza H_0 al nivel estándar $\alpha = 0,05$. Sin embargo, si como nivel de significación hubiéramos adoptado $\alpha = 0,01$, la región de aceptación sería ahora $(-\infty; 2,539]$ y la conclusión sería aceptar H_0. De hecho, hay un nivel intermedio entre $0,01$ y $0,05$ para el cual cambia el sentido del contraste que pasa de ser significativo a no serlo. En este caso ese nivel es $0,03754$, aunque con las tablas estadísticas no puede obtenerse de forma exacta, y es lo que se denomina *valor p*.

Podemos, pues, definir p como el mínimo valor de significación para el cual el contraste es significativo, es decir, rechaza la hipótesis nula. Esto es equivalente a decir que el valor p es el área (probabilidad) de la cola, en caso de contraste unilateral, o colas, en el caso bilateral, que deja el estadístico de contraste. Así, en Bioestadística ésta es la clave para decidir sobre la aceptación o rechazo de la hipótesis nula que se somete a contraste:

$$p \geq 0,05 \rightarrow \text{Aceptar } H_0 \text{ (contraste no significativo)}$$
$$p < 0,05 \rightarrow \text{Rechazar } H_0 \text{ (contraste significativo)}$$

RELACIÓN DE PROBLEMAS

1^o) Se quiere probar la eficacia de un tratamiento reductor de transaminasas (GOT) en pacientes que se encuentran en el límite saludable de 40 mU/mL. Así, se aplica a una muestra de 30 pacientes y, tras seguir el tratamiento durante un periodo de tiempo, se les midió el nivel de transaminasas resultando un promedio de 39 mU/mL con una cuasidesviación de 3 mU/mL. Suponiendo normal el nivel de transaminasas (GOT) y considerando

el nivel estándar α=0,05: a) ¿Puede admitirse una desviación típica igual a 2 mU/mL?; b) ¿es eficaz el tratamiento para redicir el nivel de transaminasas?; c) si con otra muestra de igual tamaño el promedio fuese de $39,1$ mU/mL y la cuasidesviación de 2 mU/mL ¿qué conclusión se obtendría?

Solución: a) χ^2_{exp}=65,25 Región de aceptación: [16,047;45,722]
Conclusión: Rechazar $H_0 \rightarrow$ se concluye que $\sigma \neq 2$ (p=0,0003)
b) t_{exp}=-1,826 Región de aceptación: [-1,699;∞)
Conclusión: Rechazar $H_0 \rightarrow$ se concluye que $\mu<40$ (p=0,0391)
c) t_{exp}=-1,643 Región de aceptación: [-1,699;∞)
Conclusión: Aceptar $H_0 \rightarrow$ se concluye que $\mu \geq 40$ (p=0,0556)

2^o) Para estudiar si los pacientes con hiperplasia benigna de próstata tienen una PSA superior a 4 ng/mL se eligió aleatoriamente una muestra de 10 pacientes y se realizó en ellos la determinación obteniendo los siguientes resultados: $1,4$ $3,6$ $2,2$ $3,1$ $2,2$ $3,0$ $3,2$ $1,3$ $4,6$ $3,2$. Suponiendo que la variable que representa el nivel de PSA se distribuye normalmente: a) ¿Puede aceptarse al nivel estándar una desviación superior a 1 ng/mL?; b) ¿qué puede concluirse del estudio?

Solución: a) χ^2_{exp}=9,256 Región de aceptación: [0;16,919]
Conclusión: Aceptar $H_0 \rightarrow$ se concluye que $\sigma \leq 1$ (p=0,4140)
b) t_{exp}=-3,804 Región de aceptación: [-1,833;∞)
Conclusión: Rechazar $H_0 \rightarrow$ se concluye que $\mu<4$ (p=0,0021)

3^o) Los datos obtenidos en el problema 2 se confrontan ahora con los de un grupo de pacientes con cáncer de próstata, en los cuales la determinación de la PSA ha dado los siguientes resultados: $4,1$ $2,7$ $6,5$ $5,0$ $4,4$ $2,1$ $5,2$ $8,8$ $5,9$ $7,0$ $1,8$. ¿Puede concluirse al nivel estándar que esta población de pacientes tiene una PSA superior a la de hiperplasia benigna?

Solución: Comparando σ_1 y σ_2 resulta F_{exp}=4,516 y la región de aceptación es [0,265;3,946] por lo que se concluye que $\sigma_1 \neq \sigma_2$ (p=0,0331). En cuanto a medias es t_{exp}=-2,785 y la región de aceptación es [-1,753;∞) ya que, por la aproximación de Welch f=14,51$\simeq$15 grados de libertad, por lo que se concluye que $\mu_1 < \mu_2$ (p=0,0059)

4^o) Se está realizando un estudio experimental sobre hipertensión arterial en dos grupos nutricionales: uno sigue una dieta hiperproteica y otro vegetariana, para lo cual se tomaron sendas muestras aleatorias obteniéndose los siguientes resultados:

Dieta hiperproteica: n_1=25→ $\bar{x}_1$=13,9 *mm Hg* s_1=2,4 *mm Hg*

Dieta vegetariana: n_2=28→ $\bar{x}_2$=12,5 *mm Hg* s_2=2,1 *mm Hg*

¿Qué conclusión puede deducirse del estudio al nivel estándar?

Solución: Comparando σ_1 y σ_2 resulta F_{exp}=1,306 y la región de aceptación es [0,447;2,195] por lo que se concluye que σ_1=σ_2 (p=0,4996). En cuanto a medias es t_{exp}=2,265 y la región de aceptación es (-∞;1,676], por lo que se concluye que $\mu_1 > \mu_2$ (p=0,0278)

5^o) Para comprobar la eficacia de un nuevo fármaco para reducir el colesterol se aplicó a una muestra de pacientes con hipercolesterolemia (colesterol total superior a 250 mg/dL) y se comparó con otra muestra del grupo control, obteniéndose los siguientes resultados:

Grupo control:	255	260	251	280	267	261	274	260
Grupo tratamiento	238	215	252	220	205	224	233	

¿Es eficaz el nuevo fármaco al nivel estándar?

Solución: Comparando σ_1 y σ_2 resulta F_{exp}=2,618 y la región de aceptación es [0,176;5,119] por lo que se concluye que σ_1=σ_2 (p=0,8831). En cuanto a medias es t_{exp}=5,564 y la región de aceptación es (-∞;1,771], por lo que se concluye que sí es eficaz el tratamiento (p=0,000045)

6^o) Se quiere comprobar la calidad de dos procedimientos de preparación de sobres de analgésico con un contenido de 250 mg de principio activo, para lo cual se toman muestras de sobres preparados por ambos métodos obteniendo los siguientes resultados:

Procedimiento 1	249	249	252	250	248	251	252	249	250	250
Procedimiento 2	252	244	240	251	257	259	241	262	249	245

¿Puede aceptarse al nivel estándar que ambos procedimientos son igual de precisos?

Solución: Se trata de comparar las desviaciones de ambos procedimientos mediante un contraste bilateral resultando F_{exp}=32,627 y la región de aceptación es [0,248;4,026] por lo que se concluye que $\sigma_1 \neq \sigma_2$ (p=0,000016)

Tema 4. DISEÑOS EXPERIMENTALES I: ANÁLISIS DE LA VARIANZA

4.1. DESCOMPOSICIÓN LINEAL DE LA VARIABILIDAD

Al repetir el mismo experimento en análogas condiciones observamos que, casi siempre, da resultados diferentes. Una explicación clara puede ser que determinados experimentadores o instrumentos son de mayor calidad que en otros, lo que da lugar a mejores resultados *ceteris paribus*, es decir en igualdad de las restantes condiciones que rodean el experimento. Esto puede ocurrir cuando se realiza un examen común en misma asignatura con varios grupos, en los que los alumnos se distribuyen de forma que todos tienen una composición similar en cuanto a capacidades; en tal caso, las diferencias entre las calificaciones medias de los grupos puede ser un indicador del nivel docente y científico de los profesores, pues si un grupo de composición similar a otro obtiene una nota media significativamente superior, puede achacarse a la calidad del profesor que ha impartido la materia. Sin embargo, en numerosas ocasiones subyace alguna causa secundaria que puede afectar el resultado del experimento, como ocurriría, en el ejemplo en cuestión, si un grupo se viese sistemáticamente afectado por problemas tales como cortes frecuentes de luz, convocatoria de asambleas en la hora de clase, impartición en horario de viernes a las 9 de la noche, etc. Esto provocaría que un experimentado y reconocido profesor fracasase ante la presencia de un cúmulo de tan indeseables circunstancias.

En situaciones como la expuesta, se plantea desarrollar un modelo estadístico en el cual la variabilidad total que afecta a un experimento pueda descomponerse en función de los distintos factores o causa que la provocan. Así, en su forma más simple, un modelo de estas características sería de tipo aditivo:

$$\text{Variabilidad Total} = \sum_{i=1}^{p} \text{Variabilidad debida al Factor i} + \text{Variabilidad residual}$$

La suma de variabilidades debidas a los factores controlables se denomina variabilidad *explicada*, mientras que la residual, que engloba todas aquellas causas que el investigador no puede controlar tales como la propia aleatoriedad del experimento, se denomina variabilidad *no explicada*.

Un modelo de descomposición lineal de la variabilidad de estas características se denomina Análisis de la Varianza con p factores o, abreviadamente, ANOVA p (del inglés ANalysis Of Variance). Cuando sólo existe una fuente de variablidad se dice que el análisis de la varianza es *simple* y se denota ANOVA I. Cuando existe más de una fuente de variabilidad se dice que el modelo es de tipo *factorial*.

Aunque la variable de respuesta del experimento es cuantitativa (por ejemplo, la calificación obtenida por los alumnos), los factores que afectan a la variabilidad son de tipo cualitativo (el grupo de la asignatura, la presencia o ausencia de un fenómeno inesperado, ...) y cada uno tiene, a su vez, varios niveles (grupo A, B, C,...). Cuando se consideran todos los posibles niveles de los factores se dice que el modelo ANOVA es de *efectos fijos*, mientras que si en cada factor se consideran solo algunos niveles tomados al azar de todos los posibles (quizás porque hay muchos) se dice que el ANOVA es de *efectos aleatorios*. Cuando en unos factores se consideran todos sus niveles y en otros sólo algunos muestrados al azar, se dice que el ANOVA es de *efectos mixtos*.

En conclusión, el ANOVA trata de determinar, a través del estudio de la varianza, si los distintos niveles de los factores pueden conllevar diferencias en la respuesta en los distintos grupos, contrastando para ello la igualdad de medias de la variable dependiente en dichos grupos. Para ejecutar un ANOVA se parte de tres premisas que deben cumplir los niveles de cada factor:

1. Cada grupo debe distribuirse según una ley normal
2. Los grupos han de tener igual varianza (condición de *homocedasticidad*)
3. Los grupos han de ser variables independientes entre si.

Cuando el tamaño muestral es igual entodos los grupos se dice que el diseño es *balanceado*, y en caso contrario *no balanceado.*

4.2. DISEÑOS UNIFACTORIALES: EL MODELO ANOVA I

4.2.1. Planteamiento y resolución del contraste ANOVA

Consideremos que la variabilidad de un experimento se debe a un único factor que presenta k niveles: $X_1, X_2, ..., X_k$ verificando las tres condiciones anteriores, de manera que podemos representar cada variable como $X_i \rightsquigarrow N(\mu_i; \sigma)$. El contraste ANOVA I plantea entonces las hipótesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$
$$H_1 : \textit{No todas las } \mu_i \textit{ son iguales}$$

Evidentemente, en caso de aceptar H_0 las k variables serán idénticas. El ANOVA I puede considerarse una extensión del contraste bilateral de igualdad de medias con varianzas idénticas, desarrollado en la sección 3.3 del tema anterior, con la ventaja de que, en caso de aceptarse la hipótesis nula, ahorra realizar un total de $k(k-1)/2$ contrastes del tipo:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j$$

que es el resultado de realizar las combinaciones de k elementos tomados de dos en dos. Así, en el caso de 3 grupos habría que formular 3 contrastes de este tipo, pero en el caso de 10 grupos el número de contrastes dos a dos ascendería a 45.

Para resolver el contraste se seleccionan muestras aleatorias de cada grupo o nivel X_i de la forma:

Grupos	X_1	X_2	$\cdots$	X_k
	x_{11}	x_{21}	$\cdots$	x_{k1}
	x_{12}	x_{22}	$\cdots$	x_{k2}
	$\cdot$	$\cdot$	$\cdot$	$\cdot$
	$\cdot$	$\cdot$	$\cdot$	$\cdot$
	$\cdot$	$\cdot$	$\cdot$	$\cdot$
	x_{1n_1}	x_{2n_2}	$\cdots$	x_{kn_k}
Medias	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$

Denotemos, a su vez, a la media global de todos los datos $\bar{\bar{x}}$. Cada dato x_{ij} puede expresarse en función de la supuesta media común μ, en caso de aceptar la hipótesis nula, de la forma:

$$x_{ij} = \mu_i + \varepsilon_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij} = \mu + \beta_i + \varepsilon_{ij}$$

donde los $\varepsilon_{ij} \rightsquigarrow N(0, \sigma)$ y los $\beta_i = \mu_i - \mu$, que representan el efecto grupo, verifican:

$$\sum_{i=1}^{k} \beta_i = 0.$$

Sin más que operar puede obtenerse la siguiente descomposición:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

siendo:

$VT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2$ la variabilidad total

$VE = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$ la variabilidad entre grupos (explicada)

$VR = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$ la variabilidad intra grupos (residual)

Aplicando la esperanza matemática, puede probarse que:

$$E\left[VE\right] = (k-1)\sigma^2 \qquad E\left[VR\right] = (n-k)\sigma^2 + \sum_{i=1}^{k} n_i(\mu_i - \mu)^2$$

de donde se deduce de forma inmediata que estimación insesgada de σ^2 es $ME = \frac{VE}{k-1}$; y, si H_0 es cierta, otra estimación insesgada de σ^2 es $MR = \frac{VR}{n-k}$. Se puede demostrar además el siguiente resultado:

Teorema.

Si H_0 es cierta entonces $\frac{VE}{\sigma^2} \rightsquigarrow \chi^2_{k-1}$ y $\frac{VR}{\sigma^2} \rightsquigarrow \chi^2_{n-k}$ siendo ambas independientes, por lo que $\frac{ME}{MR} \rightsquigarrow F_{(k-1,n-k)}$

En consecuencia, para la resolución del ANOVA I se construye una tabla del tipo siguiente:

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	$F_{\exp}$
Entre grupos	VE	$k-1$	$ME = \frac{VE}{k-1}$	$\frac{ME}{MR}$
Intra grupos	VR	$n-k$	$MR = \frac{VR}{n-k}$	p
Total	VT	$n-1$		

El ANOVA de una vía se considera una prueba robusta frente a la falta de normalidad, es decir tolera bien las violaciones a su supuesto de normalidad siempre que no sea excesivas. En distribuciones sesgadas o leptocúrticas (empinadas) tolera datos que no son normales con sólo un pequeño efecto sobre la probabilidad de error tipo I (nivel de significación). Sin embargo, las distribuciones platicúrticas (aplanadas) pueden tener un efecto profundo cuando los tamaños muestrales de los grupos son pequeños. Ante la falta de normalidad tenemos dos alternativas:

(1) transformar los datos para que la forma de la distribución sea normal

(2) elegir una prueba no paramétrica que no supone normalidad.

4.2.2. Contrastes de igualdad de varianzas

Una de las condiciones para poder formular un ANOVA es que todas las variables tengan igual varianza, le denominada propiedad de homocedasticidad. Existen varios métodos estadísticos para resolver el contraste:

$H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$
H_1 :*No todas las* σ_i *son iguales*

tales como el test de Levene, de Brown-Forsythe, de Bartlett, de Cochran, de Hartley, de Layard, de Fligner-Killeen, etc. Algunos de los más utilizados se describen a continuación.

Test de Levene.

Esta prueba se basa en las desviaciones absolutas: $d_{ij} = \mid x_{ij} - \bar{x}_i|$. para $i = 1, 2, ..., k$. Así, denotando

$$\bar{d}_i = \frac{\sum_{j=1}^{n_i} d_{ij}}{n_i} \qquad \bar{\bar{d}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij}}{n}$$

el estadístico de contraste es:

$$W = \frac{n-k}{k-1} \frac{\sum_{i=1}^{k} n_i \left(\bar{d}_i - \bar{\bar{d}} \right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(d_{ij} - \bar{d}_i \right)^2}$$

que se distribuye según un modelo de Fisher-Snedecor con $(k-1, n-k)$ grados de libertad.

Una modificación del test de Levene, especialmente útil cuando la distribución de las variables no es Gaussiana, consiste en considerar las desviaciones absolutas de los datos respecto a la mediana en lugar de respecto a la media, es decir: $d_{ij}^* = \mid x_{ij} - Mediana(x_i) \mid$, en cuyo caso el estadístico W también se distribuye según una $F_{(k-1,n-k)}$. Dicha prueba se denomina **test de Brown- Forsythe** y es más robusta que la de Levene.

Test de Bartlett

Consiste en evaluar el estadístico:

$$B = \frac{(n-k) \ln s_p^2 - \sum_{i=1}^{k} (n_i - 1) \ln s_i^2}{1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^{k} \left(\frac{1}{n_i - 1} \right) - \frac{1}{n-k} \right]}$$

donde

$$s_p^2 = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{n-k}$$

el cual se distribuye según una ley *chi-cuadrado* con $k-1$ grados de libertad. Cuando los datos proceden de distribuciones normales es más preciso que el de Levene.

Otros métodos alternativos para tratar la heterocedasticidad consisten en realizar transformaciones de las variables. No obstante, en estos casos, el procedimiento más adecuado consiste en utilizar la **corrección de Welch**, que consiste en ajustar el denominador MR del estadístico de contraste F_{exp} de forma que tenga la misma esperanza que el numerador ME cuando H_0 es verdadera pese a la falta de igualdad de las varianzas de los grupos.

EJEMPLO

Supongamos que X_1, X_2 y X_3 son tres variables independientes normalmente distribuidas. A partir de las muestras de la tabla siguiente se quiere estudiar si son idénticas o no lo son al nivel $\alpha = 0,05$.

	X_1	X_2	X_3
	$2,8$	$4,4$	
	$3,5$	$4,2$	$3,3$
	$3,1$	$4,7$	$3,8$
	$3,0$	$4,1$	$3,4$
	$3,2$	$4,5$	$3,1$
		$4,2$	
n_i	5	6	4
$\bar{x}_i$	$3,12$	$4,35$	$3,40$

El primer paso consistirá en contrastar la hipótesis de homocedasticidad (igualdad de varianzas), para lo cual aplicaremos la prueba de Levene. Así:

	d_1	d_2	d_3
	$0,32$	$0,05$	
	$0,38$	$0,15$	$0,10$
	$0,02$	$0,35$	$0,40$
	$0,12$	$0,25$	$0,00$
	$0,08$	$0,15$	$0,30$
		$0,15$	
$\bar{d}_i$	$0,184$	$0,183$	$0,20$

La media global de las diferencias absolutas es $\bar{\bar{d}} = 0,188$. Por tanto:

$$W = \frac{15-3}{3-1}\frac{5(0,184-0,188)^2+(0,183-0,188)^2+(0,2-0,188)^2}{(0,32-0,184)^2+(0,38-0,184)^2+\cdots+(0,30-0,20)^2} = 0,021$$

y como el valor crítico es $F^{0,05}_{(2,12)} = 3,8853$, se acepta la hipótesis de igualdad de varianzas.

Podemos proceder entonces a realizar el ANOVA. Dado que la media total es $\overline{\overline{x}} = 3,687$, las variabilidades serán:

$$VE = 5(3,12 - 3,687)2 + 6(4,35 - 3,687)2 + 4(3,40 - 3,687)2 = 4,574$$
$$VT = (2,8 - 3,687)2 + (3,5 - 3,687)2 + \cdots + (3,1 - 3,687)2 = 5,357$$
$$VR = 5,357 - 4,574 = 0,783$$

y la tabla del ANOVA

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	$4,574$	2	$2,287$	$35,05$
Intra grupos	$0,783$	12	$0,065$	$(p = 0,0000)$
Total	$5,357$	14		

rechazándose la hipótesis de igualdad de medias dado que el valor crítico en las tablas es: $F^{0,05}_{(2,12)} = 3,8853$.

4.3. COMPARACIONES MÚLTIPLES: CONTRASTES "POST HOC"

Cuando la regla de decisión del ANOVA I nos lleva a rechazar la hipótesis nula, lo único que podemos concluir es que no son iguales las medias de los k grupos considerados, pero ello no significa que sean todas distintas, pues pueden ser todas iguales menos una, o menos dos,... Así, habría que formular un total de $k(k-1)/2$ contrastes del tipo:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j$$

que es el resultado de realizar las combinaciones de k elementos tomados de dos en dos. Así, en el caso de 3 grupos habría que formular 3 contrastes de este tipo, pero en el caso de 10 grupos el número de contrastes dos a dos ascendería a 45.

Una alternativa a los test de comparación de medias basados en la distribución t de Student descrito en la sección 3.4 son los denominados con-

trastes de comparaciones múltiples o *post hoc*, que utilizan la información proporcionada por las muestras de todas las variables y no solo de las dos implicadas. Entre ellos cabe citar el de Bonferroni, el de Duncan, el de Sheffé, el de Tukey, el de Newman-Keuls, el LSD de Fisher, etc., los cuales suelen estar implemenados por los paquetes estadísticos comerciales más usuales. A continuación describiremos, de forma sucinta, los dos primeros.

Contraste de Bonferroni

El contraste de comparación de medias dos a dos con varianzas iguales utilizaba el estadístico de contraste:

$$t_{\text{exp}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

que se distribuye según una ley a t de Student con $n_1 + n_2 - 2$ grados de libertad, siendo

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

una media ponderada de las cuasivarianzas de las dos variables. El contraste de Bonferroni propone sustituir s_p^2 por MR, con lo que el estadístico de contraste:

$$t_{\text{exp}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{MR\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

se distribuye ahora según una t de Student con $n - k$ grados de libertad.

El principal inconveniente de este procedimiento radica en que, si se quieren comparar dos a dos las k variables de un estudio, la probabilidad total de cometer un error tipo I, es decir, rechazar la hipótesis nula en uno de los contrastes siendo cierta, es siempre mayor que el nivel de significación común α de los contrastes dos a dos, concretamente es, a lo sumo, $1-(1-\alpha)^k$. Por ejemplo, si se dispone de 10 variables y se considera el nivel estándar $\alpha = 0,05$ entonces la probabilidad de realizar un rechazo incorrecto es, como máximo, $1 - (1 - 0,05)^{10} = 0,40$. De esta forma, si el número k de variables fuese muy elevado, la probabilidad de un falso rechazo sería excesivamente

alta. Por ello es aconsejable realizar sólo los contrastes múltiples que se considere estrictamente necesario y obviar los evidentes.Es decir, la realización de contrastes múltiples produce una especie de acumulación de las probabilidades de errores tipo I equivalente a la que ocurriría si un médico encarga un número elevado de pruebas diagnósticas donde van sumando las probabilidades de falsos positivos concluyendo erróneamente que el paciente presenta una cierta patología que no es real. Se debe realizar sólo el mínimo número de pruebas necesarias.

Contraste de Duncan

El contraste de Duncan tiene en cuenta el orden creciente de las medias muestrales, de forma que al comparar dos medias cualesquiera se sabe el número de medias que están entre ambas y así las medias muestrales que se encuentren en medio no requieren presentar tanta diferencia entre si como la de los extremos para concluir que hay diferencias significativas entre las medias poblacionales. Esto supone un reajuste de los valores críticos.

Merece la pena resaltar que el contraste de Duncan presenta menos superposiciones en la clasificación de grupos que el de Bonferroni

EJEMPLO

Para ensayar la eficacia de cuatro tratamientos A, B, C, y D contra la hipertensión se aplicaron de forma aleatoria a un grupo de pacientes de características similares y se registró el descenso de la presión diastólica desde el estado basal hasta el estado al cabo de una semana. Los resultados obtenidos fueron los siguientes (los valores negativos indican aumento de la presión en lugar de disminución):

A	B	C	D
10	20	15	10
0	25	10	5
15	33	25	-5
-20	25	30	15
0	30	15	20
15	18	35	20
-5	27	25	0
	0	22	10
	35	11	
	20	25	

Suponiendo que las variables implicadas se distribuyen normalmente y tienen igual varianza, vamos a aplicar un ANOVA I para estudiar si existen diferencias significativas entre los cuatro tratamientos.

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Entre grupos	$2492,61$	3	$830,87$	$8,53$
Intra grupos	3020.93	31	$97,45$	$(p = 0,0003)$
Total	$5513,54$	34		

por lo que sí hay diferencias significativas entre las medias de las variables. A continuación se muestra el resultado de las agrupaciones obtenidas con los diferentes contrastes *post hoc* (los cálculos se han realizado con el programa SPSS):

Grupo	Media	Bonferroni			Duncan		
A	$2,143$	X			X		
D	$9,375$	X	X		X		
C	$21,3$		X	X			X
B	$23,3$			X			X

4.4. DISEÑOS FACTORIALES: EL MODELO ANOVA II

Supongamos ahora que la variabilidad de un experimento se debe a dos factores controlables, y que los valores muestreados de la variable de respuesta para cada uno de los factores se representan en la tabla siguiente:

	Factor 2 (columna)				
Factor 1 (fila)	C_1	C_2	...	C_k	Medias filas
F_1	x_{11}	x_{12}	...	x_{1k}	$\bar{x}_{1\bullet}$
F_2	x_{21}	x_{22}	...	x_{2k}	$\bar{x}_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	
F_h	x_{h1}	x_{h2}	...	x_{hk}	$\bar{x}_{h\bullet}$
Medias columnas	$\bar{x}_{\bullet 1}$	$\bar{x}_{\bullet 2}$	...	$\bar{x}_{\bullet k}$	$\bar{\bar{x}}$

Por su disposición, en la tabla denominaremos a los factores: fila y columna. También, en ocasiones, se les denomina factor principal y factor secundario, como ocurre cuando se quiere comparar la eficacia de varios tratamientos ante la Covid-19, pero se sospecha que existe un segundo factor que puede influir en el resultado como es el haber padecido o no previamente meningitis.

Partiendo de que se cumplen las tres condiciones previas del ANOVA, denotando μ a la media global, $\mu_{i\bullet}$ a la media de la variable fila F_i, $\mu_{\bullet j}$ a la media de la variable columna C_j, y $\alpha_i = \mu_{i\bullet} - \mu$ al efecto del nivel i-simo del factor fila y $\beta_j = \mu_{\bullet j} - \mu$ al efecto del nivel j-simo del factor columna, los contrastes a realizar en el ANOVA II son:

$$\begin{array}{ll} H_0^F: \quad \alpha_1 = \alpha_2 = \cdots = \alpha_h = 0 & H_0^C: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1^F: \quad \text{No todos los } \alpha_i \text{ son cero} & H_1^C: \quad \text{No todos los } \beta_j \text{ son cero} \end{array}$$

Como extensión inmediata a lo que ocurría en el caso unifactorial, cada dato x_{ij} puede expresarse en función de la supuesta media común μ, en caso de aceptar la hipótesis nula, de la forma:

$$x_{ij} = \mu_i + \varepsilon_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

donde los $\varepsilon_{ij} \rightsquigarrow N(0, \sigma)$, los α_i y β_j verifican:

$$\sum_{i=1}^{h} \alpha_i = 0 \qquad \sum_{j=1}^{k} \beta_j = 0.$$

Sin más que operar puede obtenerse la siguiente descomposición:

$$\sum_{i=1}^{h}\sum_{j=1}^{k_i} \left(x_{ij} - \bar{\bar{x}}\right)^2 =$$

$$k\sum_{i=1}^{h}(\bar{x}_{i\bullet} - \bar{\bar{x}})^2 + h\sum_{j=1}^{k}(\bar{x}_{\bullet j} - \bar{\bar{x}})^2 + \sum_{i=1}^{h}\sum_{j=1}^{k} \left(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{\bar{x}}\right)^2$$

siendo:

$VT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2$ la variabilidad total

$VE_F = k\sum_{i=1}^{h}(\bar{x}_{i\bullet} - \bar{\bar{x}})^2$ la variabilidad explicada por el efecto fila

$VE_F = h\sum_{j=1}^{k}(\bar{x}_{\bullet j} - \bar{\bar{x}})^2$ la variabilidad explicada por el efecto columna

$VR = \sum_{i=1}^{h}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{\bar{x}}\right)^2$ la variabilidad no exlicada (residual)

Aplicando la esperanza matemática, se deduce de forma inmediata que $ME_F = \frac{VE_F}{h-1}$ y $ME_C = \frac{VE_C}{k-1}$ son sendas estimaciones insesgadas de σ^2; y, si H_0^F y H_0^C son ciertas, otras estimaciones insesgadas de σ^2 son $MR = \frac{VR}{(h-1)(k-1)}$. Se puede demostrar además el siguiente resultado:

Teorema.

a) Si H_0^F es cierta entonces $\frac{VE_F}{\sigma^2} \rightsquigarrow \chi^2_{h-1}$ y $\frac{VR}{\sigma^2} \rightsquigarrow \chi^2_{(h-1)(k-1)}$ siendo ambas independientes, por lo que $\frac{ME_F}{MR} \rightsquigarrow F_{(h-1,(h-1)\cdot(k-1))}$

b) Si H_0^C es cierta entonces $\frac{VE_C}{\sigma^2} \rightsquigarrow \chi^2_{k-1}$ y $\frac{VR}{\sigma^2} \rightsquigarrow \chi^2_{(h-1)(k-1)}$ siendo ambas independientes, por lo que $\frac{ME_C}{MR} \rightsquigarrow F_{(k-1,(h-1)\cdot(k-1))}$

En consecuencia, para la resolución del ANOVA II se construye una tabla del tipo siguiente:

Fuente de variabilidad	Variab.	Grados de libertad	Cuadrados medios	F_{exp}
Entre filas	VE_F	$k-1$	$ME_F = \frac{VE_F}{h-1}$	$\frac{ME_F}{MR}$ (valor p)
Entre columnas	VE_C	$h-1$	$ME_F = \frac{VE_C}{k-1}$	$\frac{ME_C}{MR}$ (valor p)
Residual	VR	$(h-1)(k-1)$	$MR = \frac{VR}{(h-1)(k-.1)}$	
Total	VT	$hk-1$		

Dado que los programas estadísticos más usuales tienen implementado el ANOVA II, no consideramos importante el aspecto de cálculo de este análisis sino su planteamiento e interpretación e los resultados.

EJEMPLO

Se han sembrado 5 variedades de maíz (A, B, C, D y E) en 3 tipos de terreno (solana, umbría y mixto), obteniéndose los siguientes rendimientos expresados en Qm/Ha:

	A	B	C	D	E	Medias
Solana	280	300	310	270	330	298
Umbría	250	240	260	270	240	252
Mixto	310	304	290	280	300	296, 8
Medias	280	281, 3	286, 7	273, 3	290	282, 27

Tras realizar los cálculos oportunos, la tabla del ANOVA II sería la siguiente:

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Tipo de terreno	6874, 13	2	3437, 07	9, 08 (p=0,0088)
Variedad de maíz	494, 93	4	123, 73	0, 33 (p=0,8526)
Residual	3019, 87	8	378, 73	
Total	10398, 9	14		

En conclusión, el factor terreno sí es significativo, en cuanto que la producción de maíz depende del terreno donde se cultiva, mientras que el factor variedad de maíz no lo es.

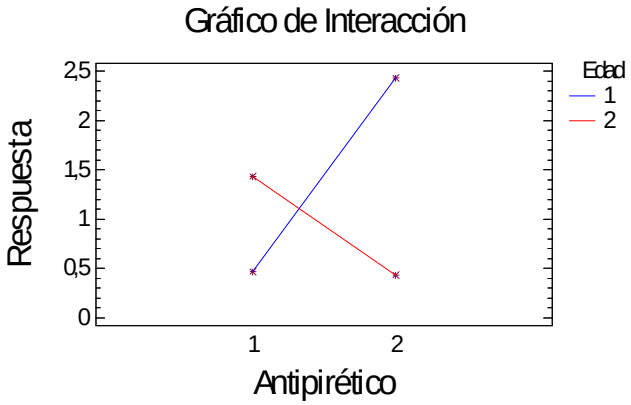
El planteamiento desarrollado en esta sección ha considerado la presencia de dos factores controlables que pueden influir en el resultado de un experimento, pero es inmediatamente extensible al caso de tres o más factores de variabilidad.

4.5. INTERACCIÓN ENTRE FACTORES

Una cuestión de gran importancia práctica en la interpretación del resultado de un ANOVA factorial es la interacción entre factores, que se produce cuando el efecto de uno de los factores sobre la variable de respuesta no es igual en todos los niveles de los demás factores, es decir, cuando el resultado de la combinación de dos o más factores es diferente a la suma de los efectos principales de esos factores.

EJEMPLO

Se quiere evaluar la eficacia de dos antipiréticos para fiebres altas (superiores a 38,5^{o}C) considerando como variable de respuesta la reducción de temperatura corporal transcurrida una hora desde su administración. Para



ello se aplican ambos medicamentos a dos grupos de población: edades inferiores a 14 años y edades a partir de 14 años, obeniéndose los siguientes resultados:

Antipirético	A			B		
Menores de 14 años	0,3	0,7	0,4	2,3	2,5	2,5
A partir de 14 años	1,6	1,2	1,5	0,5	0,6	0,2

El resultado del ANOVA II se recoge en la tabla siguiente:

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Antipirético	$0,7008$	1	$0,7008$	$0,92$ (p=0,3636)
Grupo de edad	$0,8008$	1	$0,8008$	$1,05$ (p=0,3330)
Residual	$6,8875$	9	$0,7653$	
Total	$8,3891$	11		

De ella se concluye que no hay diferencias significativas de la respuesta entre los dos antipiréticos ni entre los grupos de edad. Sin embargo, a partir de los datos se observa que la reducción de temperatura superior en menores de 14 años para el antipirético B, mientras ocurre al contrario con el A. Gráficamente sería:

Pues bien, considerando un diseño bifactorial balanceado con r observaciones en cada casilla, el efecto de la interacción aparece como un sumando más del modelo aditivo de descomposición de la varianza del tipo:

$$VI = r\sum_{i=1}^{h}\sum_{j=1}^{k}\left(\bar{x}_{ij\bullet} - \bar{\bar{x}}_{i\bullet\bullet} - \bar{\bar{x}}_{\bullet j\bullet} + \bar{\bar{\bar{x}}}\right)^2$$

donde $\bar{x}_{ij\bullet}$representa la media de los r datos de la casilla (i, j). Así, la tabla del ANOVA II con interacción sería de la forma:

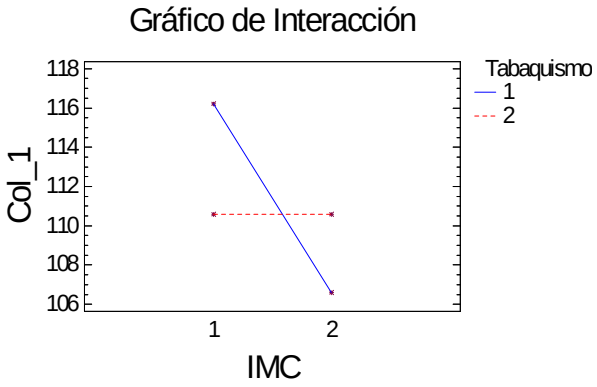
Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Entre filas	VE_F	$k-1$	$ME_F = \frac{VE_F}{h-1}$	$\frac{ME_F}{MR}$ (valor p)
Entre columnas	VE_C	$h-1$	$ME_F = \frac{VE_C}{k-1}$	$\frac{ME_C}{MR}$ (valor p)
Interacción	VI	$(h-1)(k-1)$	$MI = \frac{VI}{(h-1)(k-1)}$	$\frac{MI}{MR}$ (valor p)
Residual	VR	$hk(r-1)$	$MR = \frac{VR}{hk(r-1)}$	
Total	VT	$hkr-1$		

EJEMPLO

Sobre el mismo ejemplo anterior, la interacción se expresaría sobre la tabla del ANOVA de la forma siguiente:

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Antipirético	$0,7008$	1	$0,7008$	$19,56$ (p=0,0022)
Grupo de edad	$0,8008$	1	$0,8008$	$22,35$ (p=0,0015)
Interacción	$6,6008$	1	$6,6008$	$184,21$ (p=0,0000)
Residual	$0,2867$	8	$0,0358$	
Total	$8,3891$	11		

y cambiaría completamente la interpretación del ANOVA.



EJEMPLO

Se quiere estudiar la influencia del sobrepeso y del tabaquismo sobre el colesterol high-density (HDL) para lo cual se consideraron dos grupos según su Índice de Masa Corporal (I.M.C.) y sus hábitos de fumar. Los resultados del HDL (en mg/gL) fueron los siguientes:

IMC (kg/m^2)	No fumador					Fumador				
Superior a 25	120	116	114	117	114	112	107	110	111	113
Hasta 25	107	108	105	107	106	111	111	110	111	110

Al nivel α=0,05 se obtiene la siguiente tabla del ANOVA:

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
IMC	$115,2$	1	$115,2$	$35,18$ (p=0,0000)
Tabaquismo	$3,2$	1	$3,2$	$0,98$ (p=0,3376)
Interacción	$115,2$	1	$115,2$	$35,18$ (p=0,0000)
Residual	$52,4$	16	$3,275$	
Total	286	19		

De donde se concluye que el único factor significativo es el IMC y la interacción entre los dos factores, cuyo gráfico es el siguiente:

EJEMPLO

Suponiendo que se cumplen las hipótesis del ANOVA se quiere comparar el área bajo la curva (AUC) del precirol a dos dosis diferentes, considerando como posibles factores de variabilidad la presión y el calor.

Dosis- Pr esión	Con calor	Sin calor
	39,80	32,32
30mg-5Tons	39,77	36,86
	40,31	33,06
	40,96	36,87
30mg-10Tons	42,53	32,66
	43,88	31,49
	44,05	32,25
60mg-5Tons	44,21	31,63
	42,97	29,91
	43,87	31,58
60mg-10Tons	44,60	32,36
	43,52	28,09

Al nivel α=0,05 se obtiene la siguiente tabla del ANOVA:

Fuente de variabilidad	Variabilidad	Grados de libertad	Cuadrados medios	F_{exp}
Dosis	$0,09$	1	$0,09$	$0,03$ (p=0,8634)
Calor	$613,98$	1	$613,98$	$208,43$ (p=0,0000)
Presión	$1,157$	1	$1,157$	$0,39$ (p=0,5396)
Inter. Dosis-Calor	$46,51$	1	$46,51$	$15,79$ (p=0,0011)
Inter. Dosis-Presión	$5,255$	1	$5,255$	$1,78$ (p=0,2004)
Inter. Calor-Presión	$2,202$	1	$2,202$	$0,75$ (p=0,4000)
Inter. Dosis-Calor-Presión	$1,597$	1	$1,597$	$0,54$ (p=0,4723)
Residual	$47,131$	16	$2,946$	
Total	$717,922$	23		

Así, el único factor significativo en el AUC es el calor y la única interacción significativa es la existente entre las dosis y el calor.

4.6. DISEÑO MEDIANTE CUADRADOS LATINOS

En situaciones en las que las unidades de experimentación son limitadas, como ocurre en agronomía donde no se dispone de terreno suficiente como para poder replicar la experiencia un número suficiente de veces, resulta útil realizar un diseño en *cuadrado latino*. Se trata de una disposición en matriz cuadrada de letras de forma que cada una aparezca sólo una vez en cada fila y en cada columna. Así, cuadrados latinos de orden 2 son:

$$\begin{array}{cc} A & B \\ B & A \end{array} \qquad \begin{array}{cc} B & A \\ A & B \end{array}$$

Algunos de orden 3:

$$\begin{array}{ccc} A & B & C \\ B & C & A \\ C & B & A \end{array} \qquad \begin{array}{ccc} A & C & B \\ C & B & A \\ B & A & C \end{array} \qquad \begin{array}{ccc} B & A & C \\ A & C & B \\ C & B & A \end{array} \qquad \begin{array}{ccc} C & A & B \\ B & C & A \\ A & B & C \end{array}$$

Esta situación se presenta en una situación en la que, por ejemplo, se quiere comparar la eficacia de 3 fertilizantes en terrenos donde se va a plantar trigo, considerando como variable de respuesta la producción medida en Qm/Ha. Para evitar la variabilidad debida al suelo, pues la parcela experimental se encuentra en la ladera de una monte, ésta se divide en 3 subparcelas con dos gradientes o factores de variación: uno es según su ubicación (vega, medio monte y terreno pedregoso), y el otro es según las horas de sol que recibe (solana, umbría y mixto). Así, se elige una disposición en cuadrado latino de orden 3, que garantiza que cada uno de los 3 fertilizantes se aplica en un terreno en el que concurren cada uno de los niveles de los dos factores de variación.

En ocasiones se quiere considerar un tercer factor de variabilidad, como puede ser si el terreno es de secano, regadío por acequia o tiene riego artificial. En tal caso es necesario cruzar un cuadrado latino de orden 3 con otro también de orden 3, el cual se representa mediante letras griegas para diferenciarlo. Un ejemplo sería:

$$\begin{array}{ccc} A-\beta & B-\alpha & C-\gamma \\ B-\gamma & A-\beta & C-\alpha \\ C-\alpha & B-\gamma & A-\beta \end{array}$$

Un tal diseño se denomina en *cuadrado greco-latino*.

Tema 5. MODELOS DE REGRESIÓN

5.1. CONCEPTO DE CORRELACIÓN Y REGRESIÓN

Una de las cuestiones de mayor interés en las Ciencias Experimentales consiste en obtener un modelo matemático que relacione dos o más magnitudes variables a partir de observaciones experimentales. En ocasiones tales relaciones pueden deducirse a partir de consideraciones teóricas. Sin embargo, en la mayor parte de los fenómenos objeto de investigación experimental no es posible deducir una relación exacta entre las variables, en cuanto que la dependencia perfecta no existe en la Naturaleza. Así, por ejemplo, se sabe que entre la altura y peso de una persona existe cierta dependencia, pero ésta, claramente, no es de tipo funcional, en cuanto que el conocimiento de la altura de un individuo no nos permite deducir de forma exacta su peso, ni recíprocamente, sino tan sólo tener una idea aproximada de su valor. Se presenta, por tanto, una dependencia aproximada entre las variables, que es preciso medir numéricamente. Este tipo de dependencia se denomina *correlación*, siendo sus casos extremos la dependencia funcional o exacta, y la independencia.

Suponiendo que un variable Y está correlacionada con un conjunto de variables $x_1, x_2, ..., x_k$ el problema de la regresión consiste en estimar el valor medio de dicha variable Y, que denominaremos dependiente o de respuesta, a partir de valores de las variables x_i que se denominan independientes o explicativas, es decir, la esperanza condicionada $E[Y/x_1, x_2, ..., x_k]$. En caso de haber una única variable explicativa x se dice que la regresión es *simple*, y en caso de haber dos o más variables independientes se dice que la regresión es *múltiple*. En el planteamiento básico el investigador controla el valor de las variables independientes y, para cada valor de las x_i, puede obtener una serie de observaciones o medidas de la respuesta Y, de manera que ésta es la única variable aleatoria. Así, en el caso simple, el esquema sería el siguiente:

Valores dados a x	Valores observados de Y	Media de Y
x_1	$y_{11}\ y_{12}\cdots\ y_{1k_1}$	$\bar{y}_1$
x_2	$y_{21}\ y_{22}\cdots\ y_{2k_2}$	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$
x_n	$y_{n1}\ y_{n2}\cdots\ y_{nk_n}$	$\bar{y}_n$

y la función de regresióné estimada de Y sobre x viene dada por el conjunto de puntos $(x_i, \bar{y}_i)$.

La descomposición de la variabilidad se plantea en términos similares al ANOVA:

$$\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(y_{ij}-\bar{\bar{y}}\right)^2=\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2 k_i+\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(y_{ij}-\bar{y}_i\right)^2$$

siendo $\bar{\bar{y}}$ la media de todos los valores de la variable de respuesta y:

$VT=\sum_{i=1}^{n}\left(y_i-\bar{\bar{y}}\right)^2$ la variabilidad total

$VE=\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2 k_i$ la variabilidad explicada por la regresión

$VR=\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(y_{ij}-\bar{y}_i\right)^2$ la variabilidad residual o no explicada

Así, el cociente $R^2=\frac{VE}{VT}$, que se denomina *coeficiente de determinación*, mide la proporción en que la regresión representa a los datos. Claramente es $0\leq R^2\leq 1$, de forma que valores próximos a 1 indican que el grado de explicación de la variable de respuesta a través de las medias condicionadas es alto. Por el contrario, valores próximo a 0 indican incorrelación entre las variables que, en caso de que la respuesta se distribuya normalmente, equivaldrá a independencia. En ocasiones resulta más sencillo evaluar el coeficiente de determinación de la forma: $R^2=1-\frac{VR}{VT}$.

EJEMPLO

Consideremos un caso simple en el que se quiere explicar la actividad media de un fármaco (en mg) en función del tiempo transcurrido, para lo cual se

mide dicha actividad en varios instantes sobre tres unidades experimentales, obteniéndose una tabla como la siguiente:

Tiempo (*meses*)	Actividad (*mg*)	Actividad media (*mg*)
0	51 51 53	51,67
3	51 50 52	51
6	50 52 48	50
9	49 51 51	50,33
12	49 48 47	48
18	47 45 49	47

La función de regresión es el conjunto o nube de puntos:

$$(0; 51, 7), (3; 51), (6; 50), (9; 50, 3), (12; 48), (18; 47).$$

Como la media global de los 18 datos de actividad es: $\bar{\bar{y}} = 49, 67mg$, tenemos:

$$VT = (51 - 49, 67)^2 + (51 - 49, 67)^2 + \cdots + (49 - 49, 67)^2 = 74$$

$$VE = (51, 7 - 49, 67)^2 \cdot 3 + (51 - 49, 67)^2 \cdot 3 + \cdots + (47 - 49, 67)^2 \cdot 3 = 48, 54$$

Por tanto:

$$R^2 = \frac{48,54}{74} = 65, 6\%$$

Podría suceder que para cada valor de x se midiera un único valor de la respuesta Y, es decir, se dispusiera de una tabla de datos de la forma:

$$\begin{array}{lcccc} x: & x_1 & x_2 & \cdots & x_n \\ y: & y_1 & y_2 & \cdots & y_n \end{array}$$

En tal caso sería $\bar{y}_i = y_i$ y todas las $k_i = 1$.

Originalmente, el término regresión proviene de *regresar*, y es que a finales del siglo XIX, el antropólogo británico Sir Francis Galton, primo de Charles Darwin y cofundador junto a su discípulo Karl Pearson y a Walter Weldon de la revista *Biometrika*, en la línea de estudios genéticos y hereditarios propia de esa época, realizó un estudio sobre la transmisión de padres a hijos del

factor estatura. Él observó en un estudio que los hijos de padres altos son también altos pero, en promedio, la diferencia entre la estatura media de ellos y la de los individuos de su generación se reduce $\frac{2}{3}$ con respecto a esa diferencia en la generación de sus progenitores. Así, publicó en 1889 el libro *Natural Inheritance* (editorial MacMillan & Co., London) en el que enunció la *Ley de Regresión*, según la cual la estatura tiende a *regresar* a la media de la raza, siendo por tanto la componente racial más fuerte que la hereditaria. Como dato anecdótico, Galton nunca ocupó una cátedra universitaria sino que realizó las investigaciones por cuenta propia. Dos años antes de su fallecimiento, acaecido en 1911 a la edad de 88 años, recibió el título de *Sir*.

En las aplicaciones prácticas el problema consiste en aproximar la función de regresión mediante algún modelo matemático $\varphi(x_1, x_2, ..., x_k)$ que represente de forma óptima la tendencia mostrada por la nube de puntos, es decir:

$$E\left[Y/x\right] = \varphi(x_1, x_2, ..., x_k) + \varepsilon$$

donde ε denota el error aleatorio o desviación de los datos respecto del modelo. En general, dicho modelo dependerá de varios parámetros que será necesario estimar a partir de los datos. En el caso más sencillo, se tratará de un modelo de tipo lineal:

$$\varphi(x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Cuando, a su vez, la regresión es simple, la elección del modelo matemático $y = \varphi(x)$ con el que aproximar la regresión dependerá, en gran medida, de la tendencia observada por la nube de puntos

El esquema básico que seguiremos en este tema será el enunciado anteriormente, es decir, el investigador controla una o varias variables explicativas y, a partir de valores asignados a ellas, mide la respuesta; una tal regresión se dice que es de tipo 1. Pero puede también plantearse la situación en la que varias variables se midan simultáneamente sin controlar el valor de ninguna de antemano, como ocurre, por ejemplo, cuando a los pacientes que acuden a una consulta se les pregunta la edad y se les mide la temperatura y la tensión arterial mínima y máxima, con lo cual disponemos de 4 variables aleatorias, pudiendo realizarse regresión de cualquiera de ellas respecto de las restantes; en tal caso se dice que la regresión es de tipo 2.

5.2. REGRESIÓN LINEAL SIMPLE

En esta sección vamos a considerar el caso en que la variable de respuesta Y se explica a partir de una única variable independiente x, y la esperanza condicionada se aproxima mediante una función lineal.

5.2.1. El modelo lineal de regresión simple

Para el desarrollo del modelo partimos de las siguientes hipótesis:

1. Cada variable Y condicionada alos distintos valores de x se distribuye según una ley normal
2. Dichas variables Y/x han de tener igual varianza σ^2
3. Dichas variables Y/x han de ser variables independientes entre si
4. Los valores esperados de la respuesta condicionado a los diferentes valores de x se encuentran sobre una una recta: $E[Y/x] = \beta_0 + \beta_1 x$.

En consecuencia, la respuesta puede expresarse de la forma:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

donde ε es el error aleatorio que representa las desviaciones del modelo lineal respecto de los datos y, en base a las hipótesis anteriores, se distribuye de la forma: $\varepsilon \rightsquigarrow N(0;\sigma)$. Observemos que las tres primeras hipótesis son comunes a las del análisis de la varianza; la novedad radica en la cuarta, en la cual se supone que las medias de la respuesta se encuentran situadas sobre una recta.

La estimación de los parámetros β_0 y β_1 se realiza mediante el método de máxima verosimilitud, según el cual (ver Tema 2) es necesario maximizar el logaritmo de la función de verosimilitud a partir de una muestra $\{(x_i, y_i), i = 1, 2, ..., n\}$:

$$L(x_1, x_2, ..., x_n/\beta_0, \beta_1) = \prod_{i=1}^{n} f(x_i/\beta_0, \beta_1) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\}$$

Evidentemente, maximizar la verosimilitud en β_0 y β_1 equivale a minimizar la función:

$$\Phi(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

por lo que equivale a la estimación por mínimos cuadrados. Así, Derivando parcialmente F respecto de ambos parámetros e igualando a cero se obtiene:

$$\frac{\partial \Phi}{\partial \beta_0} = 0 \Rightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow$$
$$\Rightarrow \sum_{i=1}^{n} y_i - \beta_0 n - \beta_1 \sum_{i=1}^{n} x_i = 0 \Rightarrow \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{\partial \Phi}{\partial \beta_1} = 0 \Rightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\, x_i = 0 \Rightarrow$$
$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \Rightarrow \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

El sistema dado por estas ecuaciones proporciona los siguientes estimadores puntuales de los parámetros del modelo:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

siendo s_x^2 y s_y^2 las varianzas marginales de x e y respectivamente, y s_{xy} la covarianza entre x e y, siendo fácil probar que esta solución verifica la condición de mínimo. Así, la ecuación de la recta de regresión estimada de Y sobre x es:

$$\hat{Y} = \overline{y} + \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

la cual nos permite predecir los valores de la variable Y a partir de observaciones de x.

La diferencias entre los valores observados y_i y los estimados a través del modelo $\hat{y_i}$ se denominan residuos $\hat{\varepsilon}_i = y_i - \hat{y_i}$. A partir de ellos se define la varianza residual como:

$$s_R^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}$$

que es un estimador insesgado de la varianza del error aleatorio del modelo: $Var[\varepsilon] = \sigma^2$. Se puede demostrar la relación: $s_R^2 = s_y^2(1 - R^2)$ siendo $R = \frac{s_{xy}}{s_x s_y}$ el coeficiente de correlación lineal muestral, y claramente se verifica que $s_R^2 \leq s_y^2$. Además, despejando R^2 se obtiene la relación:

$$R^2 = 1 - \frac{s_R^2}{s_y^2}$$

que muestra efectivamente cómo a medida que la correlación entre x e y tiende a ser lineal, la varianza residual s_R^2 se aproxima a cero, por lo que R^2 tiende a ser la unidad. Por el contrario, cuando una relación de tipo lineal no es representativa de la correlación existente entre las variables, la varianza marginal de la variable de respuestas es debida principalmente al residuo, por lo que la relación s_R^2/s_y^2 es próxima a 1 y R^2 tiende a ser 0. De hecho, R^2 se interpreta como el porcentaje de variabilidad de Y debido a x.

Bajo las hipótesis del modelo lineal de regresión simple, el *Teorema de Gauss-Markov* demuestra que los estimadores de los parámetros (coeficientes) del modelo se distribuyen según una ley normal de la forma:

$$\hat{\beta}_0 \rightsquigarrow N(\beta_0; \tfrac{\sigma}{\sqrt{n}}\sqrt{1 + \left(\tfrac{\bar{x}}{s_x}\right)^2}) \ , \quad \hat{\beta}_1 \rightsquigarrow N(\beta_1; \tfrac{\sigma}{s_x\sqrt{n}})$$

por lo que las expresiones de los intervalos de confianza para β_0 y β_1 son:

$$I_\alpha(\beta_0) = \left[\hat{\beta}_0 \pm t_{n-1}^{(\alpha/2)} \tfrac{s_R}{\sqrt{n}}\sqrt{1 + \left(\tfrac{\bar{x}}{s_x}\right)^2}\right] \quad ; \quad I_\alpha(\beta_1) = \left[\hat{\beta}_1 \pm t_{n-1}^{(\alpha/2)} \tfrac{s_R}{s_x\sqrt{n}}\right]$$

ya que al estimar cada parámetro se pierde un grado de libertad. Así mismo, la respuesta media estimada también se distribuye normalmente con parámetros:

$$\hat{\beta}_0 + \hat{\beta}_1 x \rightsquigarrow N(\beta_0 + \beta_1 x; \tfrac{\sigma}{\sqrt{n}}\sqrt{1 + \left(\tfrac{x-\bar{x}}{s_x}\right)^2})$$

que gráficamente da lugar a dos bandas que cuya anchura es mínima cuando $x = \bar{x}$. Así, para un valor $x = x_0$, la expresión del intervalo de confianza para el valor esperado $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ viene dada por:

$$I_\alpha(y_0) = \left[\hat{y}_0 \pm t_{n-2}^{(\alpha/2)} \tfrac{s_R}{\sqrt{n}}\sqrt{1 + \left(\tfrac{x_0-\bar{x}}{s_x}\right)^2}\right]$$

ya que para estimar el valor medio de la respuesta es necesario estimar dos parámetros, por lo que se pierden dos grados de libertad.

En cuanto al contraste de significación del coeficiente de correlación lineal ρ, es decir

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

se resuelve a partir del valor muestral del coeficiente R evaluando el estadístico:

$$t_{\exp} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

que se distribuye según una ley t de *Student* con $n-2$ grados de libertad.

EJEMPLO

El número Y de colonias de bacterias por unidad de volumen presentes en un cultivo después de x horas viene representado en la tabla siguiente:

x :	0	1	2	3	4	5
Y :	12	19	23	34	56	62

Operando resulta:

$$\overline{x} = 2,5 \quad \overline{y} = 34,33 \quad s_x^2 = 2,92 \quad s_y^2 = 349,78 \quad s_{xy} = 31$$

y, a partir de estos cálculos, se obtienen las estimaciones $\hat{\beta}_0 = 7,78$ y $\hat{\beta}_1 = 10,62$, por lo que la ecuación de la recta de regresión estimada de Y sobre x es:

$$\hat{Y} = 7,78 + 10,62x$$

La varianza residual es ahora $s_R^2 = 19,28$ y el coeficiente de determinación $R^2 = 0,94$, lo que muestra que la recta de regresión anterior representa muy bien la relación existente entre las variables.Intervalos de confianza del 95% para los coeficientes del modelo serían:

$$I_{0,05}(\beta_0) = \left[7,78 \pm 2,571\frac{\sqrt{19,28}}{\sqrt{6}}\sqrt{1+\frac{2,5^2}{2,92}}\right] = [7,78 \pm 8,17] = [-0,39; 15,95]$$

$$I_{0,05}(\beta_1) = \left[10,62 \pm 2,571 \frac{\sqrt{19,28}}{\sqrt{2,92 \cdot 6}}\right] = [10,62 \pm 2,70] = [7,92; 13,32]$$

ya que $t_5^{(0,025)} = 2,571$, de donde concluimos que la ordenada en el origen β_0 no es significativa.

Si se quisiera realizar una predicción del número esperado de colonias de bacterias transcurridas 6 horas resultaría un total de 72 colonias, ya que:

$$\hat{Y} = 7,78 + 10,62 \cdot 6 = 71,5$$

y un intervalo de confianza del 95% para dicha predicción sería:

$$I_{0,05}(\beta_0 + \beta_1 x / x{=}6) = \left[71,5 \pm 2,776 \frac{\sqrt{19,28}}{\sqrt{6}} \sqrt{1 + \frac{(6-2,5)^2}{2,92}}\right] = [71,5 \pm 11,34] =$$

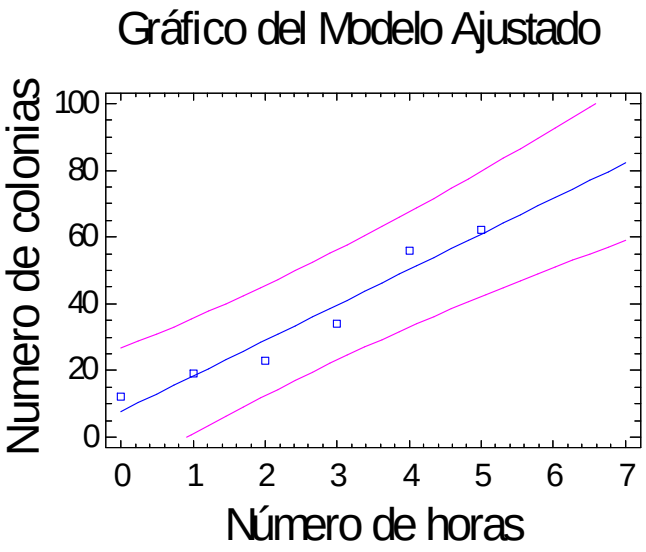
$$= [60,16; 82,84]$$

ya que $t_4^{(0,025)} = 2,776$, por lo que concluimos que al cabo de 6 horas cabe esperar, con una probabilidad de 95%, entre 60 y 83 colonias. Una representación gráfica de la banda de confianza al 95% para la repuesta esperada es la siguiente:

5.2.2. Regresión lineal por el origen

En determinadas situaciones la variable de respuesta es nula si la variable explicativa también lo es; por ejemplo, en un móvil que parte del reposo el desplazamiento es cero en el instante inicial, o la respuesta del organismo es nula si no hay dosis de fármaco. En tales casos, el modelo de regresión debe pasar por el origen de coordenadas y la ecuación de la recta no tendrá ordenada en el origen: $y = \beta x$. La estimación de la pendiente no tiene ahora la misma expresión que en el caso general, sino que se obtiene tras aplicar de nuevo el método de mínimos cuadrados, de forma que el problema consiste en minimizar la función:

$$F(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$



que depende de un único parámetro β. Derivando dicha función e igualando a cero la derivada se obtiene:

$$F'(\beta) = 0 \rightarrow -2\sum_{i=1}^{n} (y_i - \beta x_i)\, x_i = 0 \rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

que se trata de un mínimo ya que $F"(\beta) = \sum_{i=1}^{n} x_i^2 > 0$.

Bajo las hipótesis del modelo lineal de regresión el estimador del coeficiente $\hat{\beta}$ se distribuye según una ley normal de parámetros:

$$\hat{\beta} \rightsquigarrow N(\beta; \frac{\sigma}{\sum_{i=1}^{n} x_i^2})$$

y la expresión del intervalo de confianza para β es:

$$I_\alpha(\beta) = [\hat{\beta} \pm t_{n-1}^{(\alpha/2)} \frac{s_R}{\sqrt{\sum_{i=1}^{n} x_i^2}}]$$

EJEMPLO

En un estudio sobre disminución de temperatura corporal (Y) en pacientes con fiebre alta en función de la dosis (x) de antipirético administrada se han obtenido los datos siguientes:

$x(mg):$	100	200	300	400	500	600
$Y(^oC):$	0,2	0,5	0,7	1,0	1,4	1,6

Para estimar el correspondiente modelo lineal m.c. que pasa por el origen basta realizar el siguiente cálculo:

$$\hat{\beta} = \frac{2390}{910000} = 0,0026$$

Por lo que la estimación de Y a partir de x es $\hat{Y} = 0,0026x$. Así mismo, la varianza residual es $s_R^2 = 0,00472$ y un intervalo de confianza del 95% para β vendrá dado por:

$$I_{0,05}(\beta) = \left[0,0026 \pm 2,571\frac{0,069}{\sqrt{910000}}\right] = [0,0026 \pm 0,00019] = [0,00241; 0,00279]$$

5.3. REGRESIÓN LINEAL MÚLTIPLE

5.3.1. El modelo lineal de regresión múltiple

La extensión del caso simple al múltiple, aunque sobre una base común, presenta algunas peculiaridades consecuencia de la introducción de varias variables. Las hipótesis sobre las que se fundamenta son las siguientes:

1. Cada variable Y condicionada a los distintos valores de las variables explicativas x_i se distribuyesegú n una ley normal
2. Dichas variables $Y/(x_1, x_2, ..., x_p)$ han de tener igual varianza σ^2
3. Dichas variables $Y/(x_1, x_2, ..., x_p)$ han de ser independientes entre si
4. Los valores esperados de la respuesta condicionado a los diferentes valores de las x_i se encuentran sobre un hiperplano:
   $E[Y/x_1, x_2, ..., x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$
5. No pueden existir relaciones lineales entre las variables explicativas x_i
6. El número de datos ha de ser superior al de variables p.

La violación de la quinta hipótesis provoca un interesante y frecuente problema que se denomina multicolinealidad. En caso de existir una tal relación lineal entre dos variables la dimensión del problema se reduciría en una unidad. En cuanto a la última hipótesis, un tamaño muestral inferior a $p+1$ generaría problemas de singularidad en las matrices.

Bajo las hipótesis anteriores, el procedimiento de estimación por máxima versosimilitud daría lugar a una expresión del tipo:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

donde $\mathbf{X}'$ representa la matriz traspuesta de $\mathbf{X}$, siendo

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

Además, el teorema de Gauss-Markov afirma que la distribución del vector de estimadores de los parámetros es normal multivariante con media el vector $\boldsymbol{\beta}$ y matriz de varianzas-covarianzas $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. Más aún, se demuestra que:

$$\tfrac{1}{\sigma^2} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' (\mathbf{X}'\mathbf{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \rightsquigarrow \chi^2_{p+1}$$

Cuando, en el caso múltiple, se quiere estimar el grado de dependencia de una variable Y respecto de dos independientes x_1 y x_2, se introduce el denominado *coeficiente de correlación múltiple* de la forma:

$$R_{Y \cdot x_1 x_2} = \sqrt{\frac{R^2_{Y x_1} + R^2_{Y x_2} - 2 R_{Y x_1} R_{Y x_2} R_{x_1 x_2}}{1 - R^2_{x_1 x_2}}}$$

siendo $R_{Y x_1}, R_{Y x_2}$ y $R_{x_1 x_2}$ los correspondientes coeficientes de correlación lineal simple entre $(Y, x_1), (Y, x_2)$ y (x_1, x_2) respectivamente. Su cuadrado se denomina *coeficiente de determinación múltiple* y se representa simplemente R^2 $(0 \leq R^2 \leq 1)$.

Es interesante en esta situación conocer el grado de correlación existente entre la variable dependiente Y y cada una de las independientes de forma aislada, es decir, excluyendo el efecto de la otra. Se introducen así los *coeficientes de correlación parcial* de la forma:

$$r_{Yx_1 \cdot x_2} = \frac{r_{Yx_1} - r_{Yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{Yx_2}\right)\left(1 - r^2_{x_1 x_2}\right)}} \qquad r_{Yx_2 \cdot x_1} = \frac{r_{Yx2} - r_{Yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{Yx_1}\right)\left(1 - r^2_{x_1 x_2}\right)}}$$

Igual que en el caso simple, el contraste de significación sobre R^2 se resuelve mediante el estadístico del contraste:

$$t_{\text{exp}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

que se distribuye según una ley t de *Student* con $n-p-1$ grados de libertad, siendo p el número de variables explicativas del modelo.

Todo este planteamiento múltiple puede extenderse sin dificultad al caso en que se presenten más de dos variables independientes en el estudio. En general, a medida que se introducen más variables en el modelo aumenta el valor de R^2 aunque éstas no tengan un aporte significativo al mismo. Así, se introduce una modificación a dicho coeficiente de manera que, si se dispone de n muestras, se define el coeficiente de determinación lineal *corregido por grados de libertad* de la forma:

$$R^2_{corregido} = (1 - \tfrac{p}{n-p-1})R^2$$

De manera similar, los contrastes de significación de los coeficientes de correlación parcial en regresión múltiple se resuelven evaluando respectivamente los estadísticos:

$$t_{\text{exp}} = \frac{r_{Yx_1 \cdot x_2}\sqrt{n-3}}{\sqrt{1-r^2_{Yx_1 \cdot x_2}}} \qquad t_{\text{exp}} = \frac{r_{Yx_2 \cdot x_1}\sqrt{n-3}}{\sqrt{1-r^2_{Yx_2 \cdot x_1}}}$$

que siguen una distribución t de *Student* con $n-3$ grados de libertad.

EJEMPLO

Se ha medido la estatura (Y) de un grupo de niños a partir del peso (x_1) y la edad (x_2) obteniéndose los datos siguientes:

Edad (meses)	0	3	6	9	12	15	18	24
Peso (kg)	3,4	5,6	7,3	8,6	9,5	11	11,5	12,4
Estatura (cm)	50,3	59	65	70	74	77	80,5	86

Las matrices de diseño son las siguientes:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 3,4 \\ 1 & 3 & 5,6 \\ 1 & 6 & 7,3 \\ 1 & 9 & 8,6 \\ 1 & 12 & 9,5 \\ 1 & 15 & 11 \\ 1 & 18 & 11,5 \\ 1 & 24 & 12,4 \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} 50,3 \\ 59 \\ 65 \\ 70 \\ 74 \\ 77 \\ 80,5 \\ 86 \end{pmatrix}$$

Así

$$\mathbf{X'X} = \begin{pmatrix} 8 & 87 & 69,3 \\ 87 & 1395 & 921,6 \\ 69,3 & 921,6 & 6676,43 \end{pmatrix} \qquad \mathbf{X'Y} = \begin{pmatrix} 561,8 \\ 6753 \\ 5120,07 \end{pmatrix}$$

por lo que:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 8 & 87 & 69,3 \\ 87 & 1395 & 921,6 \\ 69,3 & 921,6 & 6676,43 \end{pmatrix}^{-1} \begin{pmatrix} 561,8 \\ 6753 \\ 5120,07 \end{pmatrix} = \begin{pmatrix} 40,96 \\ 0,32 \\ 2,98 \end{pmatrix}$$

y la estimación de la estatura a partir de x_1 y x_2 viene dada por:

$$\hat{Estatura} = 40,96 + 0,32 \cdot Peso + 2,98 \cdot Edad$$

A su vez, los coeficientes de correlación lineal entre las variables son:

$$R_{Yx_1} = 0,978 \qquad R_{Yx_2} = 0,996 \qquad R_{x_1x_2} = 0,968$$

por lo que: $R_{Y \cdot x_1x_2} = 0,986$. Así, como

$$t_{\exp} = \frac{0,986\sqrt{6}}{\sqrt{1-0,971}} = 14,183$$

se concluye que el coeficiente de determinación es significativo ya que el valor crítico es $t_6^{(0,025)} = 2,447$.

5.3.2. Complementos sobre regresión múltiple

Coeficientes estandarizados

La significación de cada coeficiente de regresión estimado a partir de observaciones muestrales indica la influencia de la correspondiente variable explicativa sobre las respuesta. Así, aquellas variables x_i cuyos coeficientes β_i tengan asociado un p-valor superior al nivel de significación establecido (usualmente α=0,05), no tendrán aporte significativo sobre la respuesta del modelo Y. Pero el problema consiste en interpretar el valor del coeficiente, ya que éste dependerá de las unidades en las que se expresa la correspondiente variable. Con objeto de reducir a unidades estándar los coeficientes, y poder comparar así sus valores en una misma escala, se recurre a los métodos de estandarización o escalamiento, siendo el más simple el denominado método *normal unitario*, según el cual, a partir de un modelo estimado:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

si denotamos respectivamente por $\bar{x}_j$ e $\bar{y}$ a las medias muestrales de cada variable regresora x_j y de la respuesta y, e, igualmente, por s_j^2 y s_y^2 a sus cuasivarianzas muestrales, entonces se realizan las tipificaciones siguientes:

$$z_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j} \qquad Y_i^* = \frac{y_i-\bar{y}}{s_y}\ , \ \ i = 1, 2, \cdots n, \quad j = 1, 2, , , , , k$$

dando lugar a un nuevo modelo de regresión múltiple, sin ordenada en el origen, de la forma:

$$Y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \varepsilon_i \qquad i = 1, 2, \cdots n$$

El estimador mínimo cuadrático del vector de coeficientes $\mathbf{b} \equiv (b_1 b_2 ... b_k)'$ es:

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z})^{-1}\,\mathbf{Z}'\mathbf{Y}^*$$

siendo $\mathbf{Z}$ la matriz de los z_{ij} e $\mathbf{Y}^*$ el vector de los Y_i^*.

EJEMPLO

Se quiere estimar un modelo lineal de regresión que exprese la estatura de un niño con edad no superior a 2 años en función de su edad y peso a partir de los siguientes datos muestrales:

Y : Estatura (cm)	$50,3$	59	65	70	74	77	$80,5$	86
x_1 : Edad (meses)	0	3	6	9	12	15	18	24
x_2 : Peso (kg)	$3,4$	$5,6$	$7,3$	$8,6$	$9,5$	11	$11,5$	$12,4$

El procedimiento de estimación descrito en el subapartado 5.3.1 proporciona la siguiente ecuación, donde el p-valor del estimador de cada coeficiente aparece entre paréntesis:

$$\begin{array}{c} \hat{Y} = 40,956 + 0,319x_1 + 2,979x_2 \\ \quad (0,000)\ (0,1412) \quad (0,0015) \end{array}$$

siendo $R^2 = 99,51\%$ y $R^2_{corregido} = 99,31\%$. La interpretación sería que la edad no tiene influencia significativa sobre la estatura, pero, con independencia de eso, el peso tiene un aporte sobre las estatura de más de 9 veces superior a la edad, pues $2,979/0,319 = 9,34$. Esta interpretación no es correcta, pues el peso se expresa en kilos y la edad en meses. Así, procediendo a estandarizar los coefiientes, se obtendría como modelo estimado:

$$\hat{Y}^* = 0,217x_1 + 0,786x_2$$

y ahora ya sí sería correcto afirmar que la influencia del peso sobre la estatura es, aproximadamente, tres veces y media superior a la de la edad.

Multicolinelidad

La hipótesis 5 del modelo lineal de regresión múltiple establece que no puede haber relaciones lineales entre las variables explicativas x_i ya que, en caso de existir, la matriz $\mathbf{X'X}$ sería singular y no podría invertirse; además, en tal caso, la dimensión del problema se reduciría en una unidad. Supongamos que en un modelo $Y = \beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$, la variable x_2 depende linealmente de x_1 según: $x_2 = a + bx_1$; entonces:

$$Y = \beta_0 + \beta_1x_1 + \beta_2\left(a + bx_1\right) + \varepsilon = \left(\beta_0 + \beta_2a\right) + \left(\beta_1 + \beta_2b\right)x_1 + \varepsilon$$

Este problema se conoce como *multicolineallidad*. En la práctica, esta situación no va a plantearse casi nunca, pero basta con que exista una correlación significativa entre dos variables para que puede hablarse de multicolinealidad. De hecho, en el ejemplo anterior, ésta era la situación entre peso

y edad, ya que su coeficiente de determinación es $R^2 = 93,7\%$, que es completamente significativo, y lo que explica que la edad no fuera una variable significativa, pues su información había sido absorbida por el peso.

La multicolinealidad es un problema serio que hay que evitar en regresión que, además, produce un efecto de *inflacción* de las varianzas de los estimadores, ya que en los coeficientes estandarizados se verifa entonces que $\frac{Var[\hat{b}_i]}{\sigma^2} > 1$.

Selección de variables

Al estimar un modelo de regresión lineal múltiple puede ocurrir que algunas variables no tengan aporte significtivo sobre la respuesta, lo que se pone de manifiesto en que el p-valor asociado a la estimación del parámetro correspondiente está por encima del nivel de significación. En el ejemplo anterior, en el que se pretendía estimar la estatura de los niños menores de 2 años a partir de su edad y peso, resulto no significativa la edad debido a un problema de multicolinealidad con la variable peso, por lo que el modelo final debería prescindir de dicha variable.

Pero este problema se complica a medida que el conjunto de variables explicativas es mayor, por lo que no es correcto eliminar directamente todas aquéllas cuyo p-valor exceda el nivel de significación, ya que puede ocurrir que al suprimir algunas, otreas que en principio no eran significativas pasen a serlo.

EJEMPLO

Se quiere estimar mediante un modelo lineal de regresión múltiple el beneficio anual de una empresa farmacéutica (Y) en función del volumen de activos (x_1), sueldo de los empleados (x_2) y del importe de las materias primas (x_3), todos expresado en miles de euros . Los datos recogidos de una muestra de 15 empresas fueron los siguientes:

Y	249	3334	707	477	142	301	109	167	100	84	119	35
x_1	454	2612	542	535	137	227	100	124	81	67	100	46
x_2	3358	15230	7391	6306	2075	3517	874	1267	894	978	1350	1302
x_3	166	1209	119	91	34	70	16	37	14	20	15	16

Tras estimar por mínimos cuadrados los parámetros del modelo resulta:

$$\begin{array}{c}\hat{Y} = -41,119 - 0,458x_1 + 0,084x_2 + 2,703x_3 \\ \quad (0,4018) \quad (0,6059) \quad (0,1269) \quad (0,0681)\end{array}$$

donde los p-valores figuran entre paréntesis bajo cada coeficiente estimado. El coeficiente de determinación corregido es $R^2_{corregido} = 98,82\%$. Así, la eliminación directa de x_1 y x_2, además de la constante, proporcionaría un modelo incompleto, ya que incluiría sólo la variable x_3, es decir, sería:

$$\begin{array}{c}\hat{Y} = 2,788x_3 \\ \quad (0,0000)\end{array}$$

siendo ahora $R^2_{corregido} = 97,77\%$. El modelo óptimo en este caso vendría dado por:

$$\begin{array}{c}\hat{Y} = 0,049x_2 + 2,116x_3 \\ \quad (0,0012) \quad (0,0000)\end{array}$$

Existen diversos procedimientos de selección de variables en la estimación del modelo cuyo estudio detallado queda fuera del alcance de este texto. Los más conocidos son los siguientes:

Método *forward* (hacia adelante). Consiste en ir introduciendo una a una variables en el modelo, contrastando en cada paso si la variable introducida es o no significativa.

Método *backard* (hacia atrás). A diferencia del anterior, en este método se consideran inicialmente todas las variables explicativas y se van eliminando de una en una, comenzando por la menos significativa (la que tiene un p-valor más alto) y analizando la significación de las que van quedando en el modelo, ya que, como se ha indicado, puede ocurrir que al eliminar una, otra que inicialmente no era significativa pase a serlo.

Método *stepwise* (paso a paso). Puede realizarse tanto hacia adelanta como hacia atrás y consiste en la inclusión o exclusión de las variables en el modelo de una manera secuencial, es decir, cada vez que se introduce o elimina una nueva variable, se contrasta globalmente la significación del modelo completo.

5.4. REGRESIÓN POLINÓMICA

Una extensión de la regresión lineal múltiple es la denominada regresión parabólica o polinómica, en la cual el ajuste de polinomios de grado mayor que uno, es decir, $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$, se realiza de manera similar, calculando el valor mínimo de la función:

$$\Phi\left(\beta_0, \beta_1, \beta_2, \ldots, \beta_p\right) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 - \cdots - \beta_p x_i^p\right)^2$$

Derivando parcialmente Φ respecto de cada una de los parámetros β_i, igualando a cero, y operando convenientemente, se obtiene el siguiente sistema de Cramer de orden $p+1$:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 + \cdots + \beta_p \sum_{i=1}^{n} x_i^p = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 + \cdots + \beta_p \sum_{i=1}^{n} x_i^{p+1} = \sum_{i=1}^{n} x_i y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^4 + \cdots + \beta_p \sum_{i=1}^{n} x_i^{p+2} = \sum_{i=1}^{n} x_i^2 y_i$$

.....

$$\beta_0 \sum_{i=1}^{n} x_i^p + \beta_1 \sum_{i=1}^{n} x_i^{p+1} + \beta_2 \sum_{i=1}^{n} x_i^{p+2} + \cdots + \beta_p \sum_{i=1}^{n} x_i^{2p} = \sum_{i=1}^{n} x_i^p y_i$$

que se denomina ecuaciones normales del ajuste parabólico. La expresión de la varianza residual, como estimador insesgado de la varianza del modelo, es entonces:

$$s_R^2 = \frac{\sum_{i=1}^{n} y_i^2 - \beta_0 \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i y_i - \beta_2 \sum_{i=1}^{n} x_i^2 y_i - \cdots - \beta_p \sum_{i=1}^{n} x_i^p y_i}{n - p - 1}$$

En ocasiones, los modelos polinómicos se tratan como casos particulares del modelo lineal, considerando las variables $x_i = x^i, i = 1, 2, ..., p$, si bien éstas no tienen por qué ser independientes entre si. De hecho, al no cumplirse las hipótesis del modelo lineal de regresión, la inferencia requiere métodos especiales que quedan fuera del alcance de este manual (Montgomery *et al.*, 2005)

EJEMPLO

A presión de 1 atmósfera un volumen de agua disuelve las cantidades de CO_3H_2 reflejadas en la tabla siguiente:

t:	0^o	5^o	10^o	15^o
V:	1,8	1,45	1,18	1

Ajustar a estos datos una parábola de segundo grado y hallar la varianza residual asociada.

t_i	V_i	t_iV_i	t_i^2	V_i^2	$t_i^2V_i$	t_i^3	t_i^4
0	1,8	0	0	3,24	0	0	0
5	1,45	7,25	25	2,1025	36,25	125	625
10	1,18	11,8	100	1,3924	118	1000	10000
15	1	15	225	1	225	3375	50625
30	5,43	34,05	350	7,7349	379,25	4500	61250

El sistema es entonces:

$$\left.\begin{array}{c} 4\beta_0 + 30\beta_1 + 350\beta_2 = 5,43 \\ 30\beta_0 + 350\beta_1 + 4500\beta_2 = 34,05 \\ 350\beta_0 + 4500\beta_1 + 61250\beta_2 = 379,25 \end{array}\right\} \Rightarrow \begin{array}{c} \hat{\beta}_0 = 1,8005 \\ \hat{\beta}_1 = -0,0789 \\ \hat{\beta}_2 = 0,0017 \end{array}$$

Luego la parábola de ajuste por m.c. es:

$$\hat{Y} = 1,8005 - 0,0789x + 0,0017x^2$$

La varianza residual viene dada por:

$$\begin{aligned} s_R^2 &= \frac{7,7349 - (1,8005)(5,43) - (0,0789)(34,05) - (0,0017)(379,25)}{4-3} = \\ &= 4,84 \cdot 10^{-6} \end{aligned}$$

5.5. REGRESIÓN NO LINEAL

En general, cuando se pretende ajustar una función no polinómica a unos datos experimentales por el método de los mínimos cuadrados, el sistema de ecuaciones resultante no es de tipo lineal, y su resolución directa suele presentar dificultades. Por esta razón, la técnica usual aplicable a estos casos consiste en transformar la ecuación teórica del ajuste en una recta, y, tras realizar el ajuste lineal de los datos (también transformados), deshacer los cambios efectuados. Por supuesto, al igual que ocurre con la regresión polinómica, no se cumplen las hipótesis generales del modelo lineal siendo necesario enunciar otras alternativas para poder realizar inferencia sobre estos modelos no lineales.

Las transformaciones más frecuentes que se presentan son las siguientes:

Exponencial: $y = e^{b_0+b_1x}$

Se realiza tomando logaritmos neperianos, con lo cual es: $\ln y = b_0 + b_1x$.

Multiplicativo: $y = b_0x^{b_1}$, $b_0 > 0$, $x_i > 0$.

Se realiza la misma transformación anterior:

$$\ln y = \ln b_0 + b_1 \ln x \Rightarrow Y = B_0 + b_1X$$

Inverso de x: $y = b_0 + \dfrac{b_1}{x}$, $x_i \neq 0$.

Basta considerar:

$$y = b_0 + b_1\frac{1}{x} \Rightarrow y = b_0 + b_1X$$

Inverso de y: $y = \dfrac{1}{b_0 + b_1x}$, $y_i \neq 0$.

Basta considerar:

$$\frac{1}{y} = b_0 + b_1x \Rightarrow Y = b_0 + b_1X$$

Michaeliana: $y = \dfrac{b_0x}{b_1 + x}$, $x_i \neq 0$, $y_i \neq 0$.

Se realiza el siguiente cambio:

$$\frac{1}{y} = \frac{b_1 + x}{b_0 x} = \frac{b_1}{b_0}\frac{1}{x} + \frac{1}{b_0} \Rightarrow Y = B_0 + B_1 X$$

Así: $b_0 = \dfrac{1}{B_0}$, $b_1 = \dfrac{B_1}{B_0}$.

Logística: $y = \dfrac{1}{1 + b_0 e^{b_1 x}}$, $b_0 > 0$, $b_1 < 0$, $0 < y_i < 1$.

Se realiza el siguiente cambio:

$$\begin{aligned} \frac{1}{y} &= 1 + b_0 e^{b_1 x} \Rightarrow \frac{1}{y} - 1 = b_0 e^{b_1 x} \Rightarrow \ln\left(\frac{1}{y} - 1\right) = \ln b_0 + b_1 x \\ &\Rightarrow Y = B_0 + b_1 x \end{aligned}$$

Ejercicios Resueltos

1^o) El muestreo de áreas contiguas se utiliza en Ecología para contar el número de especies distintas de plantas por área. El recuento se realiza de manera que cada área contigua tiene doble extensión que la anterior, empezando por un área de $1m^2$. El modelo que relaciona el número de especies N con el área S es $N = b_0 + b_1 \ln S$, donde b_0 es el número de especies por área unidad y b_1 un índice de diversidad. Calcular b_0 y b_1 para los siguientes datos:

S (en m^2):	1	2	4	8	16	32	64
N (num. esp.):	2	4	7	11	16	19	21

Analizar si serían también adecuados los modelos: $N = b_0 e^{b_1 S}$ y $N = b_0 S^{b_1}$.

Resolución:

Denotemos x_i a los valores de la variable independiente S, e y_i a los de la variable dependiente N. Así se tiene:

$$\sum_{i=1}^{n} x_i = 127 \qquad \sum_{i=1}^{n} x_i^2 = 5461 \qquad \sum_{i=1}^{n} y_i = 80 \qquad \sum_{i=1}^{n} y_i^2 = 1248$$

$$\sum_{i=1}^{n} \ln x_i = 14,56 \ \sum_{i=1}^{n} (\ln x_i)^2 = 43,72 \ \ \sum_{i=1}^{n} \ln y_i = 15,18 \ \ \sum_{i=1}^{n} (\ln y_i)^2 = 37,56$$

$$\sum_{i=1}^{n} y_i \ln x_i = 232,9 \quad \sum_{i=1}^{n} x_i \ln y_i = 363,87 \quad \sum_{i=1}^{n} (\ln x_i)(\ln y_i) = 39,2$$

Las ecuaciones normales del modelo son

$$\left.\begin{array}{c} 7b_0 + 14,56b_1 = 80 \\ 14,56b_0 + 43,27b_1 = 232,9 \end{array}\right\} \Rightarrow b_0 = 1,13;\ b_1 = 4,95$$

Por lo tanto, la ecuación de ajuste es: $N = 1,13 + 4,95 \ln S$, y la varianza residual es: $s_R^2 = 0,009$.

Para el modelo exponencial se obtienen los valores siguientes:

$$b_0 = 5,26;\ b_1 = 0,028$$

es decir, $N = 5,2e^{0,028S}$, siendo la varianza residual: $s_R^2 = 0,3091$; y para el modelo potencial es:

$$b_0 = 2,69;\ b_1 = 0,57$$

por lo que: $N = 2,69S^{0,57}$, siendo la varianza residual: $s_R^2 = 0,0441$.

Se observa, por tanto, que, aunque estos dos modelos son bastante aceptables, el modelo logarítmico se ajusta en mayor medida a los datos experimentales.

2^o) En estudios sobre estabilidad de fármacos se realizan ensayos en condiciones ambiente cada cierto periodo de tiempo. Además, el estudio debe realizarse sobre más de un lote de productos a fin de evitar irregularidades debidas al efecto de los aditivos, o a otras causas. Supongamos entonces que un producto farmacéutico ha sido etiquetado con una fecha de caducidade de 15 meses (desde su envase), y con una cantidad de principio activo de 50 mg. Supongamos, asimismo, que como margen de seguridad se ha preparado con un 4 por 100 de exceso de principio activo, teniendo, por tanto, en realidad 52 mg. Para analizar su actividad se realizan ensayos sobre tres lotes a los

3, 6, 9, 12 y 18 meses, obteniéndose los resultados de la tabla siguiente:

Tiempo (*meses*)	Actividad (*mg*)	Actividad media (*mg*)
0	51 51 53	51,7
3	51 50 52	51
6	50 52 48	50
9	49 51 51	50,3
12	49 48 47	48
18	47 45 49	47

Sabiendo que este tipo de fármacos se consideran eficaces si mantienen al menos un 95 por 100 de actividad, y que la ecuación cinética de pérdida puede considerarse de tipo lineal: $C(t) = C_0 - kt$, siendo $C(t)$ y C_0 las cantidades de fármaco en los instantes t e inicial, respectivamente, y k la tasa de descomposición, se desea comprobar si la fecha marcada de caducidad es correcta.

Resolución:

Matemáticamente, el problema se reduce a ajustar una recta a los puntos $(0; 51,7)$, $(3; 51)$, $(6; 50)$, $(9; 50,3)$, $(12; 48)$ y $(18; 47)$. Tras realizar las correspondientes operaciones resulta:

$$C(t) = 51,8 - 0,267t$$

siendo la varianza residual: $s_R^2 = 0,2025$. Así, para $t = 15$ resulta $C(15) = 47,795$ *mg* y como el fármaco es eficaz si mantiene una actividad igual a $50 \cdot 95\% = 47$ *mg*, concluimos que la fecha de caducidad señalada es correcta.

Resulta evidente que, si seleccionamos otra muestra distinta, el resultado final podría llegar a diferir notablemente. Esto constituye el principal defecto de los modelos deterministas, ya que mediante ellos sólo podemos obtener estimaciones puntuales del valor verdadero, pero sin disponer de medida alguna sobre la incertidumbre asociada a todo experimento regido por leyes de la Naturaleza.

3º) En el equilibrio líquido⇌gas del metanol se verifica la siguiente relación termodinámica:

$$\log k = \frac{\Delta S}{4,576} - \frac{\Delta H}{4,576}\frac{1}{T}$$

donde ΔS y ΔH son, respectivamente, las variaciones de la entropía y la entalpía durante la reacción (medidas en kcal/mol o cal/g), k es la constante de

equilibrio y T es la temperatura absoluta. Se pide ajustar dicha relación si las medidas repetidas en la presión de vapor de alcohol metílico correspondientes a ocho temperaturas distintas han sido las siguientes:

T^{o}C	17^o	22^o	27^o	32^o	37^o	42^o	47^o	52^o
1	9,6	12	15,1	18,1	25	28,9	36,5	48,6
2	10,1	12,7	15,8	19	23,4	28,2	34,8	47,5
3	9,2	11,8	14,7	18,5	24	29,6	35,7	44,3
4	9,8	13	15,5	19,5	24,5	30,3	38,1	46,4
5	9,5	12,4	14,5	19,9	25,7	31	37,3	45,3

Resolución:

Observemos que denotando: $x = \dfrac{1}{T}$, $y = \log k$, $b_0 = \dfrac{\Delta S}{4,576}$, $b_1 = -\dfrac{\Delta H}{4,576}$, la ecuación anterior queda reducida a una de tipo lineal: $y = b_0 + b_1 x$. Por tanto, el ajuste consiste en construir la recta que represente $\log P_i$ frente a $\dfrac{1}{T}$ (que para evitar cantidades excesivamente paqueñas consideraremos $\dfrac{10^3}{T}$):

T^oK	$\dfrac{1}{T}10^3$	$y_i = \overline{\log P_i}$	$\dfrac{1}{T}10^3\overline{\log P_i}$
290	3,44649	0,98386	3,3908636
295	3,38811	1,09244	3,7013068
300	3,33167	1,17874	3,9271726
305	3,27708	1,27852	4,1898123
310	3,22425	1,38928	4,4793860
315	3,17309	1,47104	4,6677423
320	3,12354	1,56184	4,8784697
325	3,07550	1,66646	5,1251977
$\sum$	26,03973	10,62218	34,359949

Tras realizar los cálculos oportunos del ajuste lineal se obtiene la relación:

$$y = 7,25975 - 1,82244 \cdot 10^3 x$$

Como $b_1 = -1,82244 \cdot 10^3 = -\dfrac{\Delta H}{4,576}$, resulta:

$$\Delta H = -4,576 b_1 = -8,34 \cdot 10^3$$

Así, el calor de evaporación del metanol, que viene expresado por la variación de entalpía en el punto de equilibrio entre los estados líquido y gaseoso, es igual a 8,34 kcal.

4^o) Se ha estudiado la pérdida de actividad de un fármaco a lo largo de 2 meses, obteniendo los datos de la tabla siguiente:

$t(días):$	0	10	30	60
$C(mg.):$	500	396,85	250	125

Se pide ajustar un modelo exponencial del tipo $C = b_0 e^{-b_1 t}$ y estudiar su precisión.

Resolución:

Aplicando la transformación logarítmica se obtiene: $lnC = lnb_0 - b_1 t$. Así los datos transformados vienen dados en la tabla siguiente:

$t:$	0	10	30	60
$y = \ln C:$	6,215	5,984	5,522	4,828

Dado que $\overline{t} = 25 \quad s_t^2 = 525 \quad \overline{y} = 5,637 \quad s_y^2 = 0,280 \quad s_{ty} = -12,130$, las estimaciones son:

$$\begin{aligned} -b_1 &= \frac{-12,13}{525} = -0,023 \rightarrow b_1 = 0,023 \\ \ln b_0 &= 5,637 - (-0,023) \cdot 25 = 6,215 \rightarrow b_0 = e^{6,215} = 500 \end{aligned}$$

por lo que el modelo de desintegración es: $C(t) = 500e^{-0,023t}$, y su precisión viene dada por: $r^2 = \frac{(-12,12)^2}{525.0,28} = 0,999$.

5^o) En un laboratorio se ha observado la reproducción de un cierto parásito partiendo de un número inicial de 30. Los resultados al cabo de 10 meses han sido los siguientes:

$meses(x):$	0	1	2	3	5	10
$num.\ parásitos(y):$	30	50	82	135	365	4452

Se pide ajustar un modelo exponencial $y = e^{b_0 + b_1 x}$ y estudiar su precisión. Estimar, asimismo, el número esperado de parásitos al cabo de 4 meses.

Resolución:

Aplicando la transformación logarítmica se obtiene: $lny = b_0 + b_1 x$. Así los datos transformados vienen dados en la tabla siguiente:

$$\begin{array}{rcccccc} x: & 0 & 1 & 2 & 3 & 5 & 10 \\ Y = \ln y: & 3,4 & 3,9 & 4,4 & 4,9 & 5,9 & 8,4 \end{array}$$

Dado que $\overline{x} = 3,5 \quad s_x^2 = 10,92 \quad \overline{Y} = 5,15 \quad s_Y^2 = 2,73 \quad s_{xY} = 5,46$, las estimaciones son:

$$b_1 = \frac{5,46}{10,92} = 0,5 \quad ; \quad b_0 = 5,15 - 0,5 \cdot 3,5 = 3,4$$

por lo que el modelo de multiplicación de parásitos es: $y = e^{3,4+0,5x}$, y su precisión viene dada por: $r^2 = \frac{(5,46)^2}{10,92.2,73} = 0,999$.

Al cabo de 4 meses se espera que haya: $y(4) = 222$ parásitos.

6^o) La ley de Boyle-Mariotte establece que, a temperatura constante, la presión ejercida P y el volumen V que ocupa un gas verifican la relación: $P.V = C$. Hallar el valor de la constante C para un sistema en el que se han obtenido las siguientes medidas:

$$\begin{array}{rcccc} P(Kg/cm^2): & 0,10 & 0,15 & 0,20 & 0,25 \\ V(litros): & 2,24 & 1,50 & 1,13 & 0,92 \end{array}$$

Resolución:

Según la ley de Boyle-Mariotte, el volumen es directamente proporcional al inverso la la presión: $P = C\frac{1}{V}$. Se trata, pues, de un caso particular de regresión lineal por el origen a partir de los datos transformados:

$$\begin{array}{rcccc} x = 1/P: & 10 & 6,66 & 5 & 4 \\ V: & 2,24 & 1,50 & 1,13 & 0,92 \end{array}$$

Por tanto: $C = \dfrac{\sum_{i=1}^{4} x_i P_i}{\sum_{i=1}^{4} x_i^2} = \frac{41,73}{185,44} = 0,225$

7^o) En Termodinámica se denomina proceso adiabático a aquél en el cual el sistema no intercambia calor con su entorno. En tal caso, la relación entre

la presión P y el volumen V viene dada por: $PV^a = b$. Se pide estimar las constantes a y b a partir de la siguiente tabla de datos observados, así como la presión que se ejerce sobre el gas cuando éste ocupa un volumen de 100 litros.

$V(litros):$	$54,3$	$61,8$	$72,4$	$88,7$	$118,6$	$194,0$
$P(atmósferas):$	$61,2$	$49,5$	$37,6$	$28,4$	$19,2$	$10,1$

Resolución:

Aplicando la transformación logarítmica se obtiene: $\log P = \log b - a \log V$. Así los datos transformados vienen dados en la tabla siguiente:

$X = \log V:$	$1,735$	$1,791$	$1,859$	$1,948$	$2,074$	$2,288$
$Y = \log P:$	$1,787$	$1,695$	$1,575$	$1,453$	$1,283$	$1,004$

Dado que $\overline{X} = 1,949$ $s_X^2 = 0,035$ $\overline{Y} = 1,466$ $s_Y^2 = 0,069$ $s_{XY} = -0,049$, las estimaciones son:

$$\begin{aligned} a &= \frac{-0,049}{0,035} = -1,404 \\ \log b &= 1,466 - (-1,404) \cdot 1,949 = 4,203 \rightarrow b = 15961,278 \end{aligned}$$

por lo que la relación adiabática viene dada por: $PV^{-1,404} = 15961,278$, y la presión estimada correspondiente a 100 litros es: $P(100) = \frac{15961,278}{100^{1,404}} = 24,835$ atmósferas.

8^o) Se quiere estudiar la relación existente entre la ingestión de grasas (x) y el nivel de colesterol en sangre (y), para lo cual se tomaron muestras de 5 pacientes, obteniendo los siguientes resultados:

$x:$	$7,4$	$8,5$	9	11	13
$y:$	38	25	$35,75$	$57,85$	$83,5$

Obtener el modelo multiplicativo de regresión: $y = b_0 x^{b_1}$ de la variable y a partir de la x y estudiar su precisión

Resolución:

Para linealizar el modelo multiplicativo es necesario aplicar la transformación logarítmica: $\log y = log b_0 + b_1 \log x$. Así la tabla de datos transformados es:

$$\begin{array}{lccccc} X = \log x: & 0,869 & 0,929 & 0,954 & 1,041 & 1,079 \\ Y = \log y: & 1,580 & 1,398 & 1,577 & 1,762 & 1,922 \end{array}$$

Dado que $\overline{X} = 0,9747$ $s_X^2 = 0,0058$ $\overline{Y} = 1,6477$ $s_Y^2 = 0,0320$ $s_{XY} = 0,0113$ las estimaciones son:

$$\begin{aligned} b_1 &= \frac{0,0113}{0,0058} = 1,943 \\ \log b_0 &= 1,6477 - 1,9426 \cdot 0,9747 = -0,246 \rightarrow b_0 = 0,568 \end{aligned}$$

por lo que el modelo multiplicativo es: $y = 0,568x^{1,943}$, y su precisión viene dada por: $r^2 = \frac{(0,0113)^2}{0,0058.0,032} = 0,682$.

Problemas Propuestos

1º) Un preparado hormonal va perdiendo actividad a lo largo del tiempo según muestra la tabla adjunta:

$Tiempo(meses):$	1	2	3	4	5
$Actividad(\%):$	90	75	42	30	21

Obtener la recta de regresión por m.c. que permita estimar el porcentaje de actividad restante del preparado hormonal en función del tiempo y calcular el coeficiente de determinación lineal. Estimar el porcentaje de actividad al cabo de 6 meses, suponiendo que se mantenga dicha tendencia.

Solución: y=106,5-18,3x; r^2=0,952; x=6→y=-3,3 (al ser negativa, se considera que no hay actividad)

2º) Según la ley de Lambert-Beer, la absorbancia de un complejo se obtiene a partir de la concentración mediante técnicas espectrofotométricas, siendo los resultados de un estudio particular los que se muestran en la tabla siguiente:

$Concentración:$	1	2	3	5	10
$Absorbancia:$	0,10	0,36	0,57	1,09	2,05

Obtener la recta de regresión por m.c. que permita estimar la absorbancia a partir de la concentración y calcular el coeficiente de determinación lineal. Estimar la absorbancia correspondiente a una concentración igual a 7.

Solución: y=0,216x-0,073; r^2=0,996; x=7→y=1,44

3º) Las calificaciones obtenidas por diez alumnos en Matemáticas y Biología han sido las siguientes:

$Matemáticas\ (x):$	6	4	8	5	3,5	7	5	10	5	4
$Biología\ (y):$	6,5	4,5	7	5	4	8	7	10	6	5

Obtener las dos rectas de regresión por m.c. asociadas a dichas calificaciones y calcular el coeficiente de determinación lineal. Estimar la calificación en Biología de un alumno que ha obtenido un 5 en Matemáticas, y la esperada en Matemáticas para un alumno que ha obtenido 7,3 en Biología.

Solución: y=0,817x+1,6; x=1,039y-0,795; r^2=0,85;
x=5→y=5,7; y=7,3→x=6,8

4^o) Obtener las dos rectas de regresión por m.c. asociadas a los puntos: $(-3;-5), (2,5;6), (-0,5;0), (0;1)$ y $(4;9)$, y calcular el coeficiente de determinación lineal.

Solución: Al ser r^2=1 ambas rectas coinciden: y=2x+1

5^o) Ajustar por m.c. un recta que pase por el origen a los datos experimentales siguientes:

$x:$	1	2	3	4	5
$y:$	1	1	0	1	3

Solución: y=0,4x

6^o) Ajustar un modelo exponencial del tipo $y = b_0.b_1^x$ y otro multiplicativo $y = b_0 x^{b_1}$ a los datos experimentales dados en la tabla siguiente y estudiar la precisión de dicho ajuste:

$x:$	2,2	2,7	3,5	4,1
$y:$	67	60	53	50

Solución: $y=92{,}27(0{,}858)^x \rightarrow r^2=0{,}978$; $y=96.27x^{-0{,}47} \rightarrow r^2=0{,}996$

7^o) Se ha observado el crecimiento de una población de bacterias en una placa de Petri obteniéndose los datos siguientes por unidad de superficie:

Días desde el cultivo $(x):$	1	2	3	4	5	6
Millones de bacterias $(y):$	1,6	4,5	13,8	40,2	125	300

Ajustar una curva exponencial del tipo $y = e^{b_0+b_1x}$ que representa dicho crecimiento y estudiar la precisión de dicho ajuste.

Solución: $y=e^{1{,}063x-0{,}583}$; r^2=0,99

8^o) Los valores obtenidos para la variable y a partir de valores de x son los siguientes:

$x:$	−2	−1	0	1	2
$y:$	6,8	4	3,1	4,2	7,1

Ajustar a dichos datos: a) un modelo lineal: $y = b_0 + b_1x$, b) un modelo cuadrático: $y = b_0 + b_1x^2$, c) estudiar la precisión de ambos modelos

Solución: a) y=5,04+0,08x; *b*) $y=3{,}126+0{,}957x^2$;
c) lineal: r^2=0,005 (s_R^2=2,563), cuadrático: r^2=0,995 (s_R^2=0,013)

9º) Partiendo de la relación adiabática: $PV^a = b$, donde a y b son constantes, estimar a partir de la siguiente tabla de datos experimentales los valores de a y b, así como la presión que corresponde a un volumen de 2 litros.

$P(kg/cm^2)$:	0,5	1	1,5	2	2,5	3
$V(litros)$:	1,65	1,03	0,74	0,61	0,53	0,45

Solución: a=1,384 b=1,011. Para V=2 es P=0,387

10º) Para los datos de la tabla siguiente obtener los modelos de regresión: a) exponencial, b) inverso-y, c) inverso-x, d) doble-inverso, e) logaritmo-x, f) multiplicativo, g) raíz-cuadrada-x, h) raíz cuadrada-y:

x :	72	70	68	86	89	85	70	68	70	70	68	80
y :	76	80	92	67	69	70	75	86	87	102	98	67

Solución: a) y=$e^{5,4861-0,0148x}$; b) y=$\frac{1}{0,0002x-0,0014}$;
c) y=-15,8215+$\frac{7141,188}{x}$; d) y=$\frac{1}{0,0281-1,1417/x}$;
e) y=476,66-91,9lnx; f) y=11658,9x-1,1566x;
g) y=260,289-20,8037$\sqrt{x}$; h) y=$(13,8812-0,0659x)^2$;

11º) Completar los datos faltantes en la siguiente tabla sabiendo que $\overline{x} = 6$, $\overline{y} = 8$, $s_{xy} = 13$, y obtener la recta de regresión m.c. de y sobre x:

x :	2	4	a	10
y :	b	c	d	e
xy :	8	f	80	g

Solución: a =8, b =4, c =4, d =10, e =14, f =16, g =140
Recta de regresión m.c.: y=1,3x+0,2.

12º) Las rectas de regresión asociadas a una variable bidimensional (x, y) tienen por ecuaciones: $x + y + 1 = 0$ y $2x + y = 0$. Calcular su centro de gravedad y su coeficiente de correlación lineal.

Solución: G≡ ($\overline{x}$=1; $\overline{y}$=-2) $r = -\frac{1}{\sqrt{2}}$

13º) Sabiendo que la longitud de una varilla depende de la temperatura ambiente mediante una relación del tipo: $l = l_0(1 + kt)$, siendo l_0 la longitud

inicial y k la constante de dilatación, hallar los valores de l_0 y k a partir de los datos siguientes:

$t(^oC)$:	20	40	50	60
$l(mm)$:	1000, 22	1000, 65	1000, 90	1001, 05

Solución: l_0=999,804mm k=212·10^{-7}

14^o) Obtener un modelo exponencial del tipo $y = ab^x$ para relacionar el incremento de biomasa (y) en un cultivo celular en función del tiempo transcurrido (x), y estudiar la precisión del mismo a partir de la tabla siguiente:

x :	0	1	2	3	4
y :	10	32	89	271	808

Solución: y=10,244·$2{,}98^x$, r^2=0,999

Tema 6. MÉTODOS NO PARAMÉTRICOS

6.1. FUNDAMENTOS

El estudio desarrollado en los temas anteriores partía de la base de que las variables objeto de estudio se distribuían según un modelo de Gauss, de forma que toda la información venía recogida en la media y en la desviación típica, por lo que todo el proceso de inferencia quedaba reducido a la formulación de hipótesis sobre estos dos parámetros; de ahí su nombre de inferncia paramétrica. Sin embargo, cuando alguna de las variables implicadas en el estudio no se distribuye normalmente, o bien cuando se trata de una variable ordinal, el estudio previo no es válido, siendo necesario recurrir a métodos de inferencia no paramétrica, que se fundamentan en la función de distribución de la variable o en parámetros más robustos que la media, como es la mediana.

Esta situación es similar a la que ocurre cuando la policía científica pretende identificar a un delincuente a partir de una huella digital. El proceso consiste en compararla con un amplio conjunto de huellas disponible en una base de datos mediante superposición, e ir analizando las coincidencias. Si en lugar de una simple huella digital se obtuviese un resto orgánico, como pelo o sangre del malechor, podría extraerse su ADN, que es como su parámetro característico (equivalente a la media y desviación típica), y todo el proceso de identificación quedaría reducido a buscar analogías entre el ADN encontrado y el que debería estar registrado en una base de datos.

Por tanto, si X_1 y X_2 son dos variables aleatorias con funciones de distribución $F_1(x)$ y $F_2(x)$ respectivamente, un conraste no paramétrico formulará hipótesis del tipo:

$$H_0 : F_1(x) = F_2(x)$$
$$H_1 : F_1(x) \neq F_2(x)$$

El procedimiento básico para resolver este contraste es el test de *Kolmogorov-Smirnov*, que se desarrollará en la sección 6.2. No obstante, como se ha indicado anteriormente, existen diversos contrastes basados en medidas de orden de las variables, concretamente en la mediana, de la forma:

$$H_0 : Mediana(X_1) = Mediana(X_2)$$
$$H_1 : Mediana(X_1) \neq Mediana(X_2)$$

los cuales tienen especial predicamento en Ciencias de la Salud, pero hay que interpretarlos correctamente y en su justa medida, ya que si un test de este tipo lleva a la decisión de rechazar la hipótesis nula, al ser diferentes las medianas, las dos variables no pueden ser idénticas, lo que indica una diferencia significativa. Pero en caso de aceptar H_0, no se puede concluir que ambas sean iguales, ya que basta observar, por ejemplo, las variables

$$\begin{array}{lccc} X_1: & 1 & 2 & 3 \\ X_1: & 1 & 2 & 3000 \end{array}$$

La media de X_1 es 2 y la de X_2 es 1001, sin embargo ambas tienen igual mediana 2.

Los contrastes no parámetricos puede aplicarse siempre, incluso aunque la variable o variables sigan una distribución normal, pero tienen menos potencia que los contrastes paramétricos basados en la media y varianza.

6.2. TEST DE KOLMOGOROV-SMIRNOV

Si $\hat{F}_1(x)$ y $\hat{F}_2(x)$ denotan las funciones de distribución empíricas de las variables X_1 y X_2 respectivamente, es decir, las estimadas a partir de los datos muestrales de tamaños n_1 y n_2, se define el estadístico:

$$D_{n_1 n_2} = \max_{-\infty<x<\infty} \mid \hat{F}_1(x) - \hat{F}_2(x) \mid$$

Entonces la función de distribución de Kolmogorov-Smirnov (K-S) viene dada por:

$$Q(x) = \lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \Pr ob\left(\sqrt{\tfrac{n_1 n_2}{n_1+n_2}} D_{n_1 n_2} \leq x\right)$$

Por tanto, cuando n_1 y n_2 son grandes, el criterio de decisión consiste en aceptar H_0 si $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1 n_2} \leq x_\alpha$ siendo $Q(x_\alpha) = 1 - \alpha$.

Mediante el test *K-S* puede contrastarse si dos variables aleatorias tienen la misma distribución, pero también si una variable se ajusta a un modelo de probabilidad pre-establecido.

EJEMPLO

Se quiere contrastar si el consumo de un suplemento de calcio retrasa la osteoporosis en mujeres post-menopáusicas, para lo cual se registran las edades de las pacientes de una consulta de nutrición, diferenciando entre las que consumen el suplemento de calcio y las que no lo hacen. Los resultados fueron los siguientes:

	Mujeres sin suplemento de calcio			Mujeres con suplemento de calcio		
Edad	Frec.	Frec. relat.	$\hat{F}_1(x)$	Frec.	Frec. relat.	$\hat{F}_2(x)$
51-55	4	0,0190	0,0190	2	0,0156	0,0156
56-60	17	0,0805	0,0995	10	0,0781	0,0938
61-65	45	0,2133	0,3128	28	0,2188	0,3125
66-70	67	0,3175	0,6306	45	0,3516	0,6641
71-75	53	0,2512	0,8815	30	0,2343	0,8984
76-80	15	0,0711	0,9526	7	0,0547	0,9531
81-85	10	0,0474	1	6	0,0469	1
TOTAL	211	1		128		

Como max $\mid \hat{F}_1(x) - \hat{F}_2(x) \mid = 0,0338$, resulta que $\sqrt{\frac{211 \cdot 128}{211+128}} \cdot 0,0338 = 0,302$. Para un nivel $\alpha = 0,05$, el valor crítico de *K-S* es 1,36 por lo que se acepta la hipótesis de que ambas variables se distribuyen de igual modo y, por tanto, el suplemento de calcio no produce ningún efecto significativo.

Cuando se quiere contrastar si una variable se ajusta a un modelo de Gauss puede suceder que se nos especifique cuáles son sus parámetros; pero, en ocasiones es necesario estimarlos a partir de los datos muesrales. En tal caso, se aplica el test *K-S* con la denominada corrección de *Lilliefors*.

EJEMPLO

Vamos a contrastar si la distribución de edades de las mujeres que no toman suplemento de calcio se ajusta a un modelo normal. Para ello tenemos, como paso previo, que estimar la media y desviación típica de la variable, resultando: $\hat{\mu} = \bar{x}_{211} = 68,52, \hat{\sigma} = s_{210} = 6,45$. Así tenemos:

$$\begin{aligned} Prob(51 \leq X < 56) &= 0,0191 \\ Prob(56 \leq X < 61) &= 0,0858 \\ Prob(61 \leq X < 66) &= 0,2117 \\ Prob(66 \leq X < 71) &= 0,3025 \\ Prob(71 \leq X < 76) &= 0,2382 \\ Prob(76 \leq X < 81) &= 0,1087 \\ Prob(81 \leq X < 86) &= 0,0340 \end{aligned}$$

y, en este caso, denotando $F(x)$ a la función de distribución de la variable de Gauss, es max $\mid \hat{F}(x)$ -$F(x) \mid = 0,0242$, por lo que $\sqrt{\frac{211 \cdot 211}{211+211}} \cdot 0,0242 = 0,249$, que es superior al valor crítico de *K-S* con corrección de *Lilliefors* 0,061 aceptándose la normalidad de la variable

6.3. AJUSTE A UN MODELO TEÓRICO

Sean $(x_1, x_2, ..., x_n)$ y $(y_1, y_2, ..., y_n)$ dos puntos del espacio n-dimensional. La expresión más conocida para calcular la distancia entre ellos es la euclídea, que viene dada por:

$$d_E = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Pero existen otros métodos para obtener una distancia, muchos de ellos basados en ponderar la distancia euclídea de formas alternativas:

$$d_w = \sqrt{\varpi_1(x_1 - y_1)^2 + \varpi_2(x_2 - y_2)^2 + \cdots + \varpi_n(x_n - y_n)^2}$$

Así, la distancia de *Mahalanobis* pondera la diferencia entre coordenadas según la dispersión de las variables implicadas, mientras que en la distancia *chi-cuadrado* los factores de ponderación son: $\varpi_i = 1/y_i$. Es precisamente esta distancia la base de este tema, la cual, además, bajo ciertas hipótesis seguirá una distribución χ^2 de Perason.

Una de las principales aplicaciones de la distancia *chi-cuadrado* es al contraste de la hipótesis de si un fenómeno observado se ajusta a una cierta ley o modelo teórico:

$$H_0 : X \rightsquigarrow Modelo$$
$$H_1 : X \not\rightsquigarrow Modelo$$

Para ello supongamos que los resultados del experimento o fenómeno en cuestión se pueden agrupar en una serie de categorías $C_1, C_2, ..., C_k$, y que se presentan con frecuencias $n_1, n_2, ..., n_k$ respectivamente. Denotemos, así mismo, $e_1, e_2, ..., e_k$ a las frecuencias teóricas si el fenómeno se ajustara al modelo. Nos encontraríamos entonces ante un esquema del tipo:

Categoría	C_1	C_2	$\cdots$	C_k
Frecuencia observada	n_1	n_2	$\cdots$	n_k
Frecuencia según modelo	e_1	e_2	$\cdots$	e_k

El estadístico del contraste viene dado por la distancia *chi-cuadrado* entre las frecuencias observadas y teóricas:

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

siendo $n = \sum_{i=1}^{k} n_i$ el número total de observaciones. Así, si H_0 es cierta, y el tamaño muestral n es suficientemente grande, el estadístico $\chi^2_{\exp}$ se distribuye según una ley *chi-cuadrado* con $k - m - 1$ grados de libertad, donde m representa el número de parámetros de la distribución del modelo; por ejemplo, si se trata de un modelo que no depende de ningún parámetro entonces $m = 0$; si se trata de uno de tipo *Poisson,* $m = 1$; si se trata de uno Gaussiano, $m = 2$. Cuando $\sum_{i=1}^{k} e_i = n$, entonces el estadístico puede calcularse de forma abreviada de la forma:

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n$$

Como norma deben ser las frecuencias $e_i \geq 5$; en caso contrario deben agruparse dos o más categorías en una sola.

EJEMPLO

Una ley de crecimiento de colonias de bacterias establece que la multiplicación debe ajustarse a un modelo exponenecial en progresión geométrica $N(t) = 2^t$, siendo t el tiempo (en días) transcurrido desde el cultivo de la primera colonia. Tras realizar el experimento se ha observado el siguiente proceso de crecimiento:

Días transcurridos	0	1	2	3	4	5	6
Frecuencia observada n_i	1	3	6	14	20	33	60
Frecuencia teórica $e_i = 2^i$	1	2	4	8	16	32	64

Para contrastar si puede aceptarse que el fenómeno se ajusta al modelo exponencial evaluaremos el estadístico:

$$\chi^2_{\text{exp}} = \frac{(1-1)^2}{1} + \frac{(3-2)^2}{2} + \frac{(6-4)^2}{4} + \frac{(14-8)^2}{8} + \frac{(20-16)^2}{16} + \frac{(33-32)^2}{32} + \frac{(60-64)^2}{64} = 7{,}281$$

Como hay 7 clases y el modelo no depende de ningún parámetro, entonces $m = 0$ y $k - 1 = 6$, y el valor crítico de la distribución *chi-cuadrado* con 6 grados de libertad para el nivel $\alpha = 0,05$ es $12,592$, de manera que el valor del estadístico queda dentro de la región de aceptación y se concluye que el crecimiento de la población se ajusta al modelo exponencial.

EJEMPLO

Según las leyes de Mendel, el cruce de guisantes de color verde y amarillo, y piel lisa y rugosa se ajusta a una proporcionalidad del tipo: $AL(9) : VL(3) : AR(3) : VR(1)$. Al realizar un experimento se obtuvieron 556 cruces con las siguientes frecuencias respectivamente: $315, 108, 101, 32$. Para contrastar si el experimento se ha ajustado a la leyes de Mendel tengamos en cuenta que, la probabilidad de aparación de cada categoría es $\frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}$. Por tanto, podemos escribir:

Clase	AL	VL	AR	VR
n_i	315	108	101	32
$e_i = \frac{p_i}{556}$	$312,75$	$104,25$	$104,25$	$34,75$

y como $\sum_{i=1}^{4} e_i = 556$, el estadístico del contraste se calcula de la forma:

$$\chi^2_{\text{exp}} = \frac{315^2}{312{,}75} + \frac{108^2}{104{,}25} + \frac{101^2}{104{,}25} + \frac{32^2}{34{,}75} - 556 = 0{,}47$$

En este caso hay 4 clases y el modelo no depende de ningún parámetro, entonces $m = 0$ y $k - 1 = 3$, y el valor crítico de la distribución *chi-cuadrado* con 3 grados de libertad para el nivel $\alpha = 0{,}05$ es $7{,}815$, de manera que el valor del estadístico queda dentro de la región de aceptación y se concluye que el cruce de guisantes se ajusta a las leyes de Mendel.

EJEMPLO

Para contrastar si la variable que representa el número de unidades de un cierto tipo de antibiótico dispensadas diariamente en una oficina de farmacia se ha realizado un registro diario a lo largo de 45 días obteniendo los siguientes resultados:

Número de días	0	1	2	3	4	5
Unidades dispensadas (n_i)	2	5	14	15	6	3
Frecuencias teóricas (e_i)	$3{,}344$	$8{,}699$	$11{,}295$	$9{,}752$	$6{,}363$	$3{,}308$

donde las e_i se han calculado multiplicando $e_i = 45p_i$ siendo las p_i las probabilidades calculadas con el modelo de *Poisson* de parámetro:

$$\hat{\lambda} = \bar{x}_{45} = \frac{1}{45}\left(0 \cdot 2 + 1 \cdot 5 + 2 \cdot 14 + 3 \cdot 15 + 4 \cdot 6 + 5 \cdot 3\right) = 2{,}6$$

La evaluación del estadístico de contraste se realiza de la forma:

$$\chi^2_{\text{exp}} = \frac{(2-3{,}344)^2}{3{,}344} + \frac{(5-8{,}699)^2}{8{,}699} + \frac{(14-11{,}295)^2}{11{,}295} + \frac{(15-9{,}752)^2}{9{,}752} + \frac{(6-6{,}363)^2}{6{,}363} + \frac{(3-3{,}308)^2}{3{,}308} = 5{,}628$$

En este caso hay 6 clases y, dado que el modelo de *Poisson* depende de un parámetro, entonces $m = 1$ y $k - m - 1 = 4$. Como el valor crítico de la distribución *chi-cuadrado* con 4 grados de libertad para el nivel $\alpha = 0{,}05$ es $9{,}492$, el valor del estadístico queda dentro de la región de aceptación y se concluye que, con un nivel de significación $\alpha = 0{,}05$, la dispensación diaria de ese tipo de antibiótico se ajusta a un modelo de *Poisson* con parámetro $2{,}6$.

EJEMPLO

Se quiere contrastar si el nivel de glucosa basal en no diabéticos se distribuye según un modelo de Gauss, para lo cual se realizó un análisis de sangre en 100 personas y los resultados, expresados en mg/dL, se agruparon en 6 intervalos de clase:

Intervalos	60-69,9	70-79,9	80-89,9	90-99,9	100-109,9	110-119,9
Frecuencias (n_i)	2	18	30	25	20	5
$e_i = 100p_i$	3,67	14,30	28,80	30,73	16,58	4,73

A partir de los datos observados se obtienen los estimadores:

$$\hat{\mu} = \bar{x}_{100} = 90,8 \qquad \hat{\sigma} = s_{99} = 12$$

y así, utilizando las tablas de la normal tipificada $Z \rightsquigarrow N(0;1)$, resulta:

p_1=Prob(60≤X<70)=Prob($\frac{60-90,8}{12}$ ≤Z<$\frac{70-90,8}{12}$)=Prob(-2,57≤Z<-1,73)=0,0367

p_2=Prob(70≤X<80)=Prob($\frac{70-90,8}{12}$ ≤Z<$\frac{80-90,8}{12}$)=Prob(-1,73≤Z<-0,90)=0,1430

p_3=Prob(80≤X<90)=Prob($\frac{80-90,8}{12}$ ≤Z<$\frac{90-90,8}{12}$)=Prob(-0,90≤Z<-0,07)=0,2880

p_4=Prob(90≤X<100)=Prob($\frac{90-90,8}{12}$ ≤Z<$\frac{100-90,8}{12}$)=Prob(-0,07≤Z<0,77)=0,3073

p_5=Prob(100≤X<110)=Prob($\frac{100-90,8}{12}$ ≤Z<$\frac{110-90,8}{12}$)=Prob(0,77≤Z<1,60)=0,1658

p_6=Prob(110≤X<120)=Prob($\frac{110-90,8}{12}$ ≤Z<$\frac{120-90,8}{12}$)=Prob(1,60≤Z<2,43)=0,0473

La evaluación del estadístico de contraste se realiza de la forma:

$$\chi^2_{\exp} = \frac{(2-3,67)^2}{3,67} + \frac{(18-14,3)^2}{14,3} + \frac{(30-28,8)^2}{28,8} + \frac{(25-30,73)^2}{30,73} + \frac{(20-16,58)^2}{16,58} + \frac{(5-4,73)^2}{4,73} = 3,556$$

En este caso hay 6 clases y, dado que el modelo de *Gauss* depende de 2 parámetros, entonces $m = 2$ y $k - m - 1 = 3$. Como el valor crítico de la distribución *chi-cuadrado* con 3 grados de libertad para el nivel $\alpha = 0,05$ es $7,815$, el valor del estadístico queda dentro de la región de aceptación y se concluye que, con un nivel de significación $\alpha = 0,05$, la concentración basal de glucosa en los no diabéticos se ajusta a un modelo normal.

6.4. COMPARACIÓN DE VARIABLES PAREADAS: TEST DE LOS SIGNOS Y TEST DE WILCOXON

Una alternativa no paramétrica a los contrastes de comparación de variables pareadas es el *test de los signos*, que tiene en cuenta únicamente el número de cambios de signo o de tendencia de cada muestra de la variable pareada. Por ejemplo, si se aplica un tratamiento y quiere analizar si ha habido un cambio de estado tras un periodo de tiempo, se marcan con signo positivo (+) los casos en los que sí se ha producido un cambio y con signo negativo (-) los que no ha habido cambio. Así, en caso de no haber diferencias, la probabilidad de un cambio sería $\frac{1}{2}$, ya que se atribuiría al azar, por lo que en una muestra de tamaño n, el número de cambios, es decir, el número de veces en los que aparece el signo +, se distribuye según un modelo binomial con parámetro $p = \frac{1}{2}$. Entonces, si k es el número de cambios en la muestra, se calcula $\text{Prob}(X \geq k)$ según dicho modelo binomial y si es mayor que el nivel de significación ($\alpha = 0,05$) se concluye que no hay diferencias significativas.

EJEMPLO

Al ensayar un nuevo tratamiento en 20 pacientes afectados de una patología se observó que 15 de ellos tenían una evolución más favorable que con otro tratamiento estándar. Para demostrar estadísticamente que el nuevo tratamiento es más eficaz con un nivel de significación $\alpha = 0,05$ consideremos una variable binomial con $n = 20$ y $p = \frac{1}{2}$. Así:

$$\text{Prob}(X \geq 15) = \sum_{k=15}^{20} \binom{20}{k} \left(\tfrac{1}{2}\right)^k \left(\tfrac{1}{2}\right)^{20-k} = \tfrac{1}{2^{20}} \sum_{k=15}^{20} \binom{20}{k} = 0,015$$

que es menor que el nivel de significación, por lo que se rechaza la hipótesis nula de no haber diferencias y se concluye que el nuevo tratamiento sí es más eficaz que el estándar.

Como se puede comprobar, el test de los signos es muy simple y sólo tiene en cuenta los cambios de estado (signo) entre una situación inicial y otra final. Una versión más precisa de este contraste consiste cuantificar el valor de la diferencia entre estados o, al menos, poder ordenar dichas diferencias, y no sólo el signo. Así, el test de los signos con rango para datos

pareados, o test de *Wilcoxon*, calcula las diferencias y las ordena de menor a mayor asignando al orden el signo + o - según la diferencia sea positiva o negativa. En los casos de empates se les asigna a las correspondientes observaciones el promedio de los rangos; y caso de difrencias nulas, estos caso no se contabilizan. Entonces se suman los rangos positivos por un lado y los negativos por otro y se considera como estadístico de contraste la menor suma en valor absoluto, comparándose con el valor crítico correspondiente de forma que si el valor del estadístico es superior se acepta la hipótesis nula, que en este caso consiste en aceptar la igualdad entre las medianas de ambas variables.

EJEMPLO

Se realiza una prueba a un grupo de 10 alumnos y, al cabo de un tiempo, tras haber recibido unas clases complementarias, se les vuelve a realizar otra prueba sobre la misma materia. fueron los siguientes:

Alumno	1	2	3	4	5	6	7	8	9	10
Nota inicial	4	6	6,5	6	5	9,5	10	8	7	5,5
Nota final	3	5,5	6	3	8,5	8	9	8,5	7	3,5

Para contrastar si las clases recibidas tuvieron una repercusión positiva en el aprendizaje se calculan las difrencias entre las puntuaciones obtenidas por los alumnos antes y después de recibir las clases complementarias, obteniéndose lo siguiente:

Alumno	1	2	3	4	5	6	7	8	9	10
Diferencia	1	0,5	0,5	3	−3,5	1,5	1	−0,5	0	2
Rango	4,5	2	2	8	9	6	4,5	2		7
Rango con signo	4,5	2	2	8	−9	6	4,5	−2		7

ya que cuando varios rangos coinciden se les asigna la media de todos ellos. A la diferencia nula no se le asigna rango. Así, la suma de los rangos positivos es $\sum R_i^+ = 34$ y la de los negativos $\sum R_i^- = -11$. El estadístico de contraste es entonces $W = \min\{34, 11\} = 11$,y como el valor crítico asociado a $n = 9$ y $\alpha = 0,05$ es 5, que es inferior al valor del estadístico de contraste, se acepta la hipótesis nula y se concluye que las clases complementarias no han resultado eficaces de cara a variar la calificación de los alumnos.

6.5. COMPARACIONES DE VARIABLES INDEPENDIENTES: TEST DE MANN-WHITNEY

Además del test *K-S* descrito en la sección 6.2, existe una extensión del test de *Wilcoxon* para variables independientes que contrasta si las variables tienen la misma mediana. Se denomina test de *Mann-Whitney* y para aplicarlo partimos de muestras de tamaños n_1 y n_2 respectivamente de ambas variables y las reunimos en una sola muestra de tamaño n_1+n_2, pero sabiendo a qué variable corresponde cada observación. Entonces se ordenan los valores de la muestra conjunta de menor a mayor y se les asigna un rango, denotando $R_i^{(1)}$ los correspondientes a la variable primera y $R_i^{(2)}$ los correspondientes a la variable segunda. El estadístico de *Mann-Whitney* se construye de la foma:

$$U = \sum_{i=1}^{n_2} R_i^{(2)} - \frac{n_2(n_2+1)}{2}$$

y se compara con el valor crítico, de forma que si éste es menor que el del estadístico U se acepta la hipótesis nula de igualdad de medianas entre ambas variables.

Cuando los tamaño muestrales son grandes, la distribución de U se puede aproximar por una normal de parámetros:

$$\mu = \frac{n_1 n_2}{2} \qquad \sigma = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

EJEMPLO

Para comparar si dos sistemas de enseñanza de inglés son igual de eficaces se les asigna la misma prueba de nivel a alumnos que han seguido uno de los dos sistemas de aprendizaje, 9 el sistema 1 (S1) y 10 el sistema 2 (S2), y se anota el orden en que van finalizando la prueba, resultando lo siguiente:

S2 S2 S2 S1 S1 S2 S1 S2 S2 S2 S2 S1 S2 S1 S1 S2 S1 S1 S1

La suma de los rangos de S2 es:

$$\sum_{i=1}^{10} R_i^{(2)} = 1+2+3+6+8+9+10+11+13+16 = 79$$

por lo que

$$U = 79 - \frac{10 \cdot 11}{2} = 24$$

y como el valor crítico correspondiente al nivel $\alpha = 0,05$ es 20, inferior a 24, se acepta la igualdad de eficacia de ambos sistemas de aprendizaje de inglés.

EJEMPLO

Se realiza un ensayo clínico para comprobar la respuesta del organismo ante dos ansiolíticos (A y B), la cual se evalúa según una escala de Likert, obteniendo los siguientes resultados

A	3	4	6	5	9	10	2	5	8	6			
B	7	6	10	7	4	10	10	8	9	7	9	10	5

Para contrastar si existen diferencias significativas entre ambos al nivel $\alpha = 0,05$ se ordenan conjuntamente de menor a mayor, anotando el orden entre paréntesis:

$2^A(1)\, 3^A(2)\, 4^A(3{,}5)\, 4^B(3{,}5)\, 5^A(6)\, 5^A(6)\, 5^B(6)\, 6^A(9)\, 6^A(9)\, 6^B(9)\, 7^B(12)\, 7^B(12)\, 7^B(12)$
$8^A(14{,}5)\, 8^A(14{,}5)\, 9^A(17)\, 9^B(17)\, 9^B(17)\, 10^A(21)\, 10^B(21)\, 10^B(21)\, 10^B(21)\, 10^B(21)$

La suma de los rangos correspondientes a las puntuaciones del ansiolítico B es:

$$\sum_{i=1}^{13} R_i^{(B)} = 3,5 + 6 + 9 + 12 + 12 + 12 + 17 + 17 + 21 + 21 + 21 + 21 = 172,5$$

por lo que el valor del estadístico de Mann-Whitney es:

$$U = 172,5 - \frac{13 \cdot 14}{2} = 81,5$$

Aunque los tamaño muestrales no sean muy grandes en este caso, aproximaremos la distribución del estadístico U mediante una normal de parámetros:

$$\mu = \frac{10 \cdot 13}{2} = 65 \qquad \sigma = \sqrt{\frac{10 \cdot 13 \cdot 24}{12}} = 16,12$$

Así, tipificando el valor de U resulta: $\frac{81{,}5-65}{16{,}12} = 1{,}026$ inferior al valor crítico $1{,}96$ por lo que se acepta la hipótesis nula de que ambas variables tienen igual mediana.

6.6. COMPARACIÓN DE VARIAS VARIABLES: TEST DE KRUSKAL-WALLIS Y TEST DE FRIEDMAN

Una alernativa no paramétrica al ANOVA I, cuando las k variables son independientes, pero no todas son Gaussianas, es el test de *Kruskal-Wallis*, que contrasta la hipótesis de igualdad de las medianas entre las variables. Para su ejecución, si denotamos $n_1, n_2, ..., n_k$ a los correspondientes tamaños muestrales y $n_1 + n_2 + ... + n_k = n$, se reunen todos los datos en una sola muestra de tamaño n, pero sabiendo a qué variable corresponde cada observación. Entonces se ordenan los valores de la muestra conjunta de menor a mayor y se les asigna un rango, denotando $R_i^{(j)}$ los correspondientes a la variable X_i y, a su vez $R_i = \sum_{j=1}^{n_i} R_i^{(j)}$. El estadístico de *Kruskal-Wallis* se define de la forma:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Si los tamaños muestrales n_i son grandes y existe homogeneidad entre las distribuciones de los k grupos, el estadístico H se distribuye según una ley *chi-cuadrado* con $k-1$ grados de libertad. Para tamaños muestrales pequeños existen tablas con los valores críticos.

EJEMPLO

Se quiere estudiar la reacción de pacientes hipertensos con diferente grupo sanguíneo al administrarles un nuevo fármaco antihipertensivo. La variable de respuesta fue la disminución de la tensión arterial máxima tras un mes de tratamiento, y los aumentos de tensión se registraron con signo regativo:

A	1	0	$1,5$	-2	0	$1,5$	$-0,5$			
B	2	$2,5$	$3,3$	$3,5$	3	$1,8$	$2,7$	0	$3,5$	2
O	$1,5$	1	$2,5$	3	$1,5$	$3,5$	$2,5$	$2,2$	$1,1$	$2,5$
AB	1	$0,5$	$-0,5$	$1,5$	2	2	0	1		

Suponiendo que las variables no son normales, para contrastar la hipótesis de homogeneidad de las variables es preciso combinarlas en una sola muestra y ordenarlas de menor a mayor:

-2^A(1) -0,5^A(2,5) -0,5^D(2,5) 0^A(5,5) 0^A(5,5) 0^B(5,5) 0^D(5,5) 0,5^A(8) 1^A(10,5) 1^C(10,5) 1^D(10,5) 1^D(10,5) 1,1^C(13) 1,5^A(16) 1,5^A(16) 1,5^C(16) 1,5^D(16) 1,5^D(16) 1,8^B(19) 2^B(21,5) 2^B(21,5) 2^D(21,5) 2^D(21,5) 2,2^C(24) 2,5^B(27) 2,5^B(27) 2,5^C(27) 2,5^C(27) 2,5^C(27) 2,7^B(30) 3^B(31,5) 3^C(31,5) 3,3^B(33) 3,5^B(34,5) 3,5^C(34,5)

La suma de los rangos para cada grupo sanguíneo es: $R_A = 57$ $R_B = 250,5$ $R_C = 226,5$ $R_D = 96$, así el valor del estadístico es:

$$H = \frac{12}{35\cdot 36}\left(\frac{57^2}{7} + \frac{250{,}5^2}{10} + \frac{226{,}5^2}{10} + \frac{96^2}{8}\right) - 3 \cdot 36 = 16,01$$

Como para un nivel $\alpha = 0,05$ el valor crítico de la distribución chi-cuadrado con 3 grados de libertad es: $\chi^2_3 = 7,815$, que es menor que el valor de H, se rechaza la hipótesis nula y se concluye que la respuesta del organismo es significativamente diferente según el grupo sanguíneo del paciente.

Cuando las variables están correladas, en el sentido de que para cada individuo muestral se realizan varias observaciones en niveles diferentes (distintos instantes o tratamientos), como, por ejemplo, cuando se mide una variable fisiológica en un grupo de pacientes en varios momentos temporales, el contraste no paramétrico adecuado para comparar la variable en todos los niveles es el test de *Friedman*, cuyo estadístico se construye de forma parecida al de Kruskal-Wallis. Así, consideremos que se mide una variable con k niveles en una muestra de n pacientes resultando:

$Nivel$ / $Muestra$	1	2	$\cdots$	k
1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

El estadístico de Friedman para contrastar si el comportamiento de la variable es el mismo en los k niveles de la variables viene dado por:

$$F = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

donde a las observaciones de cada fila se les asignan rangos de menor a mayor desde 1 hasta k; a continuación se suman los rangos correspondientes a cada columna, siendo R_j la suma correspondiente a la columna j-ésima. Cuando n es grande y H_0 es cierta, se distribuye según una ley chi-cuadrado con $k-1$ grados de libertad.

EJEMPLO

Tras administrar un antipirético a un grupo de 10 pacientes en observación con fiebre alta se ha medido la temperatura al cabo de 1 y 2 horas obteniendo los siguientes resultados:

Paciente	1	2	3	4	5	6	7	8	9	10
Temp. basal	38,8	39,0	38,6	39,1	38,5	38,9	39,6	40,0	38,7	39,0
Temp. 1 hora	38,5	39,0	37,9	38,5	38,6	38,1	39,9	38,8	38,2	38,2
Temp. 2 horas	37,5	38,2	37,0	37,2	38,0	37,6	39,7	37,5	38,3	37,5

Se pretende estudiar si es significativa al nivel $\alpha = 0,05$ la reducción de temperatura en el tiempo. Para ello, calculamos los rangos en cada paciente obteniendo

Paciente	1	2	3	4	5	6	7	8	9	10	R_j
Temp. basal	1	1,5	1	1	2	1	3	1	1	1	13,5
Temp. 1 hora	2	1,5	2	2	1	2	1	2	3	2	17,5
Temp. 2 horas	3	3	3	3	3	3	2	3	2	3	28

con lo cual, como $n = 10$ y $k = 3$:

$$F = \frac{12}{10 \cdot 3 \cdot 4} \left(13,5^2 + 17,5^2 + 28^2\right) - 3 \cdot 10 \cdot 4 = 7,25$$

y, dado que para $\alpha = 0,05$ es $\chi^2_2 = 5,99$, se rechaza la hipótesis nula y se concluye que la reducción de temperatura a lo largo de las 2 horas consideradas causada por el antipirético si es significativa.

6.7. CORRELACIÓN POR RANGOS DE SPEARMAN

Una alternativa al coeficiente de correlación de Pearson para variable ordinales es el coeficiente de correlación de *Spearman*, que parte, no de los valores de las variables, sino el orden orango que ocupan. Así, denotando

$$\begin{array}{llll} \text{Rango X} & R_x^{(1)} & R_x^{(2)} & \cdots & R_x^{(n)} \\ \text{Rango Y} & R_y^{(1)} & R_y^{(2)} & \cdots & R_y^{(n)} \end{array}$$

se define el coeficiente de correlación de *Spearman* de la forma:

$$r_s = 1 - 6\frac{\sum_{i=1}^{n}\left(R_x^{(i)} - R_y^{(i)}\right)^2}{n^3 - n}$$

cuyo valor oscila entre -1 y $+1$. El contraste de significación para r_s formulado de la forma $\begin{cases} H_0 : r_s = 0 \\ H_1 : r_s \neq 0 \end{cases}$ se resuelve evaluando el estadístico:

$$t_{\text{exp}} = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

que, cuando H_0 es cierta, se distribuye según una ley t de *Student* con $n-2$ grados de libertad.

EJEMPLO

El orden en que han quedado seleccionados 10 opositores en cada una de los dos ejercicios de que consta el examen ha sido la siguiente:

Opositor	A	B	C	D	E	F	G	H	I	J
Ejercicio 1	1	3	4	6	9	8	2	10	7	5
Ejercicio 2	1	2	5	10	7	8	3	9	6	4

con lo que:

$$r_s = 1\text{-}6\frac{(1\text{-}1)^2+(3\text{-}2)^2+(4\text{-}5)^2+(6\text{-}10)^2+(9\text{-}7)^2+(8\text{-}8)^2+(2\text{-}3)^2+(10\text{-}9)^2+(7\text{-}6)^2+(5\text{-}4)^2}{10^3\text{-}10} = 0{,}842$$

Para realizar el contraste de significación evaluaremos el estadístico

$$t_{\exp} = 0,842\sqrt{\frac{10-2}{1-0,842^2}} = 4,415$$

Como el intervalo de aceptación para el nivel $\alpha = 0,05$ es $[-2,306; 2,306]$ concluimos que se rechaza la hipótesis nula y el coeficiente es significativo.

EJEMPLO

La puntuación ordinal otorgada por dos evaluadores independientes a cinco trabajos que se presentan a un premio de investigación es la siguiente:

Trabajo	Evaluador 1	Evaluador 2
Farmacia creativa	3	4
Psicotropos de última generación	1	2
Políticas de género en Sanidad	4	5
Actividad farmacológica del gin-tonic	5	3
Antiinflamatorios en el antiguo Egipto	2	1

El coeficiente de *Spearman* es:

$$r_s = 1 - 6\frac{(3-4)^2+(1-2)^2+(4-5)^2+(5-3)^2+(2-1)^2}{5^3-5} = 0,6$$

y el estadístico para el contraste de significación:

$$t_{\exp} = 0,6\sqrt{\frac{5-2}{1-0,6^2}} = 1,299$$

Dado que, para el nivel $\alpha = 0,05$, el intervalo de aceptación es $[-3,182; 3,182]$ se concluye la aceptación de H_0 y r_s no es significativo.

Tema 7. TABLAS DE CONTINGENCIA

7.1. INDEPENDENCIA ENTRE VARIABLES CUALITATIVAS

Como se vió en el tema 5, el grado de dependencia entre dos variables cuantitativas se mide mediante el coeficiente de determinación R^2, que toma valores comprendidos entre 0 y 1. Así, mientras más próximo sea su valor a 1 indica un grado de dependencia más fuerte, mientras que valores próximos a 0 indican incorrelación entre las variables. De igual forma, el contraste de significación permite concluir si el coeficiente R^2, aunque su valor sea bajo, es o no significativo. En el caso de variables cualitativas, decidir si la dependencia entre ellas, que ahora se denomina *asociación*, es significativa se realiza de la forma que a continuación describiremos.

Denotemos por A y B a dos variables cualitativas (caracteres) que pueden tomar, respectivamentes, las modalidades $A_1, A_2, ..., A_h$ y $B_1, B_2, ..., B_k$, con frecuencias conjuntas n_{ij}, es decir el número de indiduos que presentan simultáneamente las modalidad A_i de la variable A y B_j de la variable B. Podemos representar esta situación en forma matricial mediante lo que se denomina *tabla de contingencia* $h \times k$:

$A \setminus B$	B_1	B_2	$\cdots$	B_K	Frecuencia marginal A
A_1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_h	n_{h1}	n_{h2}	$\cdots$	n_{hk}	$n_{h\bullet}$
Frecuencia marginal B	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet k}$	n

donde:

$$\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = \sum_{j=1}^{k} n_{\bullet j} = \sum_{i=1}^{h} n_{i\bullet} = n$$

El primer aspecto a tener en cuenta es que si A y B fuesen independientes, la distribución de cada uno de las dos variables condicionada a los valores de la otra sería igual en todos los casos, por lo que las filas de frecuencias serían proporcionales entre si y las columnas también lo son entre si. Eso se traduce en que cada frecuencia conjunta puede factorizarse en términos de sus frecuencias marginales de la forma:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

Por ejemplo, dada la siguiente tabla de contingencia 3×4:

$A \setminus B$	B_1	B_2	B_3	B_4	Fr. marg. A
A_1	8	2	10	6	26
A_2	20	5	25	15	65
A_3	12	3	15	9	39
Fr. marg. B	40	10	50	30	130

cualquier frecuencia conjunta admite una tal factorización:

$$5 = \frac{65 \cdot 10}{130} \quad 12 = \frac{39 \cdot 40}{130} \quad 3 = \frac{39 \cdot 10}{130} \quad 10 = \frac{26 \cdot 50}{130} \quad \cdots$$

En consecuencia, para resolver el contraste

H_0 : A y B son independientes

H_1 : Existe asociación entre A y B

se definebn las frecuencias teóricas conjuntas:

$$e_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

y se comparan con las observadas n_{ij} mediante la distancia *chi-cuadrado*:

$$\chi^2_{\text{exp}} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

que se distribuye, para tamaños muestrales grandes, según una ley *chi-cuadrado* con $(h-1)\times(k-1)$ grados de libertad si H_0 es cierta. Además, dado que $\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} e_{ij} = n$, el estadístico puede calcularse abreviadamente de la forma:

$$\chi^2_{\text{exp}} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{e_ij} - n$$

La resolución de este contraste permite decidir si se acepta o no la hipótesis de independencia, y en caso de rechazarse, algunas medidas del grado de asociación entre las variables son las siguientes:

- Coeficiente de *contingencia*: $C = \sqrt{\frac{\chi^2_{\text{exp}}}{\chi^2_{\text{exp}}+n}}$ que oscila entre $0 \leq C \leq \sqrt{\frac{P-1}{P}}$ siendo $P = \min\{h, k\}$
- Coeficiente de *Cramer*: $V = \sqrt{\frac{\chi^2_{\text{exp}}}{n\cdot(P-1)}}$ que oscila entre $0 \leq V \leq 1$

La interpretación de ambos coeficientes es obvia, pues mientras los coeficientes tomen valores más próximos a su máximo ello indicará que la asociación entre las variables cualitativas es más fuerte.

EJEMPLO

Se quiere comprobar si existe asociación entre el color del pelo y el de los ojos en las personas de una población, para lo cual, tras un muestreo se han obtenido los siguientes resultados:

Color pelo / Color ojos	Rubio	Castaño	Rojo	Negro	Total color ojos
Azul	22	16	15	12	65
Verde	15	20	9	22	66
Marrón	9	97	8	68	182
Total color pelo	46	133	32	102	313

El primer paso es obtener las frecuencias teóricas e_{ij} en caso de independencia, que representamos en la tabla siguiente:

Color pelo / Color ojos	Rubio	Castaño	Rojo	Negro	Total color ojos
Azul	$9,55$	$27,62$	$6,65$	$21,18$	65
Verde	$9,70$	$28,04$	$6,75$	$21,51$	66
Marrón	$26,75$	$77,34$	$18,60$	$59,31$	182
Total color pelo	46	133	32	102	313

El valor del estadístico de contraste se obtendrá entonces comparando las frecuencias observadas con las teóricas mediante la distancia *chi-cuadrado*:

$$\chi^2_{\text{exp}} = \frac{22^2}{9{,}55} + \frac{16^2}{27{,}62} + \cdots + \frac{8^2}{18{,}6} + \frac{68^2}{59{,}31} - 313 = 65{,}65$$

y como el valor crítico de la *chi-cuadrado* con $(3-1)\times(4-1) = 6$ grados de libertad correspondiente a un nivel $\alpha = 0{,}05$ es $12{,}592$, se rechaza la hipótesis nula de independencia y se concluye que existe asociación entre el color del pelo y el de los ojos. Por otra parte, dado que $P = \min\{3,4\} = 3$ los coeficientes de contingencia y de Cramer serían entonces:

$C = \sqrt{\frac{65{,}65}{65{,}65+313}} = 0{,}4164$, siendo su valor máximo posible $\sqrt{\frac{3-1}{3}} = 0{,}8165$

$V = \sqrt{\frac{65{,}65}{313(3-1)}} = 0{,}3238$, siendo su valor máximo posible 1

lo que indica que, aunque significativa, la asociación entre ambos caracteres es intermedia.

Cuando en una tabla de contingencia alguna frecuencia teórica e_{ij} es menor o igual a 5, la expresión del estadístico debe modificarse introduciendo la denominada *corrección de Yates*, o también corrección por continuidad, de la forma:

$$\chi^2_{\text{exp}} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(|n_{ij} - e_i j| - 0{,}5)^2}{e_i j}$$

7.2. CONCORDANCIA DIAGNÓSTICA

Una cuestión diferente al problema de la independencia frente a asociación entre variables cualitativas es la conocer si existe concordancia entre ambas. Tal situación es la que ocurre, por ejemplo, cuando dos evaluadores opinan sobre las distintas modalidades de una misma variable; o bien, cuando se quiere analizar si dos pruebas realizadas sobre los mismos alumnos tienen el mismo grado de dificultad, midiendo la calificación de cada una de ellas en una escala nominal del tipo: sobresaliente, notable, aprobado y suspenso.

Así, supongamos que los resultados de una variable se clasifican en una serie de categorías: $C_1, C_2, ..., C_h$, y que dos evaluadores independientes asignan los resultados de una muestra de n observaciones de la variable a una de esas categorías. Denotando, como anteriormente, n_{ij} a las frecuencias conjuntas, obtenemos una tabla de contingencia $h \times h$ de la forma:

Evaluador 2 / Evaluador 1	C_1	C_2	$\cdots$	C_h	Totales Evaluador 1
C_1	n_{11}	n_{12}	$\cdots$	n_{1h}	$n_{1\bullet}$
C_2	n_{21}	n_{22}	$\cdots$	n_{2h}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_h	n_{h1}	n_{h2}	$\cdots$	n_{hh}	$n_{h\bullet}$
Totales Evaluador 2	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet h}$	n

Para el estudio de la concordancia nos centraremos en las frecuencias de la diagonal principal n_{ii} que representan el número de acuerdos entre ambos evaluadores para cada una de las categorías de la variable, y a construiremos las frecuencias teóricas correspondientes a dichas concordancias:

$$e_{ii} = \frac{n_{i\bullet} n_{\bullet i}}{n} \quad i = 1.2, ..., h$$

La probabilidad de concordancias observadas es entonces:

$$p_o = \frac{1}{n} \sum_{i=1}^{h} n_{ii}$$

y la probabilidad de concordancias teóricas es:

$$p_e = \frac{1}{n} \sum_{i=1}^{h} e_{ii}$$

La medida de la concordancia entre ambos evaluadores viene dada por el coeficiente *kappa* de Cohen:

$$\kappa = \frac{p_o - p_e}{1. - p_e}$$

Si $\kappa = 1$ la concordancia es total, mientras que si $\kappa = 0$ no hay acuerdo. Podría incluso ocurrir que κ fuese negativo, si $p_o < p_e$. Se demuestra que la varianza de este estadístico es:

$$s_\kappa^2 = \frac{1}{n(1-p_e)^2}\left[p_e + p_e^2 - \frac{1}{n^2}\sum_{i=1}^{h} e_{ii}(n_{i\bullet} + n_{\bullet i})\right]$$

De tal forma, para la resolución del contraste

$$H_0 : \text{No existe concordancia entre los evaluadores}$$
$$H_1 : \text{Sí existe concordancia entre los evaluadores}$$

se construye el estadístico $\frac{\kappa}{s_\kappa}$ que, bajo hipótesis nula se distribuye según una ley normal tipificada $N(0,1)$.

EJEMPLO

Dos enólogos califican 360 muestras de vinos de una comarca con D.O. en una de estas 3 categorías: excelente (E), bueno (B) o regular (R). Los resultados de la evaluación se recogen en la tabla siguiente:

Enólogo 2 / Enólogo 1	E	B	R	Total Enól. 1
E	80	20	20	120
B	30	60	30	120
R	20	40	60	120
Total Enól. 1	130	120	110	360

Las frecuencias teóricas correspondientes a la diagonal principal son:

$$e_{11} = \frac{130\cdot 120}{360} = 43,33 \qquad e_{22} = \frac{120\cdot 120}{360} = 40 \qquad e_{33} = \frac{110\cdot 120}{360} = 36,67$$

Así:

$$p_o = \frac{80+60+60}{360} = 0,555 \qquad p_e = \frac{43,33+40+36,67}{360} = 0,333$$

y

$$\kappa = \frac{0,555-0,333}{1-0,333} = 0,334$$

De ello se deduce que el grado de concordancia entre ambos enólogos es medio-bajo. No obstante, para ver si es o no significativo, se calcula su varianza:

$$s_\kappa^2 = \frac{1}{360(1-0{,}333)^2}\left[0{,}333 + 0{,}333^2 - \frac{43{,}33\cdot 250+40\cdot 240+36{,}67\cdot 230}{360^2}\right] = 0{,}00138$$

con lo que $\kappa/s_\kappa = 8{,}959$ que no se encuentra dentro del intervalo de aceptación $[-1{,}96; 1{,}96]$ de la distribución $N(0;1)$ correspondiente al nivel $\alpha = 0{,}05$.

7.3. ANÁLISIS DE TABLAS 2×2

Dentro de las tablas de contingencia tiene un interés especial, tanto en Bioestadística como en aplicaciones epidemiológicas, el estudio de tablas $2x2$, que se representan de la forma:

$\backslash B$ / $A\backslash$	B_1	B_2	Frec. marg. A
A_1	n_{11}	n_{12}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$n_{2\bullet}$
Frec. marg. B	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Tal es el caso en el que se quiere decidir si la aplicación de una cierto tratamiento supone una mejoría en el estado del paciente; o si la diabetes tiene mayor incidencia en mujeres que en hombres. En tal caso, el estadístico del contraste para independencia frente a asociación se calcula abreviadamente de la forma:

$$\chi^2_{\text{exp}} = \frac{n(n_{11}\cdot n_{22} - n_{12}\cdot n_{21})^2}{n_{1\bullet}\cdot n_{2\bullet}\cdot n_{\bullet 1}\cdot n_{\bullet 2}}$$

el cual se distribuye según una ley *chi-cuadrado* con 1 grado de libertad. Así, para el nivel estándar $\alpha = 0{,}05$ el valor crítico es $3{,}841$, por lo que cuando el estadístico χ^2_{exp} tome valores menores se aceptará la hipótesis de independencia, y cuando sean mayores se rechazará.

La *corrección de Yates* sobre tablas 2×2, que se aplica cuando alguna $e_{ij} \leq 5$, se traduce en la siguiente expresión del estadístico de contraste:

$$\chi^2_{\text{exp}} = \frac{n(|n_{11}\cdot n_{22} - n_{12}\cdot n_{21}| - \frac{n}{2})^2}{n_{1\bullet}\cdot n_{2\bullet}\cdot n_{\bullet 1}\cdot n_{\bullet 2}}$$

En el caso de tablas 2×2 existen medidas específicas de asociación entre las variables, como son la Q y la Y de *Yule*, que se definen de la forma:

$$Q = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{n_{11} \cdot n_{22} + n_{12} \cdot n_{21}} \qquad Y = \frac{\sqrt{n_{11} \cdot n_{22}} - \sqrt{n_{12} \cdot n_{21}}}{\sqrt{n_{11} \cdot n_{22}} + \sqrt{n_{12} \cdot n_{21}}}$$

Ambas toman valores comprendidos entre -1 y 1.

EJEMPLO

Para estudiar la eficacia de la vacuna de la gripe estacional se eligieron dos grupos de personas, unos vacunados en otoño y otros no, y se comprobó si durante el invierno contraían o no la enfermedad, obteniéndose los siguientes resultados:

Vacunados / Enfermedad	*Sí*	*No*	Frec. marg.
Sí	3	10	13
No	47	40	87
Frec. marg.	50	50	100

El estadístico chi-cuadrado será:

$$\chi^2_{\text{exp}} = \frac{100(3 \cdot 40 - 10 \cdot 47)^2}{13 \cdot 87 \cdot 50 \cdot 50} = 4,332$$

que supera el valor crítico $3,841$, por lo que se rechaza la hipótesis nula y se concluye que hay asociación entre estar vacunado y contraer gripe. Sin embargo, aunque en este caso no sea necesario pues $e_{11} = 6,5$, si se hubiera aplicado la corrección de Yates se obtendría:

$$\chi^2_{\text{exp}} = \frac{100(|3 \cdot 40 - 10 \cdot 47| - 50)^2}{13 \cdot 87 \cdot 50 \cdot 50} = 3,183$$

y no podría rechazarse la hipótesis de independencia. Los coeficientes de Yule son: $Q = -0,593$ $Y = -0,329$

7.4. TEST EXACTO DE FISHER

El test chi-cuadrado para contrastar la hipótesis de independencia, al ser asintótico, puede aplicarse cuando el tamaño muestral es grande. Cuando esto no ocurre en una tabla de contingencia 2×2 y todas las casillas, salvo una a lo sumo, tienen frecuencias teóricas $e_{ij} \leq 5$, una alternativa es la aplicación del *test exacto de Fisher* que sobre la tabla

B \ A	B_1	B_2	Frec. marg. A
A_1	n_{11}	n_{12}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$n_{2\bullet}$
Frec. marg. B	$n_{\bullet 1}$	$n_{\bullet 2}$	n

se plantea de la forma siguiente:

Paso 1^o. Se obtiene el valor $p_0 = \frac{n_{1\bullet}!n_{2\bullet}!n_{\bullet 1}!n_{\bullet 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$. Si es $p_0 \geqslant 0,05$ entonces se acepta la hipótesis nula de independencia. En caso de ser $p_0 < 0,05$ se actúa según el paso siguiente.

Paso 2^o. Se forman todas las tablas 2×2 que conserven las frecuencias marginales:

B \ A	B_1	B_2	Frec. marg. A
A_1	$n_{11} \pm m$	$n_{12} \mp m$	$n_{1\bullet}$
A_2	$n_{21} \mp m$	$n_{22} \pm m$	$n_{2\bullet}$
Frec. marg. B	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Paso 3^o. Para cada tabla así obtenida se calcula el correspondiente valor $p_i = \frac{n_{1\bullet}!n_{2\bullet}!n_{\bullet 1}!n_{\bullet 2}!}{n!(n_{11}\pm m)!(n_{12}\mp m)!(n_{21}\mp m)!(n_{22}\pm m)!}$, $i = 1, 2, ...$

Paso 4^o. A p_0 se le suman los valores p_i menores o iguales a él.

Paso 5^o. Si el resultado de la suma es mayor o igual al nivel de significación $\alpha = 0,05$ se acepta la hipótesis de independencia entre A y B; en caso contrario, se rechaza y se concluye que la asociación es significativa.

EJEMPLO

Se quiere estudiar si una enfermedad *rara* es o no hereditaria, es decir, si existe asociación entre que la padezca, al menos, uno de los progenitores y alguno de sus hijos. Para ello se realizó un estudio con un grupo de 7 personas, de las cuales 4 padecían la enfermedad y los resultados fueron los siguientes:

Hijos / Progenitores	Enfermos	Sanos	Frec. marg. progen.
Enfermos	4	0	4
Sanos	0	3	3
Frec. marg. hijos	4	3	7

El valor de p_0 es:

$$p_0 = \frac{4!3!4!3!}{7!4!0!0!3!} = 0{,}0286$$

Las distintas tablas que pueden obtenerse a partir de la básica conservando las mismas frecuencias marginales, y los p_i asociados, son las siguientes:

Hijos / Progenitores	Enfermos	Sanos	Frec. marg. progen.
Enfermos	3	1	4
Sanos	1	2	3
Frec. marg. hijos	4	3	7

$$p_1 = \frac{4!3!4!3!}{7!3!1!1!2!} = 0{,}3429$$

Hijos / Progenitores	Enfermos	Sanos	Frec. marg. progen.
Enfermos	2	2	4
Sanos	2	1	3
Frec. marg. hijos	4	3	7

$$p_2 = \frac{4!3!4!3!}{7!2!2!2!1!} = 0{,}5143$$

Hijos / Progenitores	Enfermos	Sanos	Frec. marg. progen.
Enfermos	1	3	4
Sanos	3	0	3
Frec. marg. hijos	4	3	7

$$p_3 = \frac{4!3!4!3!}{7!1!3!3!0!} = 0{,}1143$$

Como todos los p_i son mayores que p_0 resulta:

$$p = p_0 = 0,0286$$

que al ser menor que $\alpha = 0,05$ se rechaza la hipótesis nula de independencia y se concluye, al nivel de significación estándar, que hay asociación entre ambas variables, es decir, que la enfermedad tiene carácter hereditario.

EJEMPLO

Se quiere comprobar si la técnica quirúrgica utilizada para cierto tipo de operación influye en el resultado, de forma que sobre una muestra de 13 pacientes intervenidos en un hospital se otuvieron los siguientes resultados:

Técnica	Resultado Positivo	Resultado Negativo	Frec. marg. técnica
Incisión	4	1	5
Laparoscopia	1	6	7
Frec. marg. resultado	5	7	12

El valor de p_0 es:

$$p_0 = \frac{5!7!5!7!}{12!4!1!1!6!} = 0{,}0442$$

Las distintas tablas que pueden obtenerse a partir de la básica conservando las mismas frecuencias marginales, y los p_i asociados, son las siguientes:

Técnica	Positivo	Negativo	Frec. marg. técnica
Incisión	5	0	5
Laparoscopia	0	7	7
Frec. resultado	5	7	12

$$p_1 = \frac{5!7!5!7!}{12!5!0!0!5!} = 0{,}0013$$

Técnica	Positivo	Negativo	Frec. marg. técnica
Incisión	3	2	5
Laparoscopia	2	5	7
Frec. resultado	5	7	12

$$p_2 = \frac{5!7!5!7!}{12!3!2!2!5!} = 0{,}2652$$

Técnica	Positivo	Negativo	Frec. marg. técnica
Incisión	2	3	5
Laparoscopia	3	4	7
Frec. resultado	5	7	12

$$p_3 = \frac{5!7!5!7!}{12!2!3!3!4!} = 0{,}4419$$

Técnica	Positivo	Negativo	Frec. marg. técnica
Incisión	1	4	5
Laparoscopia	4	3	7
Frec. resultado	5	7	12

$p_4 = \frac{5!7!5!7!}{12!1!4!4!3!}$ =0,2210

Técnica	Positivo	Negativo	Frec. marg. técnica
Incisión	0	5	5
Laparoscopia	5	2	7
Frec. resultado	5	7	12

$p_5 = \frac{5!7!5!7!}{12!0!5!5!2!}$ =0,0265

Por tanto, sumando a p_0 los valores de p_i menores o iguales a él resulta:

$$p = 0,0442 + 0,0013 + 0,0265 = 0,072$$

que al ser mayor que $\alpha = 0,05$ se acepta la hipótesis nula de independencia. Es decir, no podemos concluir al nivel de significación estándar que haya diferencias en los resultados según el tipo de intervención realizada.

7.5. TEST DE McNEMAR

Tanto el test chi-cuadrado como el exacto de Fisher parten de que las categorías de amboas variables son excuyentes entre si, es decir, un paciente no puede estar sano y enfermo a la vez, o una misma intervención no puede realizarse simultáneamente mediante dos técnicas quirúrgicas. Sin embargo, en ocasiones, es la misma característica la que se registra en dos ocasiones o situaciones diferentes, como, por ejemplo, si un paciente tiene fiebre en estado basal y transcurrida una hora desde la administración de un antipirético. Así, el test de McNemar es una alternativa a los anteriores aplicable cuando las modalidades de cada una de las dos variables consideradas en el estudio son pareadas. El esquema es el siguiente:

Variable B / Variable A	Positivo	Negativo	Frec. marg. A
Positivo	n_{11}	n_{12}	$n_{1\bullet}$
Negativo	n_{21}	n_{22}	$n_{2\bullet}$
Frec. marg. B	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Para contrastar la hipótesis nula de que A y B son independientes se evalúa el estadístico:

$$\chi^2_{\text{exp}} = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$$

que, bajo hipótesis de independencia se distribuye según una ley chi-cuadrado con 1 grado de libertad. Cuando el tamaño muestral es pequeño, digamos $n < 30$, el estadístico anterior se corrige por continuidad de la forma:

$$\chi^2_{\text{exp}} = \frac{(|n_{12}-n_{21}|-1)^2}{n_{12}+n_{21}}$$

EJEMPLO

Se ha registrado sobre una muestra de 20 personas que han estado en contacto con seropositivos de covid-19 el resultado de un test serológico en el momento inicial y transcurridos 10 días, siendo los resultados pareados los siguientes:

Instante inicial	Sí	No	Sí	No	No	No	No	Sí	No	No
Transcurridos 10 días	No	Sí	Sí	Sí	No	Sí	Sí	Sí	Sí	No
Instante inicial	No	No	Sí	No	Sí	No	Sí	Sí	No	No
Transcurridos 10 días	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No	Sí	Sí

Podemos representar estos resultados en una tabla 2×2 de la forma:

Instante inicial / Transcurridos 10 días	Sí	No	Frec. marg. 10 días
Sí	5	11	16
No	2	2	4
Frec. marg. inicial	7	13	20

Dado que el número de muestras es pequeño ($n = 20$) el estadístico de contraste se calculará de la forma:

$$\chi^2_{\text{exp}} = \frac{(|11-2|-1)^2}{11+2} = 4,923$$

y como para el nivel estándar $\alpha = 0,05$ el valor crítico de la χ^2 con 1 grado de libertad es $3,841$, se rechaza la hipótesis de independencia y se concluye que, a dicho nivel, existe asociación entre el resultado del test en el estado inicial y transcurridos 10 días.

7.6. APLICACIONES EN EPIDEMIOLOGÍA

Un estudio epidemiológico se desarrolla en dos fases: la fase de evaluación o prueba, que se realiza mediante estudios caso-control, y la de diagnóstico o predicción, que se lleva a cabo mediante estudio de cohortes. En esta sección vamos a considerar dos situaciones usuales en tales estudios, que se plantean en términos de tablas de contingencia: los estudios de sensibilidad y especificidad de una prueba diagnóstica, con la consiguiente curva ROC, y el cálculo del riesgo relativo y la razón de producto cruzado de una enfermedad frente a un factor de riesgo.

7.6.1. Sensibilidad y especificidad de una prueba: Curva ROC

Se denomina *sensibilidad* de una prueba a la probabilidad de que detecte la enfermedad cuando el paciente realmente esté enfermo; y *especificidad* es la probabilidad de que clasifique a un paciente como sano cuando no tiene la enfermedad. Es decir, si denotamos E al suceso estar enfermo y el resultado de la prueba puede ser $+$ o $-$, entonces:

$$\text{Sensibilidad} = Prob(+/E) \qquad \text{Especificidad} = Prob(-/\bar{E})$$

Ambos conceptos, que se suelen expresar en términos porcentuales, son igualmente importantes a la hora de evaluar un test diagnóstico, y no basta con que una de ellas sea muy alta (próxima a 1). Así, por ejemplo, una prueba que siempre de positiva tendrá una sensibilidad del 100% pues detectará a la totalidad de enfermos, pero también clasificará erróneamente como tales a los que estén libres de la enfermedad. Evidentemente toda prueba diagnóstica está sujeta a posibles errores, pudiendo dar lugar a falsos positivos y falsos negativos, cuyas probabilidades se denotan respectivamente mediante $\alpha = Prob(+/S)$ y $\beta = Prob(-/E)$, a similitud de los contrastes de hipótesis.

Así, el nivel de significación equivale a la probabilidad de falsos positivos, mientras que la potencia del contraste es ahora la sensibilidad de la prueba.

Al porcentaje de enfermos dentro de la población que se estudia o, equivalentemente, a la probabilidad de que un individuo de dicha población esté enfermo $Prob(E)$, se le denomina *prevalencia* de la enfermedad. A su vez, se define el valor predictivo positivo como la probabilidad de que un paciente esté realmente enfermo cuando el test da positivo, es decir $Prob(E/+)$, y el valor predictivo negativo como la probabilidad de que un paciente esté sano cuando el test da negativo, es decir $Prob(S/-)$.

Si se dispone de un estudio realizado sobre n individuos, los resultados pueden expresarse mediante una tabla de doble entrada como la siguiente:

Test	Enfermos	Sanos	Totales
$+$	n_{11}	n_{12}	$n_{1\cdot}$
$-$	n_{21}	n_{22}	$n_{2\cdot}$
Totales	$n_{\cdot 1}$	$n_{\cdot 2}$	n

La sensibilidad y la especificidad se estiman por medio de:

$$\widehat{Prob}(+/E) = \tfrac{n_{11}}{n_{\cdot 1}} \quad \widehat{Prob}(-/S) = \tfrac{n_{22}}{n_{\cdot 2}},$$

y la prevalencia por $\widehat{Prob}(E) = \frac{n_{\cdot 1}}{n}$, por lo que $\widehat{Prob}(S) = 1 - \widehat{Prob}(E)$. A su vez, la estimación de los valores predictivos será:

$$\widehat{Prob}(E/+) = \tfrac{n_{11}}{n_{1\cdot}} \quad \widehat{Prob}(S/-) = \tfrac{n_{22}}{n_{2\cdot}}$$

EJEMPLO

Se ha realizado un test diagnóstico a 100 trabajadores de una gran empresa con el fin de detectar una enfermedad contagiosa durante una epidemia habiendo obtenido los siguientes resultados:

Test	Enfermos	Sanos	Totales
$+$	30	3	33
$-$	5	62	67
Totales	35	65	100

La prevalencia estimada de la enfermedad es $\widehat{Prob}(E) = 0,35$ y la sensibilidad y especificidad estimadas:

$$\widehat{Prob}(+/E) = \tfrac{30}{35} = 85,7\% \quad \widehat{Prob}(-/S) = \tfrac{62}{65} = 95,4\%$$

A su vez, los valores predictivos positivo y negativo estimados son:

$$\widehat{Prob}(E/+) = \tfrac{30}{33} = 90,9\% \quad \widehat{Prob}(S/-) = \tfrac{62}{67} = 92,5\%$$

Con objeto de evaluar la bondad de una prueba, o bien, comparar dos pruebas para elegir la óptima existe una metodología que consiste en represenar en dos ejes coordenados la sensibilidad (ordenada) frente al complementario de la especificidad (abcisa), es decir, representa la probabilidad de + cuando el paciente está enfermo frente a cuando a está sano (falso positivo). La representación obtenida se denomina *curva ROC* (*Receiver Operating Characteristic*) y el área bajo dicha curva es la medida de la fiabilidad de la prueba.

EJEMPLO

Para diagnosticar el infarto de miocardio se dispone de dos pruebas; una es simple y consiste en medir el ritmo cardiaco en pulsaciones por minuto, y la otra es determinar el nivel de colesterol LDL del paciente en mg/dL. Así, se seleccionan 20 pacientes, 10 de ellos que han sufrido recientemente un infarto y 10 que no, obteniéndose los siguientes datos ordenados de forma creciente :

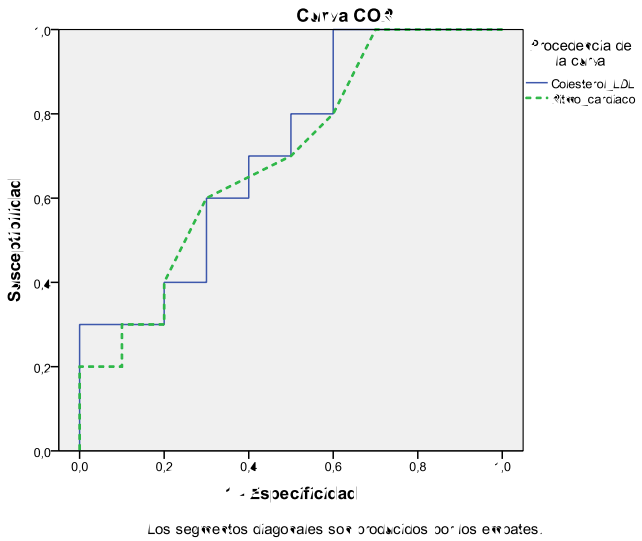
Nivel de colesterol LDL

Sin infarto	95	101	103	108	115	122	128	136	142	145
Con infarto	111	113	118	124	135	135	140	152	161	172

Ritmo cardiaco

Sin infarto	60	62	62	66	68	70	70	72	76	80
Con infarto	66	66	68	70	72	72	74	78	84	90

Para obtener las dos curvas ROC es necesario ir fijando unos puntos de corte, que son valores en ambas pruebas a partir de los cuales se considera que hay



enfermedad. Por ejemplo, si se fija 130mg/dL como punto de corte para calibrar la prueba de colesterol LDL como indicador de infarto, habría 6 pacientes infartados y 3 no infartados con niveles superiores, por lo que la sensibilidad de la prueba sería 6/10=0,6 y la especificidad 7/10=0,3. Así, podríamos construir unas tablas para cada uno de los dos criterios diagnósticos, considerando en cada caso diferentes puntos de corte:

Punto de corte Colesterol LDL	E	$1-E$	S	Punto de corte Ritmo cardiaco	E	$1-E$	S
150	1	0	0,3	85	1	0	0,1
140	0,8	0,2	0,4	80	0,9	0,1	0,2
130	0,7	0,3	0,6	75	0,8	0,2	0,3
120	0,5	0,5	0,7	70	0,5	0,5	0,7
110	0,4	0,6	1	65	0,3	0,7	1

La elección de los puntos de corte es libre y, si el cálculo del área bajo la curva ROC se realiza manualmente, mientras mayor sea el número de estos puntos mayor será la precisión de cálculo, aunque también mayor el esfuerzo de cálculo. En la representación de S frente a $1-E$ obtendríamos la siguientes áreas:

Curva Roc del colesterol LDL:

Área=0,2$\frac{0,3+0,4}{2}$+(0,3-0,2)$\frac{0,4+0,6}{2}$+(0,5-0,3)$\frac{0,6+0,7}{2}$+(0,6-0,5)$\frac{0,7+1}{2}$+(1-0,6)1=0,735

Curva Roc del ritmo cardiaco:

Área=0,1$\frac{0,1+0,2}{2}$+(0,2-0,1)$\frac{0,2+0,3}{2}$+(0,5-0,2)$\frac{0,3+0,7}{2}$+(0,7-0,5)$\frac{0,7+1}{2}$+(1-0,7)1=0,660
El nivel de colesterol LDL es, por tanto, mejor indicador del infarto quen el ritmo cardiaco. En el gráfico siguiente pueden apreciarse las curvas ROC para las dos pruebas diagnósticas.

7.6.2. Riesgo relativo y razón de producto cruzado (*odd-ratio*)

Dos conceptos de gran utilidad al realizar estudios de cohortes son el *riesgo relativo* (RR), que es el ratio entre las probabilidades de enfermedad condicionada a dar el test resultado positivo y resultado negativo respectivamente, y la *razón de producto cruzado* u *odd-ratio* (OR), que es la relación entre el ratio de las probabilidades de estar enfermo y sano para resultado del test positivo y para resultado del test negativo, es decir:

$$RR = \frac{P(E/+)}{P(E/-)} \; ; \quad OR = \frac{\frac{P(E/+)}{P(S/+)}}{\frac{P(E/-)}{P(S/-)}}$$

A patir de una tabla como la anterior se estiman mediante:

$$\widehat{RR} = \frac{\frac{n_{11}}{n_{1\cdot}}}{\frac{n_{21}}{n_{2\cdot}}} = \frac{n_{11}n_{2\cdot}}{n_{21}n_{1\cdot}} \; ; \; \widehat{OR} = \frac{\frac{n_{11}/n_{1\cdot}}{n_{12}/n_{1\cdot}}}{\frac{n_{21}/n_{2\cdot}}{n_{22}/n_{2\cdot}}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

La *odd-ratio* se puede interpretar como el riesgo relativo de una enfermedad rara, ya que al ser poco frecuente se verificará que $n_{12} \simeq n_{1\bullet}$ y $n_{22} \simeq n_{2\bullet}$. Cuando $\widehat{RR}$ sea mayor que la unidad indicará que el factor considerado es de riesgo; e igual interpretación cabe hacer para $\widehat{OR}$

Estimadores de la varianza del logaritmo neperiano del riego relativo y de la *odd-ratio* vienen dados por:

$$V\left[\ln \widehat{RR}\right] = \frac{n_{21}}{n_{11}n_{\bullet 1}} + \frac{n_{22}}{n_{12}n_{\bullet 2}} \; ; \; V\left[\ln \widehat{OR}\right] = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Así, si asumimos que tanto $\widehat{RR}$ como $\widehat{OR}$ siguen una distribución log-normal, podremos contrastar la significación de $\ln RR$ y de $\ln OR$ evaluando respectivamente los estadísticos:

$$\frac{\ln \widehat{RR}}{\sqrt{V\left[\ln \widehat{RR}\right]}} \quad \text{y} \quad \frac{\ln \widehat{OR}}{\sqrt{V\left[\ln \widehat{OR}\right]}}$$

Así cuando ambos tomen valores inferiores a $1,96$ concluiremos que no son significativos al nivel $\alpha = 0,05$ y, por tanto aceptaremos que $RR = 1$ y $OR = 1$. En caso contrario concluiremos que el factor considerado es de riesgo.

EJEMPLO

Se quiere determinar si la exposición prolongada al sol (más de 4 horas diarias) es un factor de riesgo del glaucoma, para lo cual se reunieron datos de los últimos diez años del servicio de Dermatología de un hospital y se sintetizaron en la tabla siguiente:

Glaucoma	Sí	No	Totales
Exposición alta	45	520	565
Exposición baja	102	1545	1647
Totales	147	2065	2212

El riego relativo y la *odd-ratio* se calculan de la forma:

$$\widehat{RR} = \frac{45/565}{102/1647} = 1,286 \quad ; \quad \widehat{OR} = \frac{45 \cdot 1545}{520 \cdot 102} = 1,311$$

A su vez, $\ln \widehat{RR} = 0,2515$ y $\ln \widehat{OR} = 0,2708$ siendo sus varianzas:

$$V\left[\ln \widehat{RR}\right] = \frac{520}{45 \cdot 565} + \frac{1545}{102 \cdot 1647} = 0,02965 \quad ;$$
$$V\left[\ln \widehat{RR}\right] = \frac{1}{45} + \frac{1}{520} + \frac{1}{102} + \frac{1}{1545} = 0,0346$$

por lo que

$$\frac{\ln \widehat{RR}}{\sqrt{V\left[\ln \widehat{RR}\right]}} = \frac{0,2515}{\sqrt{0,02965}} = 1,461 \quad \text{y} \quad \frac{\ln \widehat{OR}}{\sqrt{V\left[\ln \widehat{OR}\right]}} = \frac{0,2708}{\sqrt{0,0346}} = 1,456$$

y, al ser valores inferiores a $1,96$, concluimos que ni $\ln RR$ ni $\ln OR$ son significativos, o equivalentemente que podemos aceptar al nivel $\alpha = 0,05$ que $RR = 1$ y $OR = 1$, es decir, la exposición prolongada al sol no es un factor significativo en la aparición del glaucoma.