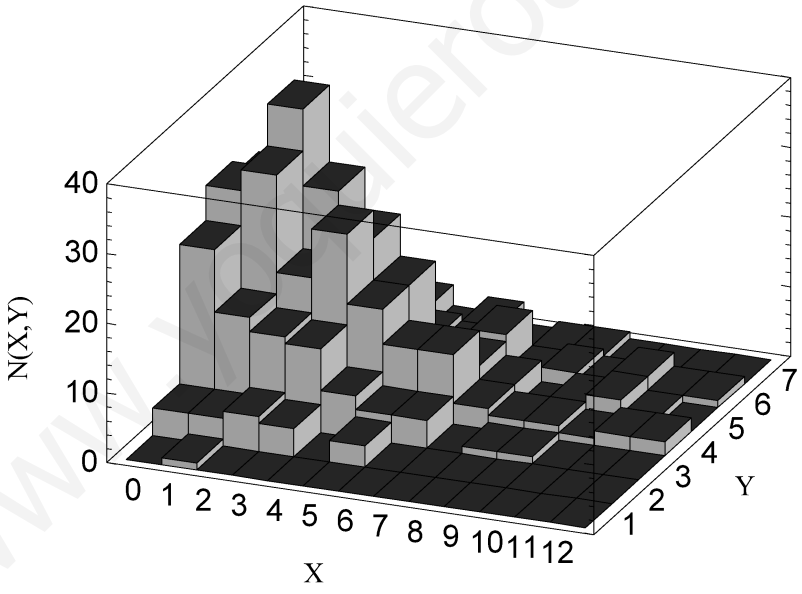
Estadística:
Apuntes para Estudiantes Universitarios



M. Ángeles Gómez Flechoso

ISBN: 978-84-695-3587-5

Índice

Capítulo 1

Leyes de probabilidad
Probabilidad. Probabilidad condicionada. Sucesos independientes. Teorema de Bayes

1.1 Experiencias aleatorias

- **Fenómenos o experiencias deterministas**: Se definen como tales aquellos que cumplen que
  - el resultado es totalmente previsible
  - realizado en las mismas condiciones sólo hay un resultado posible
  - ej.: lanzar una moneda con dos caras
- **Fenómenos o experiencias aleatorias:** Son aquellos que cumplen que
  - su resultado depende del azar (es necesaria su intervención para conocer el resultado)
  - pueden ser repetidos en las mismas condiciones
  - es posible determinar el conjunto de todos los resultados posibles: *espacio muestral*
  - ej.: lanzar un dado con seis caras distintas

1.2 Álgebra de sucesos

Consideremos E una experiencia aleatoria, definimos el *espacio muestral* como el conjunto de todos los resultados posibles.

$$\Omega = \{\omega / \omega \text{ es un resultado posible}\}$$

Ejemplos:

- Espacios muestrales discretos:
  - Tirar un dado: $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Tirar una moneda una vez: $\Omega = \{C, R\}$
  - Tirar una moneda dos veces: $\Omega = \{C, R\} \times \{C, R\} = \{CC, CR, RC, RR\}$
- Espacios muestrales continuos:
  - Ángulo de la aguja de un reloj con el eje positivo de abcisas: $\Omega = [0, 2\pi)$
  - Módulo de la velocidad de un móvil: $\Omega = [0, \infty)$

Definimos *suceso*, S, como cualquier circunstancia que una vez realizada la experiencia podemos decir si ha tenido lugar o no. A todo suceso se le puede asociar un subconjunto del espacio muestral.
Decimos que un suceso se verifica si el resultado ω de la experiencia es tal que $\omega \in S$

Suceso complementario o contrario a S: Denominado suceso complementario, $\bar{S}$ o S^c, de un suceso S a aquel que se verifica cuando no se verifica S.

Suceso imposible: Un suceso se denomina imposible cuando no se verifica nunca, $S_I = \emptyset$

Suceso seguro: Un suceso es seguro cuando se verifica siempre.

Unión de sucesos: Definimos la unión de n sucesos S_i, $i = 1, \ldots n$, y denotamos por $\bigcup_{i=1}^{n} S_i$, como el suceso que se verifica si y sólo si se ha realizado al menos un S_i

Intersección de sucesos: Definimos la intersección de n sucesos S_i, $i = 1, \ldots n$, y denotamos por $\bigcap_{i=1}^{n} S_i$, como el suceso que se verifica si y sólo si se ha realizado todos los S_i

Si se cumple que $\bigcap_{i=1}^{n} S_i = \emptyset$ decimos que los sucesos S_i son incompatibles.

Si se cumple que $S_1 \subset S_2$, tendremos que si se verifica S_1 también se verifica S_2

Ejemplo:

Experiencia de lanzar tres monedas al aire:

- Espacio muestral: $\Omega = \{CCC, CCR, CRC, RCC, CRR, RCR, RRC, RRR\}$
- Sucesos: $S_1 = \{\text{que salga una sola cara}\} = \{CRR, RCR, RRC\} \subset \Omega$;
  $S_2 = \{\text{que salga una cara o ninguna}\} = \{CRR, RCR, RRC, RRR\}$;
  $S_3 = \{\text{que todos los resultados sean iguales}\} = \{CCC, RRR\}$

- Suceso complementario a S_1: $\bar{S_1} = \{CCC, CCR, CRC, RCC, RRR\}$
- Suceso imposible, por ejemplo, $S_I = \{\text{que salgan cuatro caras}\} = \emptyset$
- Suceso seguro, por ejemplo, $S_S = \{\text{que salga, al menos, una cara o una cruz}\} = \Omega$
- $S_2 \bigcup S_3 = \{CRR, RCR, RRC, RRR, CCC\}$
- $S_1 \bigcap S_3 = \emptyset$ por lo tanto S_1 y S_3 con incompatibles; $S_2 \bigcap S_3 = \{RRR\}$
- $S_1 \subset S_2$

1.3 Concepto de probabilidad

1.3.1 Definición

La *probabilidad* permite medir la certeza o incertidumbre de un suceso de una experiencia aleatoria. En la práctica toma valores entre 0 y 1, asignándose el valor 0 a un suceso imposible y el valor 1 a un suceso seguro.

Matemáticamente la probabilidad es una función que hace corresponder a cada suceso S un número real que cumple que

1. $0 \leq p(S) \leq 1$, $\forall S$
2. $p(\Omega) = 1$
3. Si S_1 y S_2 son dos sucesos incompatibles (esto es, $S_1 \bigcap S_2 = \emptyset$), entonces $p(S_1 \bigcup S_2) = p(S_1) + p(S_2)$

Ejemplo:
Experimento E: lanzar un dado equilibrado
Espacio muestral $\Omega = \{1, 2, 3, 4, 5, 6\}$
Sucesos: $S_1 = \{\text{que salga par}\} = \{2, 4, 6\}$; $S_2 = \{6\}$; $\bar{S_2} = \{1, 2, 3, 4, 5\}$; $S_3 = \{2\}$
Probabilidades: $p(S_1) = \frac{1}{2}$; $p(S_2) = \frac{1}{6}$; $p(S_2) = \frac{1}{6}$; $p(S_2 \bigcup S_3) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$ ya que $S_2 \bigcap S_3 = \emptyset$; $p(\bar{S_2}) = 1 - \frac{1}{6} = \frac{5}{6}$ ya que $p(\Omega) = 1$ y $S_2 \bigcap \bar{S_2} = \emptyset$ y $S_2 \bigcup \bar{S_2} = \Omega$; $p(\emptyset) = 0$

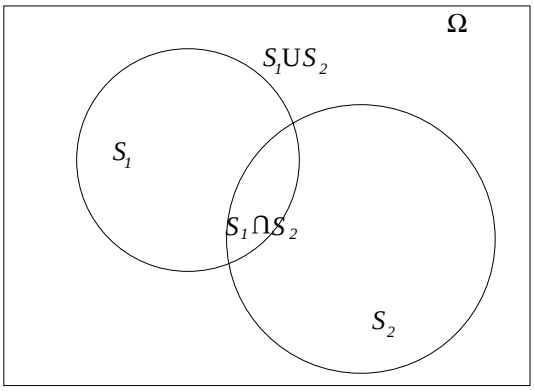


Figura 1.1: *Esquema de la unión e intersección de sucesos*

1.3.2 Propiedades

1. $p(S) = 1 - p(\bar{S})$
2. $p(\emptyset) = 0$
3. $p(\Omega) = 1$
4. $p(S_1 \bigcup S_2) = p(S_1) + p(S_2) - p(S_1 \bigcap S_2)$
5. $p(S_1 \bigcup S_2 \bigcup S_3) = p(S_1)+p(S_2)+p(S_3)-p(S_1 \bigcap S_2)-p(S_2 \bigcap S_3)-p(S_1 \bigcap S_3)+p(S_1 \bigcap S_2 \bigcap S_3)$

Ejemplo:

Considerando el experimento E anterior de lanzar un dado equilibrado, y definiendo los sucesos $S_1 = \{\text{que salga par}\} = \{2, 4, 6\}$ y $S_4 = \{\text{múltiplo de 3}\} = \{3, 6\}$, tendremos que $S_1 \bigcup S_4 = \{\text{par o múltiplo de 3}\} = \{2, 3, 4, 6\}$ y $S_1 \bigcap S_4 = \{\text{par y múltiplo de 3}\} = \{6\}$. Por lo tanto, $p(S_1) = \dfrac{1}{2}$ y $p(S_4) = \dfrac{1}{3}$, así que tendremos que las probabilidades de la intersección y de la unión serán, respectivamente, $p(S_1 \bigcap S_4) = \dfrac{1}{6}$ y $p(S_1 \bigcup S_4) = p(S_1) + p(S_4) - p(S_1 \bigcap S_4) = \dfrac{1}{2} + \dfrac{1}{3} - \dfrac{1}{6} = \dfrac{2}{3}$

1.3.3 Probabilidad en espacios muestrales discretos

Vamos a considerar, de momento, espacios muestrales discretos equiprobables, en ese caso, la probabilidad de un suceso S se puede obtener como $p(S) = \dfrac{\#\text{casos favorables}}{\#\text{casos posibles}} = \dfrac{\#S}{\#\Omega}$

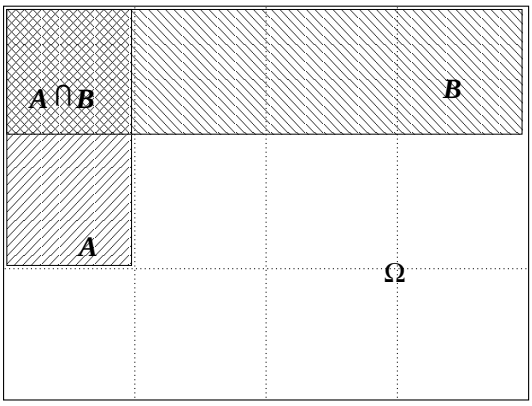


Figura 1.2: *Esquema para el cálculo de la probabilidad condicionada*

1.3.4 Probabilidad en espacios muestrales continuos

Consideremos un espacio muestral continuo equiprobable, en este caso, los sucesos elementales son igualmente probables, pero tienen una probabilidad 0. Sólo tiene sentido hablar de probabilidad de un subintervalo del espacio muestral, así si consideramos el intervalo $I = [a, b] \subset \Omega$, tendremos que $p(I) \propto (b - a)$, como queremos que $p(\Omega) = 1$, entonces tiene que cumplirse que $p(I) = \dfrac{b-a}{\text{amplitud de } \Omega}$

Ejemplo:

Si $\Omega = [0, 2\pi)$, tendremos que dado $I = [a, b]$, entonces $p(I) = \dfrac{b-a}{2\pi} = \displaystyle\int_a^b \frac{dx}{2\pi} = \int_a^b f(x)dx$

con $f(x) = \begin{cases} \dfrac{1}{2\pi} & \text{si } 0 \leq x < 2\pi \\ 0 & \text{resto} \end{cases}$

1.4 Probabilidad condicionada

La *probabilidad condicionada* permite conocer la probabilidad de un suceso A en el caso de que se haya cumplido otro suceso B, y se escribe como $p(A|B)$. Dicha probabilidad condicionada se calcula como:

$$p(A|B) = \frac{p(A \bigcap B)}{p(B)}$$

Ejemplo:

Consideremos el experimento de lanzar un dado equilibrado. Si definimos los sucesos $A =$ {que salga par} $= \{2, 4, 6\}$ y $B =$ {que salga primo} $= \{1, 2, 3, 5\}$, tendremos que, $A \bigcap B =$

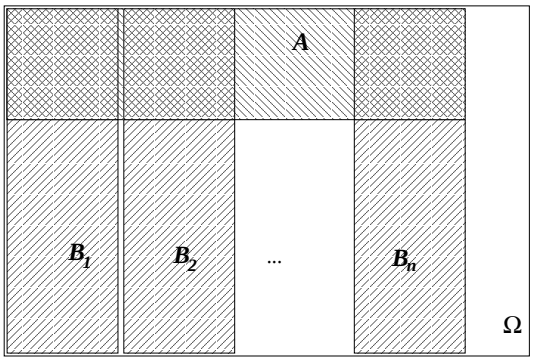


Figura 1.3*: Esquema de sucesos para la aplicación del teorema de Bayes*

{2}

Calcularemos ahora la probabilidad de que salga par, sabiendo que ha sido primo, esto es $p(A|B)$.

Dicha probabilidad podemos calcularla, por un lado como $p(A|B) = \dfrac{\#\text{casos favorables}}{\#\text{casos posibles}} = \dfrac{1}{4}$, ya que el número de casos posibles ahora es 4, puesto que se verifica B.

Por otro lado, también podríamos haber calculado dicha probabilidad como $p(A|B) = \dfrac{p(A \bigcap B)}{p(B)} = \dfrac{1/6}{4/6} = \dfrac{1}{4}$, obteniéndose, como era de esperar, el mismo resultado.

1.5 Teorema de Bayes

Vamos a considerar ahora n sucesos incompatibles dos a dos, $B_1, B_2, \ldots, B_n$, por lo tanto tendremos que $B_i \bigcap B_j = \emptyset \ \forall i \neq j$, y consideremos otro suceso A, cuya realización implica la realización de algunos de los sucesos B_i, esto es, $A \subset B = B_1 \bigcup B_2 \bigcup \cdots \bigcup B_n$. En este caso, tenemos que

$$p(A) = p(A \bigcap B_1) + p(A \bigcap B_2) + \cdots + p(A \bigcap B_n) = \sum_{i=1}^{n} p(A \bigcap B_i) = \sum_{i=1}^{n} p(A|B_i)p(B_i)$$

ya que $p(A|B_i) = \dfrac{p(A \bigcap B_i)}{p(B_i)}$

Teorema de Bayes:

Según los supuestos anteriores, esto es, considerando n sucesos incompatibles dos a dos, $B_1, B_2, \ldots, B_n$, con $B_i \bigcap B_j = \emptyset \ \forall i \neq j$, y un suceso A tal que $A \subset B = B_1 \bigcup B_2 \bigcup \cdots \bigcup B_n$,

se tiene que

$$p(B_j|A) = \frac{p(B_j \bigcap A)}{\sum_{i=1}^{n} p(A|B_i)p(B_i)} = \frac{p(B_j \bigcap A)}{p(A)}$$

Ejemplo:

Supongamos que tenemos dos bolsas con bolas blancas y rojas. En la primera bolsa hay diez bolas blancas y cuatro rojas y en la segunda bolsa hay doce bolas blancas y ocho rojas. Se elige una bolsa al azar y se extrae una bola. Calcular la probabilidad de que la bolsa elegida sea la segunda sabiendo que la bola que se ha extraído es blanca.

Dado que se eligen las bolsas al azar, la probabilidad de elegir la primera bolsa será igual a la probabilidad de elegir la segunda bolsa, o sea, $p(B_1) = p(B_2) = 1/2$.

Por otro lado, como en la primera bolsa hay 10 bolas blancas y 4 rojas, si se hubiese elegido la primera bolsa la probabilidad de que saliese blanca sería: $p(B|B_1) = \frac{10}{10+4} = \frac{5}{7}$

Por el contrario, como en la segunda bolsa hay 12 bolas blancas y 8 rojas, si se hubiese elegido la segunda bolsa la probabilidad de que saliese blanca sería: $p(B|B_2) = \frac{12}{12+8} = \frac{3}{5}$

Como queremos calcular $p(B_2|B)$, aplicando el teorema de Bayes tendremos que:

$p(B_2|B) = \frac{p(B_2 \bigcap B)}{p(B)} = \frac{p(B|B_2)p(B_2)}{p(B \bigcap B_1) + p(B \bigcap B_2)} = \frac{p(B|B_2)p(B_2)}{p(B|B_1)p(B_1) + p(B|B_2)p(B_2)} =$

$\frac{\frac{3}{5} \cdot \frac{1}{2}}{\frac{5}{7} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{2}} = \frac{21}{46} = 0.4565$

1.6 Sucesos dependientes e independientes

Dos sucesos A y B son *independientes* cuando la probabilidad de que uno ocurra no depende de que el otro haya ocurrido o no, esto es, $p(A) = p(A|B)$ y $p(B) = p(B|A)$.

Por lo tanto, como $p(A|B) = \frac{p(A \bigcap B)}{p(B)}$, si A y B son independientes tendremos que

$$p(A \bigcap B) = p(A)p(B)$$

Ejemplo:

Tenemos 4 bolas blancas y 6 bolas negras en una urna. Se extraen dos bolas sucesivamente, calcular la probabilidad de que ambas sean blancas:

Sea $S_1 = \{\text{primera bola blanca}\}$ y $S_2 = \{\text{segunda bola blanca}\}$

(a) Suponiendo que hay reemplazamiento después de la primera extracción: en este caso tendremos que $p(S_1) = \frac{4}{10} = \frac{2}{5}$ y $p(S_2) = \frac{4}{10} = \frac{2}{5}$, ya que S_1 y S_2 son sucesos independientes (al haber reemplazamiento). Por lo tanto, $p(S_1 \bigcap S_2) = p(S_1)p(S_2) = \frac{4}{25} = 0.16$

(b) Suponiendo que no hay reemplazamiento después de la segunda extracción: ahora tendremos que los sucesos S_1 y S_2 no son independientes, así que $p(S_1 \bigcap S_2) = p(S_1)p(S_2|S_1)$ y como tenemos que $p(S_1) = \frac{4}{10} = \frac{2}{5}$ y $p(S_2|S_1) = \frac{\#\text{casos favorables}}{\#\text{casos posibles}} = \frac{3}{9} = \frac{1}{3}$. Por lo tanto $p(S_1 \bigcap S_2) = \frac{2}{5} \cdot \frac{1}{3} = \frac{2}{15} = 0.13$

Problema:

Un concursante debe elegir una puerta entre tres posibles, sabiendo que detrás de una de ellas está el premio. Elegida la puerta y antes de abrirla, el presentador, que sabe en qué puerta está el premio, le muestra que detrás de una de las no escogidas no hay premio y le da la posibilidad de reconsiderar la elección, ¿qué debe hacer el concursante?

Si definimos $A_i = \{\text{el concursante elige la puerta } i\}$ y $R_i = \{\text{el premio está en la puerta } i\}$. El espacio muestral inicial son 9 sucesos posibles, $A_i \bigcap R_j$ con $i, j = 1, 2, 3$, equiprobables.

La probabilidad de acertar inicialmente es $p(R_i|A_i) = \frac{p(A_i \bigcap R_i)}{p(A_i)} = \frac{1/9}{1/3} = \frac{1}{3}$

Supongamos que el concursante elige la puerta 1 (A_1) y el presentador abre la puerta j que no tiene premio ($j = 2, 3$), denominaremos a este suceso, suceso B_j. Hay, por lo tanto, 4 sucesos posibles $\{(B_2 \bigcap R_1), (B_2 \bigcap R_3), (B_3 \bigcap R_1), (B_3 \bigcap R_2)\}$. Vamos a calcular las probabilidades de estos sucesos:

Como $p(R_1) = p(R_2) = p(R_3) = \frac{1}{3}$, si el concursante elige la puerta 1 (A_1) y el premio está en la puerta 1 (R_1) es igualmente probable que el presentador abra la puerta 2 o la puerta 3, por lo tanto, $p(B_2|R_1) = p(B_3|R_1) = \frac{1}{2} \Rightarrow p(B_2 \bigcap R_1) = p(B_2|R_1)p(R_1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} = p(B_3 \bigcap R_1) = p(B_3|R_1)p(R_1)$

Pero si el premio está en la puerta 2 (R_2) y el concursante ha elegido la puerta 1 (A_1), tendremos que $p(B_2|R_2) = 0$ y $p(B_3|R_2) = 1$, por lo tanto, $p(B_2 \bigcap R_2) = 0$ y $p(B_3 \bigcap R_2) = p(B_3|R_2)p(R_2) = \frac{1}{3}$. De forma similar se calcula para el caso de que el premio esté en la puerta 3.

Por lo tanto, tenemos el siguiente cuadro de probabilidades sabiendo que se ha verificado A_1

$\bigcap$	R_1	R_2	R_3
B_1	0	0	0
B_2	$1/6$	0	$1/3$
B_3	$1/6$	$1/3$	0

Se puede comprobar que $\sum_{i=1}^{9} p_i = 1$

La probabilidad de ganar que tiene el concursante que no cambia de elección es, supuesto que ha elegido la puerta 1 (A_1) y que el presentador ha abierto la puerta 2 (B_2)

$$p(R_1|B_2) = \frac{p(R_1 \bigcap B_2)}{p(B_2)} = \frac{p(R_1 \bigcap B_2)}{p(B_2|R_1)p(R_1) + p(B_2|R_2)p(R_2) + p(B_2|R_3)p(R_3)} = \frac{\frac{1}{6}}{\frac{1}{2}\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} = \frac{1}{3}$$

La probabilidad de ganar del concursante si cambia de puerta será $p = 1 - p(R_1|B_2) = 1 - \frac{1}{3} = \frac{2}{3}$. Vamos a comprobarlo:

$$p(R_3|B_2) = \frac{p(R_3 \bigcap B_2)}{p(B_2)} = \frac{p(R_3 \bigcap B_2)}{p(B_2|R_1)p(R_1) + p(B_2|R_2)p(R_2) + p(B_2|R_3)p(R_3)} = \frac{\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} = \frac{2}{3}$$

El motivo de que sea favorable cambiar de puerta es que la decisión del presentador de abrir una puerta, B_i, no es independiente, en general, de dónde esté el premio, R_j, y, por tanto, el suceso B_i nos da información sobre el suceso R_j.

Tenemos que $p(B_2) = p(B_3) = \frac{1}{2}$ y $p(R_1) = p(R_2) = p(R_3) = \frac{1}{3}$

El suceso R_1 es independiente de B_2 y B_3, por eso,

$p(R_1 \bigcap B_2) = p(R_1)p(B_2) = p(R_1 \bigcap B_3) = p(R_1)p(B_3) = \frac{1}{2}\cdot\frac{1}{3} = \frac{1}{6}$

Sin embargo, B_2 y B_3 no son independientes de R_2 y R_3 y por eso $p(R_2 \bigcap B_3) \neq p(R_2)p(B_3)$ y $p(R_3 \bigcap B_2) \neq p(R_3)p(B_2)$

Capítulo 2

Variables aleatorias

Variables aleatorias discretas y continuas. Función de densidad de probabilidad y función de distribución acumulativa de una variable aleatoria. Distribuciones bivariantes. Distribuciones marginales. Distribuciones condicionadas. Variables aleatorias independientes.

2.1 Variables aleatorias discretas y continuas

Vamos, en primer lugar, a definir una serie de conceptos:

- *Población:* Conjunto completo de elementos, con alguna característica común, que es el objeto del estudio. Puede ser finita o infinita. Por ejemplo, el número de habitantes de un país es una población finita, pero el número de medidas de la velocidad de la luz es una población infinita.

- *Muestra:* Un subconjunto de elementos de una población. El tamaño de la muestra es el número de elementos de la muestra. Un censo es una muestra de todos los elementos de la población.

- *Caracteres cuantitativos:* Son aquellos que toman valores numéricos, por ejemplo, la altura, el tiempo, etc

- *Caracteres cualitativos (o atributos):* Son aquellos que describen cualidades, por ejemplo, el color de pelo, estado civil, etc.

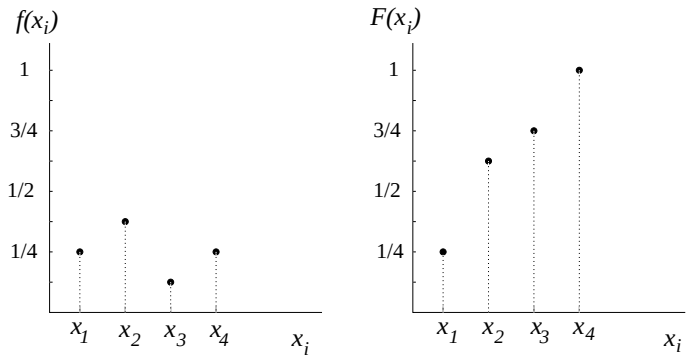


Figura 2.1: *Función de densidad de probabilidad, $f(x)$, y función de distribución acumulativa, $F(x)$, para una variable discreta.*

- *Variable estadística:* Es un símbolo que representa al dato o caracter del objeto de estudio y puede tomar un conjunto de valores. El conjunto de valores posibles de la variable aleatoria se denomina recorrido. Se utilizan letras mayúsculas (por ejemplo, X) para denotar la variable aleatoria, y letras minúsculas (por ejemplo, x_1, x_2, ...) para los valores que toma la variable. Las variables estadísticas pueden ser de varios tipos:
  - *Discreta*: Toma una cantidad numerable (finita o infinita) de valores, por ejemplo, el número de hojas de un árbol.
  - *Continua*: Puede tomar infinitos valores entre dos dados, por ejemplo, el tiempo.
  - *Unidimensional*: Sólo se mide un caracter o dato de los elementos de la muestra, por ejemplo, la temperatura.
  - *Bidimensional (tridimensional, ...)*: Se miden simultaneamente varios caracteres de cada elemento, por ejemplo, altura y peso de las personas.

2.2 Función de densidad de probabilidad y función de distribución acumulativa

2.2.1 Variables discretas

- *Función de densidad de probabilidad:* Nos informa sobre cuán probable es cada valor de X y se denota por $f(x) = p(X = x)$
  Además sabemos que, por las propiedades de la probabilidad, $f(x_i) \geq 0$ y $\sum_{\forall i} f(x_i) = 1$.

La distribución discreta de probabilidad de una variable de recorrido $\{x_1, x_2, \ldots, x_k\}$ se puede escribir como una tabla:

Valor de la variable x_i	*Función de densidad de probabilidad* $f(x_i)$	*Función de distribución acumulada* $F(x_i)$
x_1	$f(x_1) = p(X = x_1)$	$F(x_1) = f(x_1)$
x_2	$f(x_2) = p(X = x_2)$	$F(x_2) = F(x_1) + f(x_2)$
$\vdots$	$\vdots$	$\vdots$
x_k	$f(x_k) = p(X = x_k)$	$F(x_k) = F(x_{k-1}) + f(x_k) = 1$

- *Función de distribución $F(x)$ o función de probabilidad o función de distribución acumulada:* Nos informa de la probabilidad de que la variable aleatoria X tome un valor menor o igual que x, esto es $F(x) = p(X \leq x)$.
  La función $F(x)$ es no decreciente y además cumple que $F(-\infty) = 0$ y $F(\infty) = 1$

  De este modo, la probabilidad de que una variable aleatoria X esté comprendida entre x_i y x_j es

$$p(x_i < X \leq x_j) = \sum_{k=i+1}^{j} f(x_k) = F(x_j) - F(x_i).$$

Ejemplo:

Consideremos el experimento de lanzar una moneda tres veces y tomemos como variable aleatoria $X = \#$caras obtenidas.

El espacio muestral del experimento será $\{CCC, CCR, CRC, RCC, CRR, RCR, RRC, RRR\}$ y el recorrido de la variable aleatoria X es $\Omega(X) = \{0, 1, 2, 3\}$

El suceso correspondiente a $X = 0$ es $\{RRR\}$, así que tenemos que, $p(X = 0) = \dfrac{1}{8} = \dfrac{1}{2} \cdot \dfrac{1}{2} \cdot \dfrac{1}{2} = (p(R))^3$

De igual modo, tenemos que $X = 1$ corresponde a $\{CRR, RCR, RRC\}$, así,

$$p(X = 1) = \frac{3}{8} = 3 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \binom{3}{1} (p(R))^2 \, p(C)$$

Para $X = 2$ tenemos el suceso $\{CCR, CRC, RCC\}$, así,

$$p(X = 2) = \frac{3}{8} = 3 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \binom{3}{1} (p(C))^2 \, p(R)$$

Y, por último, para $X = 3$ el suceso asociado es $\{CCC\}$, por lo tanto,

$$p(X = 3) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = (p(C))^3$$

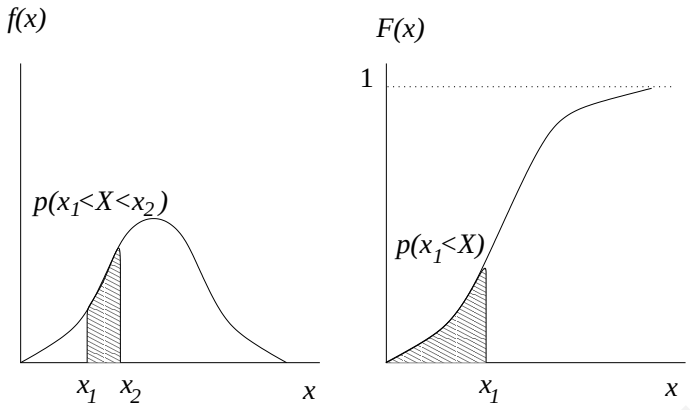


Figura 2.2: *Función de densidad de probabilidad, $f(x)$, y función de distribución, $F(x)$, para una variable continua.*

Podemos escribir la tabla de la función de densidad de probabilidad y de distribución de la forma

x_i	$f(x_i)$	$F(x_i)$
0	$1/8$	$1/8$
1	$3/8$	$1/2$
2	$3/8$	$7/8$
3	$1/8$	1

2.2.2 Variables continuas

- *Función de densidad de probabilidad:* Dado que una variable continua X puede tomar cualquier valor en un intervalo (a, b) o incluso $(-\infty, \infty)$. La probabilidad de que X tome un valor determinado es 0, no podemos definir la probabilidad de la misma forma que con las variables discretas (dando la probabilidad para cada valor de la variable), pero podemos especificar la probabilidad de que la variable esté en un intervalo dado.
  Para ello definimos la función de densidad de probabilidad $f(x)$, que cumple que $f(x) \geq 0$ y $\int_{-\infty}^{\infty} f(x)\,dx = 1$, de modo que $p(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\,dx$
- *Función de distribución o de probabilidad acumulativa $F(x)$:* Se define, como en el caso discreto, como la probabilidad de que la variable aleatoria X tome valores menores que uno dado, por lo tanto $F(x) = p(X < x) = \int_{-\infty}^{x} f(t)\,dt$

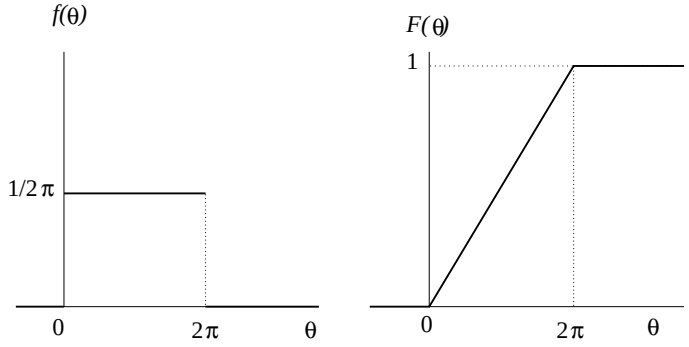


Figura 2.3: *Función de densidad de probabilidad, $f(\theta)$, y función de distribución, $F(\theta)$, para una variable aleatoria continua θ de densidad de probabilidad constante en un intervalo* $[0, 2\pi)$

Ejemplo:

Calcular la función de densidad de probabilidad y la función de distribución de que la aguja minutera de un reloj forme un ángulo θ con el eje de abcisas.

El espacio muestral en este caso es $[0, 2\pi)$, por lo tanto la función de densidad de probabilidad será

$$f(\theta) = \begin{cases} \dfrac{1}{2\pi} & \text{si } 0 \leq \theta < 2\pi \\ 0 & \text{en el resto} \end{cases}, \text{ por lo tanto, } F(\theta) = \begin{cases} 0 & \text{si } -\infty < \theta < 0 \\ \dfrac{\theta}{2\pi} & \text{si } 0 \leq \theta < 2\pi \\ 1 & \text{si } 2\pi \leq \theta < \infty \end{cases}$$

2.3 Distribuciones bivariantes. Distribución marginal. Distribución condicionadas. Variables aleatorias independientes.

Consideremos ahora una variable aleatoria bidimensional (X, Y) que tomará valores (x, y) de un espacio bidimensional real. Analizaremos a continuación algunas características de estas variables y de sus funciones de densidad de probabilidad y de distribución.

2.3.1 Función de densidad de probabilidad conjunta $f(x, y)$

- *Variables discretas:* Para variables discretas definimos la función de densidad de probabilidad $f(x_i, y_j)$ como la probabilidad de que la variable X tome el valor x_i y la variable Y tome el valor y_j, esto es, $f(x_i, y_j) = p(X = x_i, Y = y_j)$. Sabemos además que $\sum_{\forall i} \sum_{\forall j} f(x_i, y_j) = 1$.

- *Variables continuas:* Para variables continuas definimos la función de densidad de probabilidad como una función $f(x,y) \geq 0$ que nos permite calcular la probabilidad de las variables X e Y en los intervalos (x_1, x_2) e (y_1, y_2), respectivamente. Por lo tanto, $p(x_1 < X < x_2, y_1 < Y < y_2) = \int_{x_1}^{x_2} dx \int_{y_1}^{y_2} dy\, f(x,y)$, sabiendo que $\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x,y) = 1$

2.3.2 Distribuciones de densidad marginales

- *Variables discretas:* Definimos las funciones de densidad marginales, $f_X(x_i)$ y $f_Y(y_j)$, como:
$f_X(x_i) = P(X = x_i) = \sum_{\forall j} f(x_i, y_j)$ y $f_Y(y_j) = P(Y = y_j) = \sum_{\forall i} f(x_i, y_j)$

X \ Y	y_1	y_2	$\cdots$	y_m	*Marginales*
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\cdots$	$f(x_1, y_m)$	$f_X(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\cdots$	$f(x_2, y_m)$	$f_X(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_n	$f(x_n, y_1)$	$f(x_n, y_2)$	$\cdots$	$f(x_n, y_m)$	$f_X(x_n)$
Marginales	$f_Y(y_1)$	$f_Y(y_2)$	$\cdots$	$f_Y(y_m)$	1

- *Variables continuas:* Definimos las funciones de densidad marginales de forma similar al caso discreto, de modo que:
$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$ y $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$

2.3.3 Función de distribución conjunta

Es equivalente a la función de distribución acumulada del caso univariable, pero para variables aleatoria bivariables.

- *Variables discretas:* Se define como
$F(x,y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j)$

- *Variables continuas:* En este caso tenemos que
$F(x,y) = P(X < x, Y < y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} dv\, f(u,v)$

2.3.4 Función de distribución marginal

Son las funciones de distribución asociadas a las funciones de densidad marginal, por lo tanto

- *Variables discretas:* Se definen como
  $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} \sum_{\forall y_j} f(x_i, y_j)$ y $F_Y(y) = P(Y \leq y) = \sum_{y_j \leq y} \sum_{\forall x_i} f(x_i, y_j)$

- *Variables continuas:* En este caso tenemos que

  $$F_X(x) = P(X < x) = \int_{-\infty}^{x} du \int_{-\infty}^{\infty} dy\, f(u,y) \text{ y } F_Y(y) = P(Y < y) = \int_{-\infty}^{y} dv \int_{-\infty}^{\infty} dx\, f(x,v)$$

2.3.5 Distribuciones condicionadas

Al igual que cuando trabajamos con probabilidades condicionadas de un suceso, podemos definir funciones de probabilidad condicionada (para variables discretas) y funciones de densidad de probabilidad condicionada (para variables continuas) que nos den información sobre la probabilidad de que se verifique un valor para una de las componentes de la variable aleatoria bivariante, sabiendo el valor que toma la otra componente de la variable aleatoria. De este modo, definiremos:

- *Variables discretas:* La función de probabilidad condicionada nos permitirá calcular la probabilidad de que en la variable aleatoria (X, Y) se satifaga que $X = x$ sabiendo que $Y = y$, o viceversa. Así tendremos que
  $$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \Rightarrow f(x|y) = \frac{f(x,y)}{f_Y(y)}$$
  y
  $$p(Y = y|X = x) = \frac{p(X = x, Y = y)}{p(X = x)} \Rightarrow f(y|x) = \frac{f(x,y)}{f_X(x)}$$
  De este modo, para calcular la probabilidad $p(x_1 \leq X \leq x_2|Y = y)$ haremos
  $$p(x_1 \leq X \leq x_2|Y = y) = \sum_{x_1 \leq x_i \leq x_2} f(x_i|y),$$
  y de forma similar, la probabilidad $p(y_1 \leq Y \leq y_2|X = x)$ vendrá dada por
  $$p(y_1 \leq Y \leq y_2|X = x) = \sum_{y_1 \leq y_j \leq y_2} f(y_j|x)$$

- *Variables continuas:* La función de densidad condicionada nos permitirá calcular probabilidades de que una de las componentes de la variable aleatoria se encuentre en un intervalo, conocido el valor que toma la otra componente de la variable aleatoria. Dichas funciones de densidad de probabilidad condicionada serán
  $f(x|y) = \dfrac{f(x,y)}{f_Y(y)}$ y $f(y|x) = \dfrac{f(x,y)}{f_X(x)}$
  De modo que, $p(x_1 < X < x_2|Y = y) = \displaystyle\int_{x_1}^{x_2} f(x|y)\,dx$ y $p(y_1 < Y < y_2|X = x) = \displaystyle\int_{y_1}^{y_2} f(y|x)\,dy$

2.3.6 Variables independientes

Decimos que dada una variable aleatoria bivariante (X, Y), sus componentes X e Y son independientes cuando el conocimiento de una de las componentes no aporta información sobre la otra, esto es equivalente a decir que las distribuciones de densidad condicionadas son iguales a las marginales, o sea, $f(x|y) = f_X(x)$ y $f(y|x) = f_Y(y)$.

Esto se puede ver fácilmente comprobando que

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{\infty} f(x|y)f_Y(y)\,dy = f(x|y)\int_{-\infty}^{\infty} f_Y(y)\,dy = f(x|y)$$

donde se ha aplicado que $f(x|y)$ no depende de y, dado que hemos supuesto que la información sobre la probabilidad de X no depende del valor que tome Y para sacarlo fuera de la integral.

Teniendo todo esto en cuenta, podemos ver que en el caso de que las variables X e Y sean independientes, se cumplirá que

$$f(x,y) = f_X(x)f_Y(y)$$

Y del mismo modo, para función de distribución tendremos que en el caso de variables independientes

$$F(x,y) = F_X(x)F_Y(y)$$

Ejemplo:

Dadas las siguientes funciones de densidad de probabilidad para una variable aleatoria bivariante, analizar si son variables dependientes o independientes, construir las funciones de densidad marginales y las funciones de distribución asociadas. Calcular también $p(X = 0|Y = 1)$ y $p(Y = 2|X = 1)$

(a) Consideremos la función de densidad de probabilidad dada por la siguiente tabla

		Y			
	$f(x,y)$	0	1	2	$f_X(x)$
X	0	$1/6$	$1/12$	$1/12$	$f_X(0) = 1/3$
	1	$1/3$	$1/6$	$1/6$	$f_X(1) = 2/3$
	$f_Y(y)$	$f_Y(0) = 1/2$	$f_Y(1) = 1/4$	$f_Y(2) = 1/4$	1

Se puede ver que se cumple en todos los casos que $f(x,y) = f_X(x)f_Y(y)$, por lo tanto, las dos variables son independientes. Podemos calcular la función de distribución $F(x,y)$, que vendrá dada por la siguiente tabla:

		Y		
	$F(x,y)$	0	1	2
X	0	$1/6$	$1/4$	$1/3$
	1	$1/2$	$3/4$	1

Teniendo en cuenta estos datos podemos calcular $p(X = 0|Y = 1) = \dfrac{f(0,1)}{f_Y(1)} = \dfrac{1/12}{1/4} = \dfrac{1}{3}$ y $p(Y = 2|X = 1) = \dfrac{f(1,2)}{f_X(1)} = \dfrac{1/6}{2/3} = \dfrac{1}{4}$

(b) Consideremos ahora la función de densidad de probabilidad dada por la siguiente tabla

		Y			
	$f(x,y)$	0	1	2	$f_X(x)$
X	0	$1/5$	$2/15$	$2/5$	$f_X(0) = 11/15$
	1	0	$1/5$	$1/15$	$f_X(1) = 4/15$
	$f_Y(y)$	$f_Y(0) = 1/5$	$f_Y(1) = 1/3$	$f_Y(2) = 7/15$	1

Se puede ver que en este caso no se cumple que $f(x,y) = f_X(x)f_Y(y)$, por lo tanto, las dos variables son dependientes.

Calculamos ahora la función de distribución $F(x,y)$, que vendrá dada por la siguiente tabla:

		Y		
	$F(x,y)$	0	1	2
X	0	$1/5$	$1/3$	$11/15$
	1	$1/5$	$8/15$	1

Teniendo en cuenta estos datos podemos calcular $p(X=0|Y=1)=\dfrac{f(0,1)}{f_Y(1)}=\dfrac{2/15}{1/3}=\dfrac{2}{5}$ y $p(Y=2|X=1)=\dfrac{f(1,2)}{f_X(1)}=\dfrac{1/15}{4/15}=\dfrac{1}{4}$

Ejemplo:

Dadas las siguientes funciones $f(x,y)$, comprobar que definen funciones de densidad de probabilidad y calcular las funciones de densidad marginales y las funciones de distribución correspondientes. Analizar si las variables aleatorias X e Y son dependientes o independientes.

(a) Sea $f(x,y)=\begin{cases}\dfrac{3}{8}x(1-y^2) & \text{, si } 0<x<2 \text{ y } -1<y<1\\ 0 & \text{, resto}\end{cases}$

Podemos comprobar que $f(x,y)\geq 0$ para cualquier valor de x e y.

Además $\displaystyle\int_{-\infty}^{\infty}dy\int_{-\infty}^{\infty}dx\,f(x,y)=\frac{3}{8}\int_{-1}^{1}dy(1-y^2)\int_0^2 x\,dx=\frac{3}{8}\left[y-\frac{y^3}{3}\right]_{-1}^{1}\left[\frac{x}{2}\right]_0^2=$

$\dfrac{3}{8}\cdot 2\cdot\dfrac{2}{3}\cdot 2=1$

Las funciones de densidad de probabilidad marginales serán:

$$f_X(x)=\int_{-\infty}^{\infty}dy\,f(x,y)=\frac{3}{8}x\int_{-1}^{1}dy(1-y^2)=\frac{3}{8}x\left[y-\frac{y^3}{3}\right]_{-1}^{1}=\frac{x}{2}\Rightarrow$$

$$f_X(x)=\begin{cases}\dfrac{x}{2} & \text{, si } 0<x<2\\ 0 & \text{, resto}\end{cases}$$

$$f_Y(y)=\int_{-\infty}^{\infty}dx\,f(x,y)=\frac{3}{8}(1-y^2)\int_0^2 x\,dx=\frac{3}{4}(1-y^2)\Rightarrow$$

$$f_Y(y)=\begin{cases}\dfrac{3}{4}(1-y^2) & \text{, si } -1<y<1\\ 0 & \text{, resto}\end{cases}$$

Como se cumple que $f(x,y)=f_X(x)\cdot f_Y(y)$, las componentes X e Y de la variable aleatoria serán independientes.

Calculamos ahora la función de distribución:

$$F(x,y)=\int_{-\infty}^{y}dv\int_{-\infty}^{x}du\,f(u,v)=\frac{3}{8}\int_{-1}^{y}(1-v^2)dv\int_0^x u\,du=\frac{x^2}{16}(3y-y^3+2)\Rightarrow$$

$$F(x,y)=\begin{cases}0 & \text{, si } x<0 \text{ ó } y<-1\\ \dfrac{x^2}{16}(3y-y^3+2) & \text{, si } 0<x<2,-1<y<1\\ \dfrac{1}{16}(3y-y^3+2) & \text{, si } x>2,-1<y<1\\ \dfrac{x^2}{4} & \text{, si } y>1,0<x<2\\ 1 & \text{, si } x>2,y>1\end{cases}$$

(b) Sea $f(x,y)=\begin{cases}x+y & \text{, si } 0<x<1 \text{ y } 0<y<1\\ 0 & \text{, resto}\end{cases}$

Podemos comprobar que $f(x,y)\geq 0$ para cualquier valor de x e y.

Además $\displaystyle\int_{-\infty}^{\infty}dy\int_{-\infty}^{\infty}dx\,f(x,y)=\int_0^1 dy\int_0^1(x+y)\,dx=\int_0^1 dy\left[\frac{x^2}{2}+xy\right]_0^1=$

$\displaystyle\int_0^1 dy\left(y+\frac{1}{2}\right)=\left[\frac{y^2+y}{2}\right]_0^1=1$

Las funciones de densidad de probabilidad marginales serán:

$$f_X(x)=\int_{-\infty}^{\infty}dy\,f(x,y)=\int_0^1 dy(x+y)=\left[yx+\frac{y^2}{2}\right]_0^1=x+\frac{1}{2}\Rightarrow$$

$$f_X(x)=\begin{cases}x+\dfrac{1}{2} & \text{, si } 0<x<1\\ 0 & \text{, resto}\end{cases}$$

$$f_Y(y)=\int_{-\infty}^{\infty}dx\,f(x,y)=\int_0^1 dx(x+y)=\left[yx+\frac{x^2}{2}\right]_0^1=y+\frac{1}{2}\Rightarrow$$

$$f_Y(y)=\begin{cases}y+\dfrac{1}{2} & \text{, si } 0<y<1\\ 0 & \text{, resto}\end{cases}$$

Como se cumple que $f(x,y)\neq f_X(x)\cdot f_Y(y)$, las componentes X e Y de la variable aleatoria serán dependientes.

Calculamos ahora la función de distribución:

$$F(x,y)=\int_{-\infty}^{y}dv\int_{-\infty}^{x}du\,f(u,v)=\int_0^y dv\int_0^x(u+v)\,du=\int_0^y\left(xv+\frac{x^2}{2}\right)dv=\frac{x^2y+y^2x}{2}\Rightarrow$$

$$F(x,y)=\begin{cases}0 & \text{, si } x<0 \text{ ó } y<0\\ \dfrac{x^2y+y^2x}{2} & \text{, si } 0<x<1, 0<y<1\\ \dfrac{y+y^2}{2} & \text{, si } x>1, 0<y<1\\ \dfrac{x^2+x}{2} & \text{, si } y>1, 0<x<1\\ 1 & \text{, si } x>1, y>1\end{cases}$$

Capítulo 3

Función de variable aleatoria
Esperanza. Varianza. Covarianza. Coeficientes de correlación y determinación. Momentos.

3.1 Variables aleatorias unidimensionales

3.1.1 Esperanza matemática, valor esperado o media

La *esperanza matemática o media*, $E(X)$ ó μ, de una variable aleatoria X, es un valor teórico que nos da una idea de dónde se centra la distribución. La forma de calcularla es la siguiente:

- *Variable aleatoria discreta:*
  $\mu = E(X) = \sum_{\forall i} x_i p(X = x_i) = \sum_{\forall i} x_i f(x_i)$

- *Variable aleatoria continua:*
  $\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

Es fácil comprobar que la esperanza matemática es una función lineal, por lo tanto, $E(aX+b) = aE(X) + b$, donde a y b son constantes.

3.1.2 Varianza y desviación típica

La media no da una adecuada descripción del comportamiento de la variable aleatoria X, por eso también es importante determinar la dispersión o variación de los valores de la variable aleatoria. Por ejemplo, las variables aleatorias X e Y descritas a continuación:

X, Y	0	1	2	3	4
$p(X)$	0	$1/4$	$1/2$	$1/4$	0
$p(Y)$	$1/8$	$1/4$	$1/4$	$1/4$	$1/8$

tienen ambas el mismo valor para la esperanza matemática, $E(X) = E(Y) = 2$, sin embargo, la variable X tiene menos dispersión que la variable Y.

Por este motivo resulta útil definir la *varianza,* $Var(X) = \sigma^2 = E\left((X - E(X))^2\right)$, que nos permite analizar la dispersión de la variable con respecto a la media.

La forma de calcular dicha varianza es

- *Variables discretas:*
  $Var(X) = E\left((X - E(X))^2\right) = \sum_{\forall i}(x_i - E(X))^2 f(x_i)$
- *Variables continuas:*
  $Var(X) = E\left((X - E(X))^2\right) = \int_{-\infty}^{\infty}(x - E(X))^2 f(x)\,dx$

Debido a que el operador esperanza matemática es un operador lineal, es fácil comprobar que: $Var(X) = E\left((X - E(X))^2\right) = E\left(X^2 - 2XE(X) + E(X)^2\right) = E\left(X^2\right) - \left(E(X)\right)^2$

Se define también, a partir de la varianza, la *desviación típica,* $\sigma = \sqrt{Var(X)}$ que da una mejor idea de la dispersión de la variable con respecto a la media, al tener las mismas unidades que la variable aleatoria.

3.1.3 Momentos

La media y la varianza son dos casos particulares de la definición de un concepto más general, como es el de *momento de orden r con respecto al parámetro c* de una variable aleatoria X.

Dada una variable aleatoria X definimos el momento de orden r con respecto al parámetro c como la esperanza matemática de $(X - c)^r$, que se calculará de la siguiente manera:

- *Variable discreta:*
  $E\left((X - c)^r\right) = \sum_{\forall i}(x_i - c)^r f(x_i)$
- *Variable continua:*
  $E\left((X - c)^r\right) = \int_{-\infty}^{\infty}(x - c)^r f(x)\,dx$

De especial interés resultan los *momentos respecto al origen,* $m_r = E(X^r)$, que se obtienen al tomar $c = 0$. Como se puede ver, los primeros definidos así representan parámetros de la variable ya conocidos anteriormente:

$m_0 = 1; \qquad m_1 = E(X) = \mu; \qquad m_2 = E(X^2) = \sigma^2 + \mu^2$

De igual interés resultan también los *momentos centrales o momentos respecto a la media,* $\mu_r = E\left((X-\mu)^r\right)$, obtenidos al tomar $c = \mu$. En este caso, los primeros momentos son:
$\mu_0 = 1; \qquad \mu_1 = E(X-\mu) = 0; \qquad \mu_2 = E\left((X-\mu)^2\right) = \sigma^2$

Función generadora de momentos: Dada una variable aleatoria X, se define para cualquier t real la función generadora de momentos como $M_X(t) = E\left(e^{tX}\right)$, de modo que tenemos:

- *Variable discreta:*
  $M_X(t) = \sum_{\forall i} e^{tx_i} f(x_i)$
- *Variable continua:*
  $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$

La importancia de dicha función generadora de momentos es que permite calcular todos los momentos respecto al origen, ya que $m_r = \left.\dfrac{d^r M_X(t)}{dt^r}\right|_{t=0}$

Parámetros poblacionales: Se definen los parámetros poblacionales como toda constante que puede definirse, en general, a partir de los momentos (respecto al origen) de la variable aleatoria X. Los más comunes son:

- *Esperanza matemática:*
  $\mu = E(X) = m_1$
- *Varianza:*
  $Var(X) = \sigma^2 = E(X^2) - (E(X))^2 = m_2 - m_1^2$
- *Coeficiente de variación:* permite hacerse una idea de la dispersión de la variable en unidades de la media (o sea, normalizado a un valor característico de la variable X)
  $cv = c_v = \dfrac{\sigma}{\mu} = \dfrac{\sqrt{m_2 - m_1^2}}{m_1}$
- *Coeficiente de asimetría:* es una medida de la asimetría de la distribución de la variable aleatoria X. Si este coeficiente es positivo, los valores de la variable X se extienden más hacia valores superiores a la media, y si es negativo, la variable X se extiende más en la zona de valores inferiores a la media
  $\gamma_1 = \dfrac{E\left((X-\mu)^3\right)}{\sigma^3} = \dfrac{m_3 - 3m_1m_2 + 2m_1^3}{\sqrt{\left(m_2 - m_1^2\right)^3}}$

- *Curtosis:* como veremos más adelante, la curtosis nos da una medida del apuntamiento de la distribución de densidad. Está normalizada a la distribución normal, que es una de las más comunes en estadística. Si la curtosis es positiva, la distribución es más apuntada que la distribución normal, y si es negativa, es menos apuntada que la normal.
  $\gamma_2 = \dfrac{E\left((X-\mu)^4\right)}{\sigma^4} - 3$

Ejemplos

Calcular la media y la varianza de las siguientes variables aleatorias:

(a) Considerar la variable aleatoria X cuya función de densidad de probabilidad viene descrita por:

x	0	1	2	3
$f(x)$	0.5	0.1	0.2	0.2

La media de dicha variable aleatoria es:

$$\mu = E(X) = \sum_{i=0}^{3} x_i f(x_i) = 0 \cdot 0.5 + 1 \cdot 0.1 + 2 \cdot 0.2 + 3 \cdot 0.2 = 1.1$$

Como se puede ver, el valor de la esperanza matemática puede no coincidir con uno de los elementos del espacio muestral.

La varianza de X es:

$$Var(X) = E\left((X-\mu)^2\right) = E\left(X^2\right) - \mu^2 = 0^2 \cdot 0.5 + 1^2 \cdot 0.1 + 2^2 \cdot 0.2 + 3^2 \cdot 0.2 - 1.1^2 = 1.49$$

(b) Considerar, ahora, la variable aleatoria continua X cuya función de densidad de probabilidad es:

$$f(x) = \begin{cases} \dfrac{1}{2}x & , \ 0 < x \leq 2 \\ 0 & , \ \text{resto} \end{cases}$$

Se puede comprobar que $f(x)$ es una función de densidad de probabilidad, ya que cumple que $f(x) \leq 0$ para todo valor de la variable aleatoria y además

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^2 \frac{1}{2}x\,dx = \left[\frac{x^2}{4}\right]_0^2 = 1$$

La media de dicha variable aleatoria es:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^2 \frac{1}{2}x^2\,dx = \left[\frac{x^3}{6}\right]_0^2 = \frac{4}{3}$$

La varianza de X es:

$Var(X) = E\left((X-\mu)^2\right) = E\left(X^2\right) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \frac{16}{9} = \int_0^2 \frac{1}{2}x^3\,dx - \frac{16}{9} = \left[\frac{x^4}{8}\right]_0^2 - \frac{16}{9} = 2 - \frac{16}{9} = \frac{2}{9}$

3.2 Variable aleatoria bidimensional

3.2.1 Media o esperanza matemática

Cuando trabajamos con una variable aleatoria bidimensional (X,Y) cuya función de densidad conjunta es $f(x,y)$, podemos definir la *media o esperanza matemática* para cada una de sus variables, de la siguiente forma:

- *Variable discreta:*
  $E(X) = \mu_X = \sum_{\forall i}\sum_{\forall j} x_i f(x_i, y_j); \qquad E(Y) = \mu_Y = \sum_{\forall i}\sum_{\forall j} y_j f(x_i, y_j)$
- *Variable continua:*
  $E(X) = \mu_X = \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx\, x\, f(x,y); \;\; E(Y) = \mu_Y = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, y\, f(x,y)$

Como ya hemos visto, la esperanza matemática es un operador lineal, por lo que $E(aX+bY) = aE(X) + bE(Y)$, siendo a y b constantes.

3.2.2 Varianza. Covarianza. Coeficiente de correlación y coeficiente de determinación.

Varianza: La *varianza* de una variable bidimensional (X,Y) se puede definir para cada una de sus componentes de la siguiente forma:

- *Variable discreta:*
  $Var(X) = \sigma_X^2 = E\left((X-\mu_X)^2\right) = \sum_{\forall i}\sum_{\forall j}(x_i-\mu_X)^2 f(x_i,y_j)$
  $Var(Y) = \sigma_Y^2 = E\left((Y-\mu_Y)^2\right) = \sum_{\forall i}\sum_{\forall j}(y_j-\mu_Y)^2 f(x_i,y_j)$
- *Variable continua:*
  $Var(X) = \sigma_X^2 = E\left((X-\mu_X)^2\right) = \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx\,(x-\mu_X)^2 f(x,y)$

$$Var(Y) = \sigma_Y^2 = E\left((Y-\mu_Y)^2\right) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\,(y-\mu_Y)^2\, f(x,y)$$

Covarianza: Un parámetro importante en las variables bidimensionales (X,Y) es la *covarianza,* que permite analizar la relación entre las componentes X e Y. Se define como:
$Cov(X,Y) = \sigma_{XY}^2 = E\left((X-\mu_X)(Y-\mu_Y)\right) = E(XY) - \mu_X\mu_Y$

Por lo tanto, se calculará de la siguiente forma:

- *Variable discreta:*
  $Cov(X,Y) = \sigma_{XY}^2 = E\left((X-\mu_X)(Y-\mu_Y)\right) = \sum_{\forall i}\sum_{\forall j}(x_i-\mu_X)(y_j-\mu_Y)f(x_i,y_j)$
- *Variable continua:*
  $Cov(X,Y) = \sigma_{XY}^2 = E\left((X-\mu_X)(Y-\mu_Y)\right) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\,(x-\mu_X)(y-\mu_Y)\, f(x,y)$

Con estas definiciones, es fácil demostrar que:
$Var(aX+bY) = \sigma_{aX+bY}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}^2$
siendo a y b constantes. Con esto se demuestra que el operador varianza no es un operador lineal.

Vamos a comprobar ahora que la covarianza nos da información sobre la dependencia o independencia de las variables X e Y. Para ello, supongamos que ambas variables son independientes, en ese caso tendremos que $f(x,y) = f_X(x)f_Y(y)$. Calcularemos ahora la covarianza de las variables X e Y para demostrar que en este caso $Cov(X,Y) = 0$

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y = \sum_{\forall i}\sum_{\forall j} x_i y_j f(x_i,y_j) - \mu_X\mu_Y = \sum_{\forall i}\sum_{\forall j} x_i y_j f_X(x_i) f_Y(y_j) - \mu_X\mu_Y$$

$$= \left(\sum_{\forall i} x_i f_X(x_i)\right)\left(\sum_{\forall j} y_j f_Y(y_j)\right) - \mu_X\mu_Y = 0$$

Coeficiente de correlación: Definimos el *coeficiente de correlación* de una variable bidimensional (X,Y) como $\rho = \rho_{XY} = \dfrac{Cov(X,Y)}{\sigma_X\sigma_Y}$

Este parámetro cumple que $-1 \leq \rho_{XY} \leq 1$ y da una medida de la dependencia lineal o correlación de las variables X e Y. Cuando $\rho_{XY} = 0$ ambas variables son independientes, mientras que si $\rho_{XY} = 1$ ó $\rho_{XY} = -1$ las variables siguen una relación lineal de la forma $Y = aX + b$, con $a > 0$ ó $a < 0$, respectivamente.

Coeficiente de determinación: Definimos el coeficiente de determinación de una variable aleatoria bidimensional (X,Y) como $\rho^2 = \rho_{XY}^2$. Dicho coeficiente, que cumple que $0 \leq \rho^2 \leq 1$, nos permite conocer cómo de conocida (o determinada) es una componente de la variable cuando

se conoce la otra. Si $\rho^2 = 0$, el conocimiento de una de las componentes de la variable no nos da información sobre el valor que puede tomar la otra componente; mientras que si $\rho^2 = 1$, dado un valor de una de las componentes de la variable, la otra queda perfectamente determinada.

Ejemplos

Calcular las medias, varianzas y covarianzas de la variable aleatoria bidimensional definida por las siguientes tablas de densidad de probabilidad

(a) Consideremos la siguiente función de densidad de probabilidad:

		Y			
	$f(x,y)$	0	1	2	$f_X(x)$
X	0	$1/6$	$1/12$	$1/12$	$f_X(0) = 1/3$
	1	$1/3$	$1/6$	$1/6$	$f_X(1) = 2/3$
	$f_Y(y)$	$f_Y(0) = 1/2$	$f_Y(1) = 1/4$	$f_Y(2) = 1/4$	1

Las esperanzas matemáticas de las componentes de la variable aleatoria (X, Y) son:

$$\mu_X = E(X) = \sum_{i=1}^{2} x_i f_X(x_i) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}$$

$$\mu_Y = E(Y) = \sum_{i=1}^{3} y_i f_Y(y_i) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} = \frac{3}{4}$$

Las varianzas de las componentes son:

$$Var(X) = E(X^2) - \mu_X^2 = 0^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{2}{3} - \frac{4}{9} = \frac{2}{9}$$

$$Var(Y) = E(Y^2) - \mu_Y^2 = 0^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{4} - \frac{9}{16} = \frac{11}{16}$$

Y, finalmente, la covarianza es:

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y = \sum_{i=1}^{2} \sum_{j=1}^{3} x_i y_j f(x_i, y_j) - \mu_X \mu_Y =$$

$$= 0 \cdot \left(0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12}\right) + 1 \cdot \left(0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6}\right) - \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{2} - \frac{1}{2} = 0$$

La covarianza sale 0 porque las variables X e Y son independientes.

(b) Consideremos ahora la función de densidad de probabilidad siguiente:

		Y			
	$f(x,y)$	0	1	2	$f_X(x)$
X	0	$1/5$	$2/15$	$2/5$	$f_X(0) = 11/15$
	1	0	$1/5$	$1/15$	$f_X(1) = 4/15$
	$f_Y(y)$	$f_Y(0) = 1/5$	$f_Y(1) = 1/3$	$f_Y(2) = 7/15$	1

Las esperanzas matemáticas de las componentes de la variable aleatoria (X, Y) son:

$$\mu_X = E(X) = \sum_{i=1}^{2} x_i f_X(x_i) = 0 \cdot \frac{11}{15} + 1 \cdot \frac{4}{15} = \frac{4}{15}$$

$$\mu_Y = E(Y) = \sum_{i=1}^{3} y_i f_Y(y_i) = 0 \cdot \frac{1}{5} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{7}{15} = \frac{19}{15}$$

Las varianzas de las componentes son:

$$Var(X) = E(X^2) - \mu_X^2 = 0^2 \cdot \frac{11}{15} + 1^2 \cdot \frac{4}{15} - \frac{16}{225} = \frac{44}{225}$$

$$Var(Y) = E(Y^2) - \mu_Y^2 = 0^2 \cdot \frac{1}{5} + 1^2 \cdot \frac{1}{3} + 2^2 \cdot \frac{7}{15} - \frac{361}{225} = \frac{134}{225}$$

Y, finalmente, la covarianza es:

$$Cov(X, Y) = E(XY) - \mu_X \mu_Y = \sum_{i=1}^{2} \sum_{j=1}^{3} x_i y_j f(x_i, y_j) - \mu_X \mu_Y =$$

$$= 0 \cdot \left(0 \cdot \frac{1}{5} + 1 \cdot \frac{2}{15} + 2 \cdot \frac{2}{5}\right) + 1 \cdot \left(0 \cdot 0 + 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{15}\right) - \frac{4}{15} \cdot \frac{19}{15} = -\frac{1}{225}$$

La covarianza es distinta de 0 porque las variables X e Y son dependientes. Podemos ver que el coeficiente de correlación es $\rho_{XY} = \dfrac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}} = -\dfrac{1/225}{\sqrt{44 \cdot 134}/225} = -0.013$

Capítulo 4

Distribuciones discretas

Distribución binomial. Distribución de Poisson o ley de sucesos raros. Aproximación de la distribución binomial a la Poisson.

4.1 Introducción

El comportamiento de una variable aleatoria X queda descrito, en general, por su distribución de probabilidad o función de densidad de probabilidad. Vamos a centrarnos, en este capítulo, en las variables aleatorias discretas, analizando las distribuciones de probabilidad , $f(x) = P(X = x)$, más características.

4.2 Distribución discreta uniforme

Se caracteriza porque la probabilidad de la variable aleatoria es constante, o sea, $f(x_i) = p(x_i) = cte$. En temas anteriores ya hemos visto ejemplos de esta distribución de probabilidad, por ejemplo, tirar un dado equilibrado o lanzar una moneda.

Dado el recorrido de la variable aleatoria X, esto es $X(\Omega) = \{x_1, x_2, \ldots, x_n\}$, todos los valores x_i de la variable aleatoria tienen la misma probabilidad, por lo tanto, la función de distribución de probabilidad será:

$$f(x_i) = \begin{cases} \dfrac{1}{n} & \text{, si } i = 1, 2, ..., n \\ 0 & \text{, resto} \end{cases}$$

Por lo tanto, tendremos que:

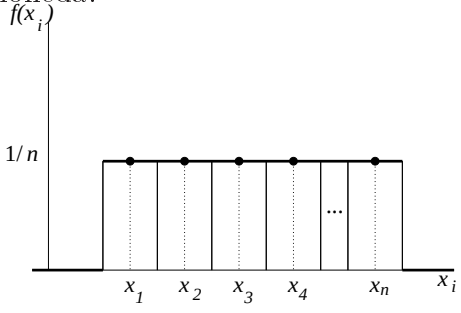


Figura 4.1*: Función de distribución de probabilidad de una variable discreta uniforme*

- La *media* o *esperanza matemática* es:
  $\mu = \sum_{i=1}^{n} x_i f(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$

- La *varianza* es:
  $Var(X) = \sum_{i=1}^{n}(x_i - \mu)^2 f(x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)$

4.3 Distribución de Bernoulli

Vamos a considerar ahora una variable aleatoria asociada a un suceso A que puede tener lugar (o no) con una probabilidad p.

Consideraremos que $X = 1$ si tiene lugar el suceso A y $X = 0$ si tiene lugar su complementario.

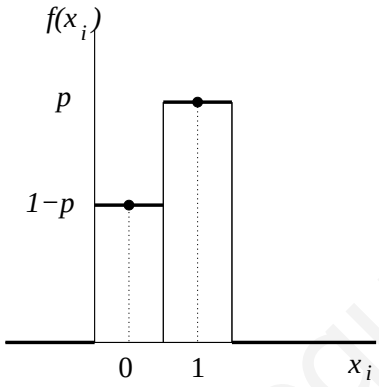


Figura 4.2*: Función de distribución de probabilidad de Bernoulli*

Por lo tanto, la función de distribución de probabilidad será

$$f(x) = \begin{cases} p^x(1-p)^{1-x} & \text{, si } x = 0, 1 \\ 0 & \text{, resto} \end{cases}$$

Por lo tanto, tendremos que:

- La *media* o *esperanza matemática* es:
  $\mu = E(X) = p$

- La *varianza* es:
  $Var(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$

4.4 Distribución binomial

Consideremos ahora un experimento en el cual la realización de un suceso A supone un éxito, en ese caso, la variable aleatoria es $X = 1$, y en caso de que se verifique el suceso complementario A^c consideramos que supone un fracaso y en ese caso $X = 0$ (distribución de Bernoulli). Si el suceso A tiene una probabilidad p y repetimos el experimento n veces, tendremos una distribución binomial cuya función de distribución de probabilidad será:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{, si } x = 0, 1, ..., n \\ 0 & \text{, resto} \end{cases}$$

El factor $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ nos indica el número de veces que podemos tener x éxitos entre los n intentos. Viene dado por el binomio de Newton, que representa la combinación de n elementos tomadas de x en x.

En una distribución binomial se cumplirá que:

- La *media* o *esperanza matemática* es: $\mu = E(X) = np$
- La *varianza* es: $Var(X) = E(X^2) - (E(X))^2 = np(1-p)$

Como se puede ver fácilmente, un modelo binomial es equivalente a un modelo de Bernoulli repetido n veces.

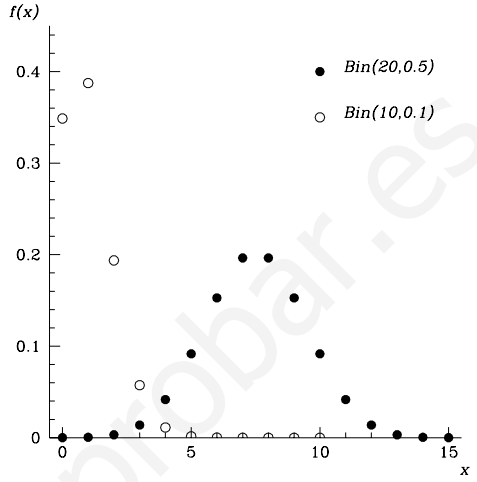


Figura 4.3: *Ejemplos de funciones de distribución de probabilidad binomiales $Bin(n,p)$. Se han considerado dos modelos: $Bin(20, 0.5)$ y $Bin(10, 0.1)$*

Ejemplo:

Un jugador de baloncesto tira 3 tiros libres, se sabe que su probabilidad de encestar es 80%, calcular la función de probabilidad y la función de distribución.

Consideremos el suceso S, que se satisface cuando el jugador encesta, y su complementario $S^c = N$, que se verifica cuando el jugador no encesta. La variable aleatoria será X="número de veces que encesta". Como se hacen tres tiros libres, el recorrido de dicha variable será $X(\Omega) = \{0, 1, 2, 3\}$. Además tenemos que la probabilidad del suceso S es $p = 0.8$, por lo tanto, estamos ante una distribución binomial $Bin(3, 0.8)$. Vamos a calcular ahora la probabilidad para cada uno de los valores de la variable aleatoria:

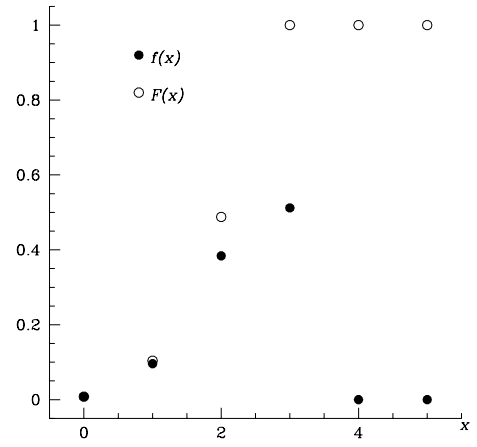


Figura 4.4: *Distribución de probabilidad, $f(x)$, y función de distribución, $F(x)$, de una $Bin(3, 0.8)$*

Si $X = 0$, quiere decir que el resultado del experimento es $\{NNN\}$, por lo tanto, la probabilidad de dicho valor de la variable aleatoria será:

$p(X = 0) = (1-p)^3 = \binom{3}{0} p^0(1-p)^{3-0} = 0.2^3 = 0.008$

Si $X = 1$, quiere decir que el resultado del experimento es $\{SNN,\ NSN, NNS\}$, por lo tanto, la probabilidad de dicho valor de la variable aleatoria será:

$p(X = 1) = 3p(1-p)^2 = \binom{3}{1} p^1(1-p)^{3-1} = 3 \cdot 0.8 \cdot 0.2^2 = 0.096$

Si $X = 2$, quiere decir que el resultado del experimento es $\{SSN,\ SNS, NSS\}$, por lo tanto, la probabilidad de dicho valor de la variable aleatoria será:

$p(X = 2) = 3p^2(1-p) = \binom{3}{2} p^2\,(1-p)^{3-2} = 3 \cdot 0.8^2 \cdot 0.2 = 0.384$

Si $X = 3$, quiere decir que el resultado del experimento es $\{SSS\}$, por lo tanto, la probabilidad de dicho valor de la variable aleatoria será:

$p(X = 3) = p^3 = \binom{3}{3} p^3(1-p)^{3-3} = 0.8^3 = 0.512$

Así pues podemos escribir una tabla con la función de probabilidad $f(x) = p(x)$ y la función de distribución $F(x) = p(X \leq x)$

x	0	1	2	3
$f(x)$	0.008	0.096	0.384	0.512
$F(x)$	0.008	0.104	0.488	1

La media y la varianza de esta distribución son, respectivamente, $\mu = np = 3 \cdot 0.8 = 2.4$ aciertos y $\sigma^2 = np(1-p) = 3 \cdot 0.8 \cdot 0.2 = 0.48 \Rightarrow \sigma = 0.693$

4.5 Distribución de Poisson

Consideremos ahora una variable aleatoria X que mide el número de sucesos en un intervalo continuo. El intervalo puede ser, por ejemplo, de tiempo, midiendo el número de partículas desintegradas en una unidad de tiempo; de espacio, el número de poros por milímetro cuadrado de piel; etc. Diremos que tenemos una distribución de Poisson (o ley de sucesos raros) si se cumple que:

- El número de resultados en un intervalo es independiente de lo que ocurre en otro intervalo (el proceso no tiene memoria).

- La probabilidad de que un suceso ocurra es proporcional al tamaño del intervalo y además es constante (proceso estable), por lo que se puede definir un número medio de resultados por unidad de intervalo.

- La posibilidad de que ocurra más de un resultado en un intervalo suficientemente pequeño es despreciable.

Bajo estas condiciones, la variable aleatoria X tendrá una variabilidad $X = \{0, 1, 2, ...\}$ y sigue una distribución de Poisson.

Podemos considerar esta distribución como una binomial en la que el número de ensayos se hace muy grande, $n \to \infty$, y la probabilidad de éxito muy pequeña, $p \to 0$, de modo que la media se mantiene constante, $\lambda = np = E(X) = cte$. Así podemos escribir la función de probabilidad como:

$$f(x) = p(X = x) = \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} =$$

$$\lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} =$$

$$\lim_{n\to\infty} \frac{n(n-1)\ldots 2\cdot 1}{n^x(n-x)(n-x-1)\ldots 2\cdot 1} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^{n} = \frac{\lambda^x}{x!} e^{-\lambda}$$

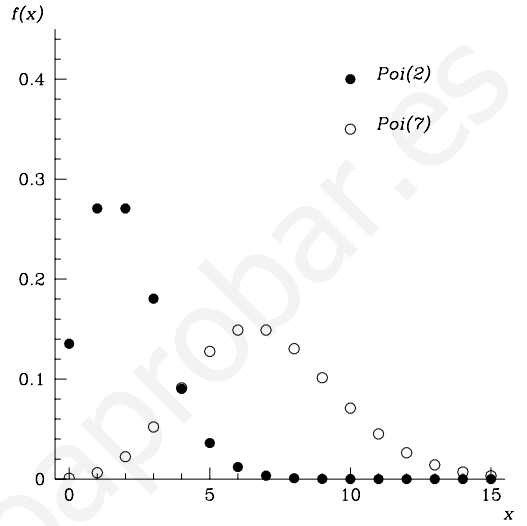


Figura 4.5: *Ejemplos de distribución de Poisson, $f(x) \to Poi(\lambda)$. Se han considerado dos casos: $\lambda = 2$ y $\lambda = 7$*

Por lo tanto, la función de distribución será:

$$f(x) = p(x;\lambda) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda} & \text{, si } x = 0, 1, 2, ... \\ 0 & \text{, resto} \end{cases}$$

En una distribución de Poisson se cumplirá que:

- La *media* o *esperanza matemática* es:
  $\mu = E(X) = \lambda$

- La *varianza* es:
  $Var(X) = E(X^2) - (E(X))^2 = \lambda$

Estos valores se pueden obtener fácilmente a partir de la función generadora de momentos,

$$M_X(t) = E(e^{tx}) = \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}}{x!}\lambda^x = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} = e^{\lambda(e^t-1)}$$

También se puede ver que si calculamos el coeficiente de asimetría, γ_1, obtenemos $\gamma_1 = \frac{1}{\sqrt{\lambda}}$. Con ello se puede ver que si λ es grande, la distribución tiende a ser simétrica.

Ejemplo:

Un detector astronómico observa una media de 3 fotones por segundo. Calcular la probabilidad de que lleguen $X = 0, 1, 2, 3, 4, ...$ fotones por segundo.

En este problema tenemos una distribución de Poisson de parámetro $\lambda = 3$, por lo tanto $X \sim Poi(3)$ y tendremos que $f(x) = p(x;3) = \frac{3^x e^{-3}}{x!}$. Con esta expresión podremos construir una tabla de las probabilidades y la función de distribución para varios valores de la variable aleatoria X

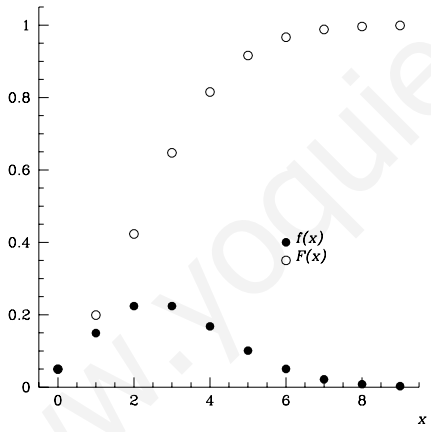


Figura 4.6: *Distribución de probabilidad, $f(x)$, y función de distribución, $F(x)$, para un modelo de Poisson de parámetro $\lambda = 3$*

x	$f(x)$	$F(x)$
0	0.050	0.050
1	0.149	0.199
2	0.224	0.423
3	0.224	0.647
4	0.168	0.815
5	0.101	0.916
6	0.050	0.966
7	0.022	0.988
8	0.008	0.996
9	0.003	0.999

En este ejemplo, la media y la varianza son, respectivamente, $\mu = \lambda = 3$ fotones por segundo y $\sigma^2 = \lambda = 3$, por lo que $\sigma = \sqrt{\lambda} = \sqrt{3} = 1.73$ fotones por segundo.

4.6 Aproximación de la distribución binomial a la de Poisson

Hemos visto que la distribución de Poisson se puede obtener a partir de una binomial en el límite $n \to \infty$ y $p \to 0$ con $np = cte$. Así, en determinados casos podremos aproximar la binomial por una distribución de Poisson. La aproximación será buena si se cumple que $n > 50$ y $p < 0.1$, y será tanto mejor cuanto mayor sea n y menor p.

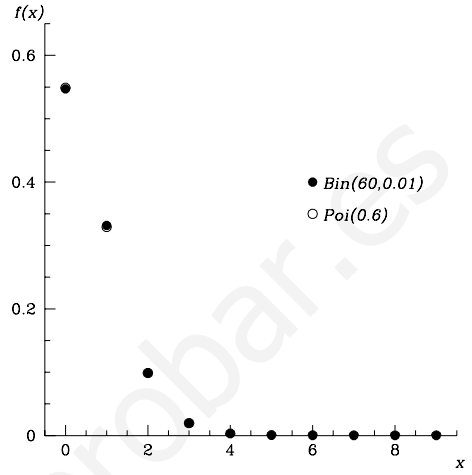


Figura 4.7: *Comparación entre la función de distribución de probabilidad $f(x)$ de un distribución binomial $Bin(60, 0.01)$ y una de Poisson $Poi(0.6)$*

Ejemplo:

Considerar una distribución binomial en la cual la probabilidad de éxito es $p = 0.01$ y el número de veces que se repite el experimento es $n = 60$. Calcular de forma exacta y utilizando la aproximación a una distribución de Poisson la probabilidad de tener tres éxitos. Calcular también la probabilidad de tener más de tres éxitos.

La distribución que tenemos es una binomial $Bin(60, 0.01)$ que dado que $n = 60 > 50$ y $p = 0.01 < 0.1$, podríamos aproximarla por una distribución de Poisson de parámetro $\lambda = np = 0.6$, esto es, $Poi(0.6)$

Vamos a calcular de forma exacta la probabilidad de tener 3 éxitos, o sea, $p(X = 3)$:

$$p(X = 3) = \binom{60}{3} 0.01^3 \cdot 0.99^{57} = 0.0193$$

Para calcular ahora la probabilidad de tener más de 3 éxitos, tendremos que calcular $p(X > 3) = 1 - p(X \leq 3)$, así que calcularemos también $p(0), p(1)$ y $p(2)$:

$$p(X = 0) = \binom{60}{0} 0.01^0 \cdot 0.99^{60} = 0.5472$$

$$p(X = 1) = \binom{60}{1} 0.01^1 \cdot 0.99^{59} = 0.3316$$

$$p(X = 2) = \binom{60}{2} 0.01^2 \cdot 0.99^{58} = 0.0988$$

Por lo tanto $p(X > 3) = 1 - 0.5472 - 0.3316 - 0.0988 - 0.0193 = 0.0031$

Repetiremos ahora los cálculos considerando la aproximación a una distribución de Poisson

con $\lambda = 0.6$

$$p(X = 3) = \frac{0.6^3 e^{-0.6}}{3!} = 0.0198$$

$$p(X = 2) = \frac{0.6^2 e^{-0.6}}{2!} = 0.0988$$

$$p(X = 1) = \frac{0.6^1 e^{-0.6}}{1!} = 0.3293$$

$$p(X = 0) = \frac{0.6^0 e^{-0.6}}{0!} = 0.5488$$

Por lo tanto, $p(X > 3) = 1 - 0.0198 - 0.0988 - 0.3293 - 0.5488 = 0.0034$

Capítulo 5

Distribuciones continuas I

Distribución uniforme. Distribución normal. Teorema del límite central. Aproximaciones de las distribuciones binomial y de Poisson a la normal. Corrección de continuidad.

5.1 Introducción

En este tema vamos a estudiar la función de densidad de probabilidad continua más utilizada, la distribución normal. En el siguiente tema analizaremos otras funciones de densidad continua que también resultan de interés estadístico.

5.2 Distribución continua uniforme

Antes de empezar a analizar la distribución normal, vamos a recordar algunos aspectos útiles en el estudio de las funciones de distribución. Para ello comenzaremos analizando la distribución continua uniforme.

Una variable aleatoria continua X sigue una distribución continua uniforme cuando su densidad de probabilidad toma un valor constante en un intervalo $[a, b]$. Según esto, $f(x) = k$ si $a < x < b$, como tiene que cumplirse que:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow \int_{a}^{b} kdx = k(b-a) = 1 \Rightarrow k = \frac{1}{b-a}$$

Por lo tanto, la función de densidad de probabilidad será:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{, si } a < x < b \\ 0 & \text{, resto} \end{cases}$$

A partir de aquí podemos calcular la función de distribución $F(x) = \int_{-\infty}^{x} f(t)dt$ y obtenemos:

$$F(x) = \begin{cases} 0 & \text{, si } -\infty < x \leq a \\ \dfrac{x-a}{b-a} & \text{, si } a < x < b \\ 1 & \text{, si } x \geq b \end{cases}$$

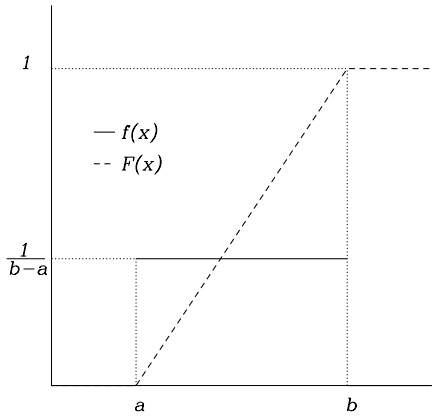


Figura 5.1*: Función de densidad de probabilidad, $f(x)$, y función de distribución, $F(x)$, para una distribución continua uniforme.*

Calculamos ahora la *media* de esta variable aleatoria:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)dx = \frac{1}{b-a}\int_{a}^{b} x\,dx = \frac{b^2-a^2}{2(b-a)} \Rightarrow E(X) = \frac{b+a}{2}$$

La *varianza* será:

$$Var(X) = \sigma^2 = E(X^2) - (E(X))^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left(\frac{b+a}{2}\right)^2 = \frac{1}{b-a}\int_{a}^{b} x^2\,dx - \frac{1}{4}\left(b^2+a^2+2ab\right) = \frac{1}{12}\left(b^2-2ab+a^2\right) \Rightarrow Var(X) = \frac{(b-a)^2}{12}$$

5.3 Distribución normal

5.3.1 Densidad de probabilidad y función de distribución de una distribución normal

La distribución de probabilidad más usada en estadística es la *distribución normal o gaussiana* (gran parte de los fenómenos físicos, naturales y sociales siguen este tipo de distribución).

Se define como:

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

5.3.2 Parámetros poblacionales de una distribución normal

Dicha función de distribución se caracteriza porque su *media* es $E(X) = \mu$ y su *varianza* $Var(X) = \sigma^2$. Cuando nos referimos a una distribución normal de media μ y de desviación típica σ la denominamos $N(\mu, \sigma)$.

Se puede comprobar fácilmente que si tenemos n variables aleatorias X_i, que siguen cada una de ellas una distribución $N(\mu_i, \sigma_i)$, y definimos una nueva variable aleatoria X tal que $X = \sum_{i=1}^{n} X_i$, dicha variable X seguirá una distribución $N(\mu, \sigma)$ con $\mu = \sum_{i=1}^{n} \mu_i$ y $\sigma = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$

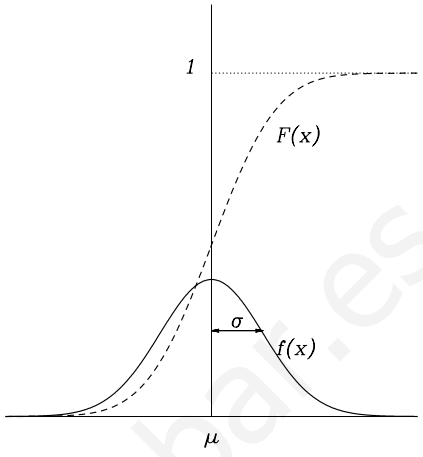


Figura 5.2: *Función de densidad de probabilidad, $f(x)$, y función de distribución, $F(x)$, para una distribución normal $N(\mu, \sigma)$*

Otros parámetros de interés de esta función de distribución son:

- el *coeficiente de asimetría*, que será $\gamma_1 = \dfrac{E\left((X-\mu)^3\right)}{\sigma^3} = 0$, es 0 porque es una función simétrica,

- y la *curtosis*, $\gamma_2 = \dfrac{E\left((X-\mu)^4\right)}{\sigma^4} - 3 = 0$, toma valor 0 porque se usa esta definición de la curtosis para medir el apuntamiento de las funciones de distribución utilizando la normal como distribución patrón.

5.3.3 Tipificación de una distribución normal

Supongamos ahora que queremos calcular la probabilidad de que la variable X, que sigue una distribución $N(\mu, \sigma)$ esté en un intervalo (a, b). En este caso tenemos que calcular $p(a < X < b) = \int_a^b f(x)\,dx$, por lo tanto $p(a < X < b) = \int_a^b f(x)\,dx = \dfrac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$. Haciendo el cambio de variable $y = x - \mu$ tenemos que $p(a > X < b) = \dfrac{1}{\sigma\sqrt{2\pi}} \int_{a-\mu}^{b-\mu} e^{-\frac{y^2}{2\sigma^2}}\,dy$. Usando ahora el

cambio de variable $z = \frac{y}{\sigma} = \frac{x-\mu}{\sigma}$ tendremos, finalmente,

$$p(a < X < b) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz$$

Por lo tanto, calcular la probabilidad de que la variable $X \sim N(\mu, \sigma)$ esté en el intervalo (a, b) es equivalente a calcular la probabilidad de que la variable $Z \sim N(0, 1)$ esté en el intervalo $\left(\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right)$.

Este procedimiento que permite, mediante un cambio de variable, pasar de una variable aleatoria que sigue una distribución $N(\mu, \sigma)$ a otra variable aleatoria que sigue una distribución $N(0, 1)$ se denomina *tipificación de la variable aleatoria.*

La ventaja de tipificar una variable aleatoria X que se comporta como una $N(\mu, \sigma)$ es que cuando queremos calcular probabilidades en un intervalo, como la integral $p(a < X < b) = \int_a^b f(x)dx$ no se puede resolver analíticamente y hay que recurrir la integración numérica o a las tablas con los resultados de dicha integración, basta con disponer de las tablas para la distribución normal $N(0, 1)$ y utilizar la variable tipificada $Z = \frac{X-\mu}{\sigma}$para hacer el cálculo de $p\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = p\left(a < X < b\right)$

Vamos ahora a ver cómo utilizar las tablas. Generalmente cuando se publican tablas se suele dar o bien el valor z_α tal que $\alpha = F(z_\alpha) = p(Z < z_\alpha)$, o sea, el valor de la variable aleatoria cuya cola inferior encierra un área α; o bien el valor z_α tal que $\alpha = p(Z > z_\alpha) = 1 - F(z_\alpha)$, esto es, el valor de la variable aleatoria cuya cola superior encierra un área α. Nosotros mostramos, en el apéndice A, una tabla para este último caso.

Por lo tanto, si nosotros necesitamos calcular $p(a < X < b)$ donde X es una variable $N(0, 1)$ lo que haremos será:

- Primero tipificamos nuestra variable, para ello definimos $Z = \frac{X-\mu}{\sigma}$ y calculamos $z_a = \frac{a-\mu}{\sigma}$ y $z_b = \frac{b-\mu}{\sigma}$.
- Segundo, buscamos en las tablas los valores α_a y α_b tales que $\alpha_a = p(Z > z_a)$ y $\alpha_b = p(Z > z_b)$, por lo que se cumplirá que $\alpha_a > \alpha_b$.
- Finalmente, como queremos $p(a < X < b) = p(z_a < Z < z_b) = p(Z < z_b) - p(Z < z_a) = F(z_b) - F(z_a)$, y sabemos que $\alpha_a = p(Z > z_a) = 1 - F(z_a)$ y $\alpha_b = p(Z > z_b) = 1 - F(z_b)$, tendremos que $p(a < X < b) = \alpha_a - \alpha_b$

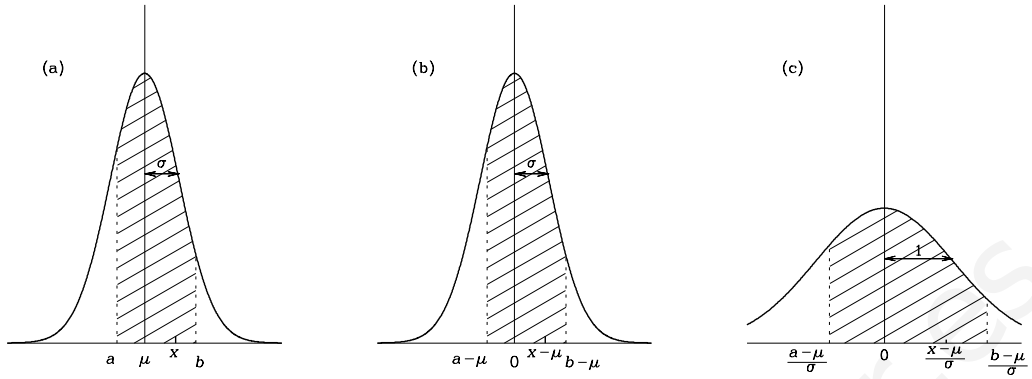


Figura 5.3: *Esquema del procedimiento de tipificación. (a) Distribución normal inicial,* $N(\mu, \sigma)$. *(b) Distribución normal,* $N(0, \sigma)$, *obtenida mediante una traslación* $X \to Y = X - \mu$. *(c) Distribución normal,* $N(0, 1)$, *obtenida al normalizar la variable* $Y = X - \mu$ *al valor de la desviación típica* σ, *de modo que* $Y = X - \mu \to Z = \dfrac{x - \mu}{\sigma}$

Ejemplo:

Calcular la probabilidad $p(-1.2 < X < 0.8)$ sabiendo que X es una variable aleatoria que sigue una distribución normal de media $\mu = 0.5$ y de varianza $\sigma^2 = 0.36$.

En primer lugar, tipificamos la variable, definiendo $Z = \dfrac{X - 0.5}{\sqrt{0.36}}$, por lo que, para esta variable, los límites del intervalo serán $z_1 = \dfrac{-1.2 - 0.5}{0.6} = -2.833$ y $z_2 = \dfrac{0.8 - 0.5}{0.6} = 0.5$.

Ahora buscamos en la tabla de la distribución $N(0, 1)$ los valores de las áreas asociadas a las colas superiores de los anteriores z_1 y z_2.

Para $z_2 = 0.5$ obtenemos que $p(Z > 0.5) = 0.3085$.

Para $z_1 = -2.833$, como en la tabla sólo dan los valores para $z \geq 0$, tenemos que calcularlo teniendo en cuenta que la distribución normal es simétrica, por lo que si tenemos un $z_\alpha > 0$ tal que $p(Z > z_\alpha) = \alpha$ entonces podemos escribir $p(Z > -z_\alpha) = 1 - \alpha$. Así que, en nuestro caso buscamos $p(Z > 2.833)$ obteniendo $p(Z > 2.833) \simeq 0.00231$ (donde hemos interporlado a partir de las tablas el valor sabiendo que $p(Z > 2.83) = 0.00233$ y $p(Z > 2.84) = 0.00226$), por lo que $p(Z > -2.833) = 1 - 0.00231 = 0.99769$

Finalmente tendremos que $p(-1.2 < X < 0.8) = 0.99769 - 0.3085 = 0.68919$.

5.4 Teorema del límite central

Este teorema nos permitirá, en determinadas circunstancias, aproximar cualquier distribución de probabilidad a una distribución normal.

Teorema del límite central: Sean $X_1, X_2, ..., X_n$ n variables aleatorias independientes con medias $\mu_1, \mu_2, ..., \mu_n$ y desviaciones típicas $\sigma_1, \sigma_2, ..., \sigma_n$, respectivamente, que siguen distribuciones de probabilidad cualesquiera. Si definimos la variable aleatoria $X = \sum_{i=1}^{n} X_i$, en el límite de n grande (esto es, $n \to \infty$) se comportará como una distribución normal de media $\mu = \sum_{i=1}^{n} \mu_i$ y desviación típica $\sigma = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$.

5.5 Aproximación de las distribuciones binomial y de Poisson a la normal.

5.5.1 Aproximación de una distribución binomial $Bin(n,p)$ a una distribución normal $N(\mu,\sigma)$

Hemos visto en el capítulo anterior que una variable aleatoria X que sigue una distribución binomial $Bin(n,p)$ se puede considerar como la suma de n variables de Bernoulli. Por este motivo, teniendo en cuenta el teorema del límite central, en el límite $n \to \infty$ la variable X se comportará siguiendo una distribución normal de media $\mu = np$ y varianza $\sigma^2 = np(1-p)$, por lo tanto, X seguirá una distribución $N(np, \sqrt{np(1-p)})$.

La aproximación $Bin(n,p) \approx N(np, \sqrt{np(1-p)})$ es tanto mejor cuanto mayor sea n, siendo p ni demasiado grande ni demasiado pequeño. En la práctica, se observa que la aproximación es relativamente buena si $np \geq 5$ y $n(1-p) \geq 5$

5.5.2 Aproximación de una distribución de Poisson $Poi(\lambda)$ a una distribución normal $N(\mu,\sigma)$

De forma similar al caso anterior, la distribución de Poisson de parámetro λ, $Poi(\lambda)$, se puede aproximar a una distribución normal cuando λ es grande (ya vimos que la distribución de Poisson tiende a ser simétrica para valores grandes de λ). Así, si tenemos una variable aleatoria

X que sigue una distribución de Poisson $Poi(\lambda)$ con λ grande, podremos aproximarla por una normal de media $\mu = \lambda$ y de varianza $\sigma^2 = \lambda$, por lo tanto, $Poi(\lambda) \approx N(\lambda, \sqrt{\lambda})$ si λ grande.

En la práctica, se observa que la aproximación es válida si $\lambda \geq 5$.

Igualmente, si tenemos n variables aleatorias X_i, que cada una de ellas sigue una distribución de Poisson de parámetro λ, la variable $X = \sum_{i=1}^{n} X_i$ seguirá una distribución de Poisson de parámetro $n\lambda$ y para valores grandes de $n\lambda$ se podrá a proximar por una normal $N(n\lambda, \sqrt{n\lambda})$

5.6 Corrección de continuidad

Cuando aproximamos una función discreta por una continua, por ejemplo $Poi(\lambda) \approx Nor(\lambda, \sqrt{\lambda})$ ó $Bin(n,p) \approx Nor(np, \sqrt{np(1-p)})$, cometemos un pequeño error por el hecho de pasar de una distribución discreta por una continua.

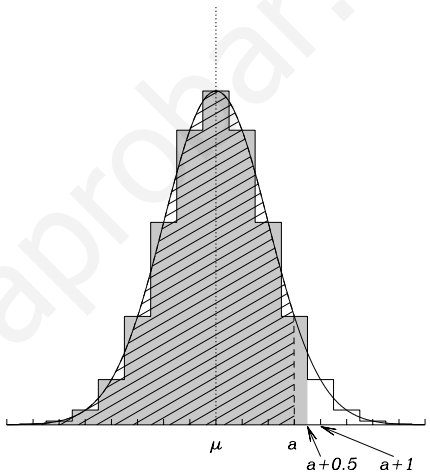


Figura 5.4: *Esquema para la corrección de continuidad en la aproximación de una distribución discreta por una continua.*

Para entender mejor esto, vamos a fijarnos en la Figura 5.4. Cuando calculamos para una distribución discreta la probabilidad de que una variable aleatoria X tome un valor menor o igual que una cierta cantidad a, esto es, $p(X \leq a)$ estaríamos calculando el área de la parte sombreada en gris de la Figura 5.4, que recorrería valores de x entre $-\infty$ y $a + 0.5$. Sin embargo, si suponemos que la variable X se puede aproximar por una distribución continua y calculamos $p(X \leq a)$, estaríamos calculando el área de la región de la Figura 5.4 que está rayada, y que recorrería valores de x entre $-\infty$ y a. Por este motivo, cuando hacemos la aproximación de la distribución discreta por la distribución continua, en el cálculo de la probabilidad estaríamos cometiendo un error del orden del área comprendida entre $x = a$ y $x = a + 0.5$.

Siguiendo este razonamiento, podemos establecer como debemos corregir los límites del intervalo utilizado para calcular la probabilidad de una determinada variable aleatoria X cuando su distribución de probabilidad es discreta y la aproximamos por una distribución continua para minimizar el error. La forma en la que debemos realizar la modificación de los límites está resumida en la siguiente tabla:

CORRECCIÓN DE CONTINUIDAD	
Cálculo de la probabilidad	
Distribución discreta	*Distribución continua*
$p(X \leq x)$	$p(X < x + 0.5)$
$p(X < x)$	$p(X < x - 0.5)$
$p(X \geq x)$	$p(X > x - 0.5)$
$p(X > x)$	$p(X > x + 0.5)$

5.7 Ejemplos

1. Supongamos que X es una variable aleatoria que sigue una distribución normal de media $\mu = 3$ y de desviación típica $\sigma = 5$. Calcular la probabilidad de que X esté comprendida entre 2 y 8.

   Sabemos que $X \sim N(3, 5)$, así que para poder calcular $p(2 < X < 8)$ primero tenemos que tipificar la variable. Para ello definimos una nueva variable $Z = \dfrac{X - \mu}{\sigma} = \dfrac{X - 3}{5}$, así tendremos que $p(2 < X < 8) = p(z_1 < Z < z_2)$ donde $z_1 = \dfrac{2 - 3}{5} = -0.2$ y $z_2 = \dfrac{8 - 3}{5} = 1$, siendo Z una variable aleatoria que sigue una distribución $N(0, 1)$.
   Como $p(z_1 < Z < z_2) = p(-0.2 < Z < 1) = p(Z < 1) - p(Z < -0.2) = p(Z < 1) - p(Z > 0.2) = 1 - p(Z > 1) - p(Z > 0.2) = 1 - 0.1587 - 0.4207 = 0.4206$

2. Consideremos una variable aleatoria X que sigue una distribución binomial $Bin(n = 100, p = 0.1)$. Calcular la probabilidad de que X sea estrictamente mayor que 17, o sea, $p(X > 17)$.

   Para poder calcular dicha probabilidad es necesario aplicar la aproximación a una distribución normal, puesto que de no hacerlo tendríamos que calcular:

   $$p(X > 17) = 1 - \sum_{k=0}^{17} p(X = k) = 1 - \sum_{k=0}^{17} \binom{100}{k} 0.1^k \, 0.9^{100-k}$$

   En este caso, como $n = 100$ y $p = 0.1$, tendremos que $np = 10 > 5$ y $n(1-p) = 90 > 5$ y, por lo tanto, tiene sentido aproximar $Bin(n, p) \approx N(np, \sqrt{np(1-p)}) \Rightarrow Bin(100, 0.1) \approx N(10, 3)$.
   Además, debemos considerar la corrección por continuidad para minimizar el error de la

aproximación, por lo tanto, como queremos calcular $p(X > 17)$, cuando aproximamos a la distribución continua, consideraremos $p(X > 17.5)$.

Para realizar el cálculo, en primer lugar tipificamos la variable, definiendo $Z = \dfrac{X-10}{3}$ que sigue una $N(0,1)$. Ahora calcularemos $p(X > 17.5) = p(Z > z_1)$ con $z_1 = \dfrac{17.5-10}{3} = 2.5$. Mirando en las tablas de la distribución $N(0,1)$, obtenemos $p(X > 17.5) = p(Z > 2.5) = 0.00621$

3. Consideremos una variable aleatoria X que se comporta siguiendo una distribución de Poisson $Poi(10)$. Calcular de forma exacta y utilizando la aproximación a una distribución normal la probabilidad de que X sea menor o igual que 6, $p(X \leq 6)$

   Para calcular $p(X \leq 6)$ utilizando la distribución de Poisson, calcularemos $p(X \leq 6) = p(0)+p(1)+p(2)+p(3)+p(4)+p(5)+p(6)$. Como $p(x) = \dfrac{\lambda^x}{x!}\, e^{-\lambda}$, con $\lambda = 10$ en este caso, tendremos que $p(X \leq 6) = e^{-10}\left(1 + 10 + \dfrac{10^2}{2} + \dfrac{10^3}{6} + \dfrac{10^4}{24} + \dfrac{10^5}{120} + \dfrac{10^6}{720}\right) = 0.1301$

   Como $\lambda = 10 > 5$ podemos aplicar la aproximación a una distribución normal $N(\lambda, \sqrt{\lambda}) = N(10, \sqrt{10})$. Aplicaremos además la corrección de continuidad para minimizar los errores, por lo tanto, como queremos calcular $p(X \leq 6)$, cuando utilicemos la distribución continua calcularemos $p(X < 6.5)$. Primero tipificamos la variable, definiendo $Z = \dfrac{X-10}{\sqrt{10}}$, que seguirá una distribución $N(0,1)$.
   Por lo tanto, $p(X \leq 6) = p(X < 6.5) = p(Z < -1.107) = p(Z > 1.107)$, mirando en las tablas e interpolando obtenemos que $p(X \leq 6) = 0.1341$, que vemos que difiere sólo en torno a un 3% del valor exacto.

Capítulo 6

Distribuciones continuas II

Tiempo de espera en un proceso de Poisson o distribución exponencial. Distribución gamma. Distribución χ^2 de Pearson. Distribución t de Student. Distribución F de Fisher.

6.1 Introducción

En este capítulo vamos a estudiar otras distribuciones continuas que son de gran utilidad en el análisis estadístico y que necesitaremos en capítulos posteriores.

6.2 Tiempo de espera en un proceso de Poisson o distribución exponencial

En muchos casos, cuando utilizamos una distribución de Poisson, lo que se analiza es la frecuencia de un suceso en un intervalo de tiempo. Resulta razonable, en estos casos, preguntarnos la probabilidad de tener que esperar un tiempo t para que el suceso se origine.

En una distribución de Poisson la función de densidad de probabilidad es de la forma:

$$P(X = x) = f(x) = \frac{(\alpha t)^x e^{-\alpha t}}{x!} \text{ donde } \alpha t = \lambda$$

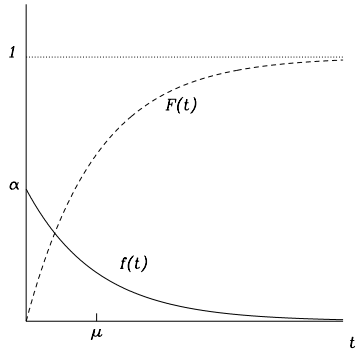


Figura 6.1: *Función de densidad de probabilidad, $f(t)$, y de distribución, $F(t)$, para el tiempo de espera en un proceso de Poisson.*

siendo λ el parámetro de la distribución de Poisson y α la tasa de ocurrencias del suceso (esto es, el número de veces que se verifica el suceso por unidad de tiempo).

La probabilidad de que tengamos que esperar un tiempo mayor que t será $p(T > t) = p(X = 0 \text{ en tiempo } t) = e^{-\alpha t} = 1 - F(t)$ siendo $F(t)$ la función de distribución para la variable T que mide el tiempo de espera. Como tenemos que $F(t) = 1 - e^{-\alpha t} \Rightarrow f(t) = \dfrac{dF(t)}{dt}$, obtenemos que la función de densidad de probabilidad para la variable tiempo de espera en un proceso de Poisson,T, es:

$$f(t) = \alpha e^{-\alpha t}$$

Calculamos ahora la *media* para esta distribución:

$$\mu = E(T) = \int_0^\infty \alpha\, t\, e^{-\alpha t} dt = \frac{1}{\alpha} \int_0^\infty u e^{-u} du \Rightarrow \mu = E(T) = \frac{1}{\alpha}$$

Y la *varianza:*

$$Var(T) = \sigma^2 = E(T^2) - (E(T))^2 = \frac{1}{\alpha^2} \int_0^\infty u^2 e^{-u} du - \frac{1}{\alpha^2} \Rightarrow Var(T) = \frac{1}{\alpha^2}$$

6.3 Distribución gamma

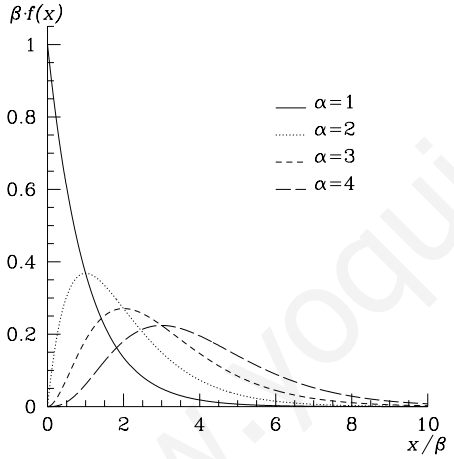


Figura 6.2*: Función de densidad, $f(x)$, para una distribución gamma para distintos valores del parámetro α. En el eje de abcisas se ha dibujado x/β y en el eje de ordenadas $\beta\, f(x)$.*

Decimos que una variable aleatoria continua X sigue una distribución gamma de parámetros $\alpha, \beta > 0$ si tiene una función de densidad de probabilidad de la forma:

$$f(x) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

donde $\Gamma(\alpha) = \displaystyle\int_0^\infty x^{\alpha-1} e^{-x} dx$ es la función gamma.

Se puede comprobar que la variable aleatoria X que sigue una distribución gamma cumple que su *media* es $\mu = E(X) = \alpha\beta$ y su *varianza* es $Var(X) = \sigma^2 = \alpha\beta^2$

Es fácil comprobar que la distribución exponencial es un caso particular de la distribución gamma, tomando $\alpha = 1$.

Otro caso particular de la distribución gamma es la distribución χ^2 de Pearson que veremos a continuación.

6.4 Distribución χ^2 de Pearson

Si en una distribución gamma consideramos los parámetros $\alpha = \frac{n}{2}$, siendo n un número natural, y $\beta = 2$ obtenemos una distribución de probabilidad denominada χ^2 (chi-cuadrado) con n grados de libertad, por lo tanto, su función de densidad de probabilidad es:

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

El interés de esta función de probabilidad es que si consideramos una variable aleatoria χ^2_n construída como suma del cuadrado de n variables aleatorias con una distribución normal $N(0,1)$, esto es, $\chi^2_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2$, dicha variable aleatoria sigue una distribución chi-cuadrado con n grados de libertad.

La variable χ^2_n sólo toma valores positivos y su distribución depende del parámetro n.

Dado que es un caso particular de una distribución gamma, es fácil ver que se caracteriza porque la *media* es $E(\chi^2_n) = \mu = n$ y la *varianza* es $Var(\chi^2_n) = \sigma^2 = 2n$.

El *número de grados de libertad* es el número de elementos independientes que contiene la variable, esto es, el número de elementos del conjunto menos el número de relaciones que existen entre ellos. En los siguientes capítulos analizaremos como obtener, en la práctica, el número de grados de libertad de una variable aleatoria que siga una distribución χ^2 para ejemplos concretos de interés estadístico.

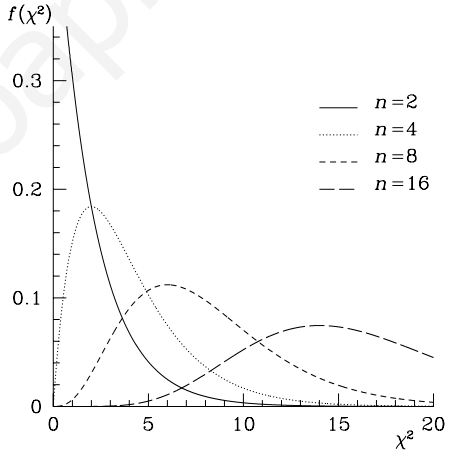


Figura 6.3: *Función de densidad de probabilidad, $f(\chi^2)$, de una distribución chi-cuadrado para varios grados de libertad ($n = 2$, 4, 8 y 16)*

Para $n \geq 30$ se puede aproximar la variable $\sqrt{2\chi^2_n}$ por una normal $N\left(\sqrt{2n-1}, 1\right)$.

Además, si $\chi^2_{n_1}$ y $\chi^2_{n_2}$ son dos variables aleatorias que siguen una distribución χ^2 con n_1 y n_2 grados de libertad, respectivamente, entonces la variable aleatoria $\chi^2_n = \chi^2_{n_1} + \chi^2_{n_2}$ sigue una distribución χ^2 con $n = n_1 + n_2$ grados de libertad.

Igualmente, según el teorema central del límite, para n grande podremos aproximar χ^2_n por una distribución normal $N(n, \sqrt{2n})$. En la práctica se acepta que la aproximación es buena para $n > 30$.

Como el cálculo de la probabilidad encerrada por una función de distribución χ^2_n entre dos valores de las coordenadas de abcisas hay que calcularlo numéricamente, en el apéndice B presentamos, para varios n, una tabla con algunos valores de abcisas, $\chi^2_{\alpha,n}$, que encierran a su derecha una probabilidad α.

6.5 Distribución t-Student

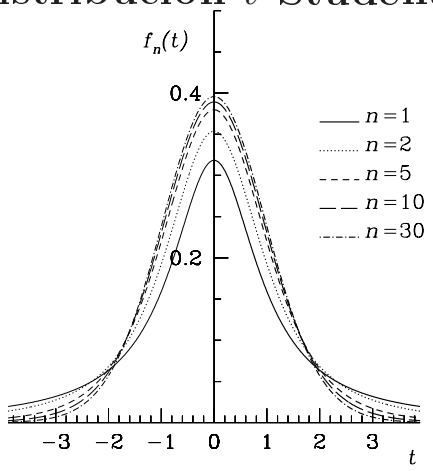


Figura 6.4*: Función de densidad de probabilidad, $f_n(t)$, para una distribución t-Student con n grados de libertad (para $n = 1, 2, 5, 10$ y 30)*

Supongamos que tenemos una variable aleatoria X que sigue una distribución χ^2 con n grados de libertad y otra variable aleatoria Z que sigue una distribución normal $N(0,1)$, si definimos una nueva variable aleatoria T tal que $T = \dfrac{Z}{\sqrt{X/n}}$, dicha variable aleatoria seguirá una distribución de probabilidad denominada t de Student con n grados de libertad, t_n.

Por lo tanto, si tenemos $n+1$ variables aleatorias independientes, Z_1, Z_2, ..., Z_n, Z, que siguen cada una de ellas una distribución $N(0,1)$, y construímos a partir de ellas, la variable aleatoria $T = \dfrac{Z}{\sqrt{\dfrac{1}{n}\sum_{i=1}^{n} Z_i^2}}$ seguirá una distribución t_n (t-Student con n grados de libertad).

La función de densidad de probabilidad de la distribución t-Student con n grados de libertad es:

$$f_n(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Dicha distribución cumple que su *media* es $E(T) = \mu = 0$ y su *varianza* es $Var(T) = \sigma^2 = \dfrac{n}{n-2}$ para $n > 2$.

Se puede comprobar que en el límite $n \to \infty$ la distribución $f_n(t)$ tiende a una distribución normal: $\lim_{n\to\infty} f_n(t) = \lim_{n\to\infty} \dfrac{1}{\sqrt{n\pi}} \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \dfrac{t^2}{n}\right)^{-\frac{n+1}{2}} = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \to N(0,1)$

Cuando n aumenta $f_n(t)$ se va haciendo cada vez más apuntada, tendiendo a la distribución de densidad normal tipificada, $N(0,1)$, cuando $n \to \infty$. En la práctica, la aproximación de la distribución t-Student por una distribución normal tipificada es válida cuando $n \geq 30$.

Como la función de probabilidad de una t de Student de n grados de libertad hay que obtenerla numéricamente, para facilitar el cálculo mostramos, en el apéndice C, una tabla con las abcisas, $t_{\alpha,n}$, de función de distribución t de Student que encierran a su derecha una probabilidad α (para varios n).

6.6 Distribución F de Fisher (Snedecor-Fisher)

Si tenemos dos variables aleatorias X e Y que siguen distribuciones χ^2 con n_1 y n_2 grados de libertad, respectivamente, entonces la variable aleatoria F definida como $F = \frac{X/n_1}{Y/n_2}$ sigue una distribución F de Fisher con n_1 grados de libertad en el numerador y n_2 grados de libertad en el denominador, $F_{(n_1,n_2)}$. Dado que las variables X e Y siguen distribuciones χ^2 que sólo toman valores positivos, la variable F sólo tomará valores positivos.

La función de densidad de probabilidad de la distribución F de Fisher será:

$$f_{n_1,n_2}(x) = \begin{cases} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} x^{\frac{n_1}{2}-1} \left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} & \text{, si } x > 0 \\ 0 & \text{, si } x \leq 0 \end{cases}$$

Se puede comprobar que $F_{(n_1,n_2)} = \frac{1}{F_{(n_2,n_1)}}$, por lo que es fundamental establecer correctamente el orden de los índices.

Una distribución F de Fisher se caracteriza porque su *media* es $E(F) = \mu = \frac{n_1}{n_2 - 2}$ para $n > 2$ y su *varianza* es $Var(F) = \sigma^2 = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}$ para $n_2 > 4$.

Se puede ver que las distribuciones χ^2 y t-Student son casos particulares de la F de Fisher, ya que $F_{(1,n)} = (t_n)^2$ y $F_{(n,\infty)} = \frac{\chi_n^2}{n}$

En el apéndice D se ofrece una tabla para facilitar el cálculo de probabilidades para una función de distribución F de Fisher.

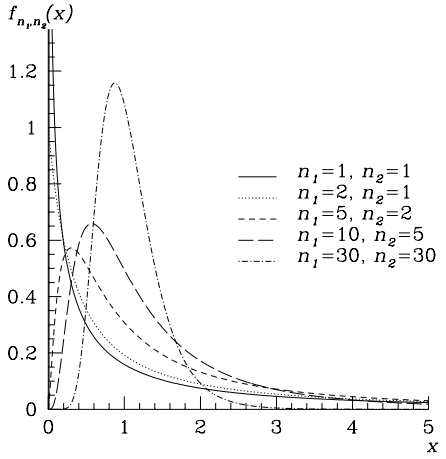


Figura 6.5: *Función de densidad de probabilidad, $f_{n_1,n_2}(x)$, para una distribución F de Fisher, para varios grados de libertad n_1 y n_2.*

Capítulo 7

Muestreo aleatorio y distribuciones muestrales

Muestreo aleatorio. Muestra aleatoria. Estadística descriptiva. Estadísticos muestrales. Media y varianza muestrales. Distribuciones muestrales.

7.1 Introducción al muestreo aleatorio

Vamos a recordar ciertos conceptos mencionados en capítulos anteriores y a definir otros nuevos que necesitaremos utilizar en este capítulo y en los próximos:

- *Población:* Es el conjunto completo de elementos, por ejemplo, los sucesos de un experimento. Puede ser finita o infinita.
- *Muestra:* Es un subconjunto de la población. El número de elementos de una muestra se denomina *tamaño muestral.*
- *Muestreo:* proceso de obtener muestras.
- *Inferencia estadística:* Son los métodos necesarios para extraer o inferir conclusiones válidas e información sobre una población a partir del estudio experimental de una muestra. Dependiendo de nuestro conocimiento sobre la muestra podemos utilizar:
  - *Métodos paramétricos:* usados cuando se conoce la forma de la distribución de la población y queremos encontrar el valor de los parámetros que la definen (por ejemplo, la media y la varianza para una distribución normal).
  - *Métodos no paramétricos:* se usan cuando la distribución poblacional es desconocida y el problema es encontrar la forma y características de la distribución.

Nosotros nos centraremos primero en los métodos paramétricos.

- *Muestreo aleatorio:* Para poder estudiar una población mediante inferencia estadística es fundamental que la muestra esté bien escogida, o sea, que sea representativa de la población. Una forma de hacerlo es asegurar que todos los elementos de la población tengan la misma probabilidad de ser escogidos, es lo que se denomina un muestreo aleatorio.

- *Parámetros poblacionales:* Suponiendo que conocemos para una población la distribución $f(x)$ que sigue una variable aleatoria X, para conocer la población necesitamos calcular los parámetros que caracterizan la distribución, dichos parámetros se denominan parámetros poblacionales. Por ejemplo, la media y la varianza en la distribución normal, la media en una distribución de Poisson, etc.

- *Estadísticos:* Dada una variable aleatoria X, en un muestreo aleatorio obtendremos n medidas de X. Cada una de esas medidas es una variable aleatoria X_1, X_2, ..., X_n, para las cuales en nuestra muestra obtendremos los valores numéricos x_1, x_2, ..., x_n. Un estadístico es cualquier función de las n variables aleatorias de la muestra $g\left(X_1, X_2, ..., X_n\right)$.
  Los parámetros poblacionales los designaremos, en general, con letras griegas (μ, σ, λ, ...) y los estadísticos por letras romanas ($\bar{X}$, S, ...)

- *Estimadores:* A cada parámetro poblacional le corresponderá un estadístico de la muestra que permitirá hacer una estimación del parámetro poblacional.
  Por ejemplo, la media poblacional o esperanza matemática μ se puede estimar mediante el estadístico media muestral $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (lo veremos más adelante). Para una muestra dada x_1, x_2, ..., x_n podemos obtener la estimación de la media poblacional calculando la media muestral $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, el valor $\hat{\mu} = \bar{x}$ es una estimación de μ obtenida a partir del estadístico $\bar{X}$ que es un estimador de μ.

7.2 Estadística descriptiva

7.2.1 Tablas de frecuencias

Supongamos que tenemos una población para la que definimos una variable aleatoria X y obtenemos una muestra de n valores. Podemos analizar y describir la población estudiando los valores obtenidos en la muestra y calculando estadísticos muestrales que nos den información sobre los parámetros poblacionales. En primer lugar vamos a ver cómo analizar la muestra.

(a) Tablas de frecuencia de una variable discreta:

Cuando tenemos una variable aleatoria discreta X que puede tomar los valores x_1, x_2, ..., x_k podemos estudiar la frecuencia con que dichos valores aparecen en nuestra muestra de modo que obtengamos tanto las frecuencias absolutas n_i (número de veces que aparece el valor x_i en nuestra muestra), como las relativas $f_i = \dfrac{n_i}{n}$, así como las frecuencias acumuladas, tanto absolutas, $N_i = \displaystyle\sum_{j=1}^{i} n_j$, como relativas, $F_i = \displaystyle\sum_{j=1}^{i} \frac{N_i}{n}$, tal y como se describe en la tabla siguiente:

Valor de la variable	x_1	x_2	$\cdots$	x_k	
Frecuencia absoluta	n_1	n_2	$\cdots$	n_k	$\displaystyle\sum_{i=1}^{k} n_i = n$
Frecuencia relativa	$f_1 = \dfrac{n_1}{n}$	$f_2 = \dfrac{n_2}{n}$	$\cdots$	$f_k = \dfrac{n_k}{n}$	$\displaystyle\sum_{i=1}^{k} f_i = 1$
Frecuencia absoluta acumulada	$N_1 = n_1$	$N_2 = n_2 + n_1$	$\cdots$	$N_k = \displaystyle\sum_{i=1}^{k} n_i$	$N_k = n$
Frecuencia relativa acumulada	$F_1 = \dfrac{N_1}{n}$	$F_2 = \dfrac{N_2}{n}$	$\cdots$	$F_k = \dfrac{N_k}{n}$	$F_k = 1$

Ejemplo:

Supongamos que realizamos un muestreo en 20 familias midiendo el número de hijos en la familia, obteniéndose los siguientes valores $\{2, 1, 1, 3, 1, 2, 5, 1, 2, 3, 4, 2, 3, 2, 1, 4, 2, 3, 2, 1\}$.
La tabla de frecuencias para esta muestra será:

x_i	n_i	$f_i = \dfrac{n_i}{20}$	N_i	$F_i = \dfrac{N_i}{20}$
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

(b) Agrupamientos en intervalos de clase:

En determinadas ocasiones, cuando el número de valores distintos que toma la variable estadística es demasiado grande o la variable es continua, se definen intervalos $[a_i, a_{i+1})$, que quedan representados por el punto medio del intervalo $c_i = \dfrac{a_i + a_{i+1}}{2}$.

El número de intervalos se suele definir entre 5 y 20. En general, suele usarse el entero más cercano $\sqrt{n}$, donde n es el número de medidas.

(c) Tablas de frecuencia de doble entrada:

Cuando la variable aleatoria que queremos analizar es bidimensional, (X, Y), y obtenemos en nuestra muestra n pares de medidas se usan tablas de frecuencia de doble entrada.

Al igual que en el caso de una variable unidimensional podemos definir las frecuencias absolutas n_{ij} (número de veces que obtenemos el par (x_i, y_j)) y las frecuencias relativas $f_{ij} = \dfrac{n_{ij}}{n}$.

A partir de estos valores, podemos definir las frecuencias relativas condicionadas $f(x_i|y_j) = \dfrac{n_{ij}}{n_{y_j}}$, siendo $n_{y_j} = \displaystyle\sum_{\forall i} n_{ij}$.

Las tablas de frecuencias de este tipo de muestras tienen la estructura siguiente:

x \ y	y_1	y_2	$\cdots$	y_m	
x_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{x_1} = \displaystyle\sum_{j=1}^{m} n_{1j}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{x_2} = \displaystyle\sum_{j=1}^{m} n_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{x_k} = \displaystyle\sum_{j=1}^{m} n_{kj}$
	$n_{y_1} = \displaystyle\sum_{i=1}^{k} n_{i1}$	$n_{y_2} = \displaystyle\sum_{i=1}^{k} n_{i2}$	$\cdots$	$n_{y_m} = \displaystyle\sum_{i=1}^{k} n_{im}$	n

7.2.2 Estadísticos muestrales

(a) Medidas de centralización:

Para poder describir una variable aleatoria de una población uno de los aspectos que son de interés es saber en torno a qué valores se agrupa dicha variable aleatoria. Para ello utilizaremos estadísticos que nos den una medida de la centralización de la distribución (en torno a qué valores está centrada). Aquí vamos mencionar algunos de lo más utilizados.

- *Media muestral:* se define como $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  En el ejemplo anterior sobre el número de hijos por familia tendremos que $\bar{x} = 2.5$

- *Mediana:* M_e, define una medida central tal que, con los datos ordenados de menor a mayor, el 50% de los datos son inferiores a M_e y el 50% de los datos tiene valores superiores.
  En el ejemplo sobre el número de hijos por familia, la mediana es $M_e = 2$

- *Moda:* M_o, es el valor de la variable que tiene una frecuencia mayor.
  De nuevo en el ejemplo del número de hijos por familia tendremos que la moda es $M_o = 2$

- *Cuartiles:* son una generalización del concepto de mediana.
  Primer cuartil $Q_{1/4}$: el 25% de los datos son menores que $Q_{1/4}$
  Segundo cuartil $Q_{1/2} = M_e$
  Tercer cuartil $Q_{3/4}$: el 75% de los datos son menores que $Q_{3/4}$
  De forma similar se pueden definir los deciles (dividen la muestra en 10 partes iguales) y los percentiles (dividen la muestra en 100 partes iguales), por ejemplo, p_k deja el k% de los datos por debajo de p_k.

(b) Medidas de dispersión:

Al igual que es importante saber en torno a qué valores está centrada una distribución, también es necesario conocer la dispersión de la distribución. Para ello se utilizan algunos estimadores que nos dan una medida de la dispersión. Pasamos a enumerar aquí algunos de los más utilizados:

- *Recorridos o rangos:* son la diferencia entre el valor máximo y mínimo de la variable estadística.
  Recorrido intercuartílico: diferencia entre el tercer y primer cuartil, $R_I = Q_{3/4} - Q_{1/4}$

- *Desviación media:* se define como la media aritmética de las diferencias absolutas respecto a la media, $D_m = \dfrac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}\right|$
- *Varianza:* es la media aritmética de los cuadrados de las diferencias con la media, $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- *Desviación típica:* es la raíz cuadrada positiva de la varianza, $S = \sqrt{S^2}$
- *Coeficiente de variación:* es el cociente entre la desviación típica y la media muestral, $CV = \dfrac{S}{\bar{X}}$.
- *Cuasivarianza muestral:* se define como la varianza multiplicada por el factor $\dfrac{n}{n-1}$, por lo tanto, $\mathcal{S}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- *Cuasidesviación típica:* es la raíz cuadrada positiva de la cuasivarianza, $\mathcal{S} = \sqrt{\mathcal{S}^2}$

(c) Momentos:

Al igual que definimos los momentos de una distribución, podemos definir los estadísticos que nos permitan evaluar dichos momentos de la distribución a partir de nuestra muestra.

- *Momentos respecto al origen:* el momento de orden r respecto al origen está definido como $a_r = \dfrac{1}{n}\sum_{i=1}^{n}X_i^r$
- *Momentos respeto a la media:* el momento de orden r respecto a la media se define como $m_r = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^r$

(d) Características de forma:

Existen estadísticos que nos permiten analizar la forma de nuestra distribución, los más característicos son los coeficientes de asimetría o sesgo y los de curtosis o aplastamiento. Vamos a mencionar los más utilizados:

- *Coeficiente de asimetría de Fisher:* se define como $g_1 = \frac{m_3}{S^3}$. Si $g_1 > 0$, la muestra presenta una cola alargada a la derecha (sesgo positivo). Si $g_1 < 0$, la muestra presenta una cola alargada a la izquierda (sesgo negativo). Y, finalmente, si $g_1 = 0$, es simétrica.

- *Coeficiente de curtosis de Fisher:* definido como $g_2 = \frac{m_4}{S^4} - 3$. Si $g_2 > 0$, la muestra es más apuntada que una distribución normal (leptocúrtica) . Si $g_2 < 0$, es más aplastada que una normal (platicúrtica). Y, finalmente, si $g_2 = 0$, es igual de apuntada que la normal (mesocúrtica).

7.3 Media muestral

Uno de los estadísticos más importantes es la media muestral:

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

Cuando las variables aleatorias X_i tomen en una muestra los valores particulares x_i, el valor que tendrá la media muestral será $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Si tomamos varias muestras, los valores x_i diferirán y, por lo tanto, el valor de la media muestral obtenida, $\bar{x}$, será distinto.

7.3.1 Distribución muestral de la media

Por lo comentado anterioremente se puede ver que, al ser la media muestral una combinación de variables aleatorias, es en si misma una variable aleatoria. De tal forma que si nosotros obtenemos k muestras (cada una de ella con n medidas), las medias muestrales obtenidas $\bar{x}_1, \bar{x}_2$, ..., $\bar{x}_k$ serán, en general, diferentes unas de otras. Si hacemos que k tienda a ∞, los valores de $\bar{x}_i$ seguirán una distribución de probabilidad $f(\bar{x})$ que se denomina *distribución muestral de la media.*

Vamos a analizar la forma que tendrá dicha distribución muestral de la media. Por el hecho de ser el estadístico $\bar{X}$ una variable aleatoria, podemos calcular su esperanza matemática, $E(\bar{X}) = \mu_{\bar{X}}$, y su varianza, $Var(\bar{X}) = \sigma^2_{\bar{X}}$.

(a) Esperanza matemática de la media muestral:

Vamos a calcular la media o esperanza matemática de la media muestral, para ello utilizaremos la propiedad de linealidad del operador esperanza matemática:

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}\left(E(X_1) + E(X_2) + \cdots + E(X_n)\right)$$

Como $E(X_i) = \mu$, ya que X_i sigue la misma distribución que la variable aleatoria X (siempre y cuando nuestro muestreo sea aleatorio), tendremos que $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mu \Rightarrow$

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

(b) Varianza de la media muestral:

Pasemos ahora a calcular la varianza de la distribución de probabilidad de la media muestral $\bar{X}$. Para ello tendremos en cuenta que dado que las variables aleatorias X_i siguen la misma distribución que la variable aleatoria X, se cumplirá que $Var(X_i) = Var(X) = \sigma^2$. Además, las variables X_i son independientes, por lo tanto $Cov(X_i, X_j) = 0$ si $i \neq j$. Con todo esto, tendremos que:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(X_i, X_j) = \frac{1}{n^2} n Var(X) \Rightarrow$$

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

(c) Forma de la distribución muestral de la media:

La forma de la distribución muestral de la media dependerá, en principio de la población de partida, pero, teniendo en cuenta el postulado del teorema del límite central, se puede establecer que $\bar{X}$, que es la suma de n variables aleatorias, seguirá una distribución que tiende asintóticamente a una normal de media μ y de varianza $\frac{\sigma^2}{n}$.

De este modo, si $\bar{X}$ es la media de una muestra aleatoria de tamaño n para una población con distribución cualquiera de media μ y de varianza σ^2, entonces la variable aleatoria tipificada $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ tiende a una distribución normal $N(0,1)$ cuando n tiende a infinito.

En la práctica se considera que $\bar{X}$ sigue una distribución normal si $n > 30$. Además, en el caso particular de que la variable aleatoria inicial X (y, por tanto, también las X_i) siguiera una distribución normal, entonces $\bar{X}$, al ser la suma de n variables aleatorias normales, seguirá una distribución normal independientemente del tamaño n de la muestra.

7.3.2 Distribución muestral de la diferencia de medias

Como consecuencia de lo anterior, vamos a analizar el comportamiento que tendrá una variable aleatoria definida como la diferencia de medias de dos poblaciones. Supongamos que tenemos dos poblaciones, la primera de ellas con media μ_1 y varianza σ_1^2 y la segunda con media μ_2 y varianza σ_2^2, y que extraemos muestras independientes de cada una de ellas, con tamaños n_1 y n_2, respectivamente. Los estadísticos $\bar{X}_1$ y $\bar{X}_2$ representarán las medias muestrales de dichas poblaciones. A partir de dichos estadísticos, podemos definir un nuevo estadístico consistente en la diferencia de las medias muestrales $\bar{X}_1 - \bar{X}_2$. Al ser dicho estadístico la suma dos variables aleatorias, será una nueva variable aleatoria que seguirá una distribución de probabilidad denominada *distribución muestral de la diferencia de medias.*

Podemos calcular la media o esperanza matemática para la diferencia de medias y, dado que el operador esperanza matemática es un operador lineal, se obtiene que $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) \Rightarrow$

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

Por otro lado, si calculamos la varianza de la diferencia de medias, teniendo en cuenta que $\bar{X}_1$ y $\bar{X}_2$ son variables aleatorias independientes, tendremos que $Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) - 2Cov(\bar{X}_1, \bar{X}_2) = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} \Rightarrow$

$$Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Respecto a la forma de la distribución, debido al teorema del límite central, cuando n_1 y n_2 tienden a infinito, la variable aleatoria $\bar{X}_1 - \bar{X}_2$ sigue una distribución normal de media $\mu_1 - \mu_2$ y de varianza $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$, de modo que la variable $Z = \dfrac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ se comporta como una variable normal tipificada, $N(0,1)$. En la práctica, si se cumple que $n_1 + n_2 > 30$ con $n_1 \approx n_2$ se puede aplicar la aproximación normal. Además, si las poblaciones iniciales fuesen normales, entonces $\bar{X}_1 - \bar{X}_2$ seguiría una distribución normal, independientemente de los tamaños muestrales n_1 y n_2

7.3.3 Distribución muestral de una proporción

Vamos a analizar ahora una población que sigue un proceso de Bernoulli, es decir, se realizan n ensayos y cada uno de ellos tiene una probabilidad p de éxito y $1 - p$ de fracaso.

Consideremos el estadístico P que define la proporción de éxitos, esto es, el número de éxitos entre el tamaño de la muestra n. Este estadístico P puede considerarse la media muestral de

una variable de Bernoulli, y seguirá una distribución de probabilidad denominada *distribución muestral de una proporción*, que será un caso particular de la distribución muestral de la media.

Para analizar esta distribución hay que recordar que una variable de Bernoulli tiene una media dada por $\mu = p$ y una varianza $\sigma^2 = p(1-p)$.

Por lo tanto, la media y la varianza de la distribución de una proporción las podemos calcular aplicando lo que sabemos sobre la distribución muestral de la media que hemos visto anteriormente. Así tendremos que:

$$E(P) = \mu_P = \mu = p$$

$$Var(P) = \sigma_P^2 = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$$

Dado que la distribución muestral de la proporción es un caso particular de una distribución muestral de la media, sabemos que tiende a una distribución normal para número de ensayos n tendiendo a infinito. En ese caso, la variable $Z = \dfrac{P-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ se comportará como una normal tipificada $N(0,1)$. En la práctica la aproximación se considera válida cuando $n > 30$ ó, aplicando los criterios de aproximación de una variable binomial a una normal, cuando $np > 5$ y $n(1-p) > 5$ (dado que nP se comportará como una binomial de media np y de varianza $np(1-p)$).

7.4 Cuasivarianza muestral

Otro de los estadísticos importantes en el análisis de una población es la cuasivarianza muestral:

$$\mathcal{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

donde $\bar{X}$ es la media muestral. Más adelante analizaremos los motivos por los cuales resulta más interesante utilizar la cuasivarianza muestral en lugar de la varianza muestral.

Cuando las variables aleatorias X_i tomen en una muestra los valores particulares x_i, el valor que tendrá la cuasivarianza muestral será $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ con $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

7.4.1 Distribución muestral de la cuasivarianza

Dado que $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, podemos ver que si tomamos varias muestras, los valores x_i diferirán y, por lo tanto, el valor de la cuasivarianza muestral obtenida, s^2, será distinto. Por lo tanto $\mathcal{S}^2$ se comportará como un estadístico que seguirá una distribución de probabilidad denominada *distribución muestral de la cuasivarianza.*

Vamos a calcular ahora el valor esperado del estadístico $\mathcal{S}^2$:

$$E(\mathcal{S}^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}E\left((X_i - \bar{X})^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}E\left(\left((X_i - \mu) - (\bar{X} - \mu)\right)^2\right) =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}E\left((X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left((X_i - \mu)^2\right) + nE\left((\bar{X} - \mu)^2\right) - \right.$$

$$\left. 2E\left((\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu)\right)\right) = \frac{1}{n-1}\left(n\sigma^2 + n\frac{\sigma^2}{n} - 2nE\left((\bar{X} - \mu)^2\right)\right) = \frac{1}{n-1}\left(n\sigma^2 + \sigma^2 - 2n\frac{\sigma^2}{n}\right) \Rightarrow$$

$$E(\mathcal{S}^2) = \mu_{\mathcal{S}^2} = \sigma^2$$

Como podemos ver, el interés de analizar la cuasivarianza muestral (en lugar de la varianza muestral) es que se comporta como una variable aleatoria cuya media es la varianza poblacional y, por tanto, nos puede servir para estimar dicho parámetro.

7.4.2 Distribución muestral de $(n-1)\dfrac{\mathcal{S}^2}{\sigma^2}$

Vamos a definir ahora el estadístico $(n-1)\dfrac{\mathcal{S}^2}{\sigma^2}$ y estudiar su distribución, ya que nos resultará más cómodo que trabajar con $\mathcal{S}^2$. Esta variable aleatoria es, por lo tanto,

$$(n-1)\frac{\mathcal{S}^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2}$$

Para analizar como se comporta esta distribución vamos a considerar que la variable aleatoria X o bien sigue una distribución normal o bien el tamaño muestral es suficientemente grande ($n > 30$) para poder aproximarla a una normal. En este caso, la variable aleatoria $Z_i = \dfrac{X_i - \mu}{\sigma}$ se comportará como una normal tipificada y, por lo tanto, $\chi^2 = \sum_{i=1}^{n}\dfrac{(X_i - \mu)^2}{\sigma^2}$ será la suma de los cuadrados de n variables normales tipificadas, por lo tanto seguirá una distribución chi-cuadrado con n grados de libertad, χ^2_n. Como

$$\chi^2=\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}=\sum_{i=1}^{n}\frac{\left((X_i-\bar{X})+(\bar{X}-\mu)\right)^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}+n\frac{(\bar{X}-\mu)^2}{\sigma^2}+2\frac{\bar{X}-\mu}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})\Rightarrow$$

$$\chi^2=(n-1)\frac{\mathcal{S}^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

Como la variable aleatoria $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ se comporta como una normal tipificada, según hemos visto anteriormente, entonces $\left(\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ seguirá una distribución chi-cuadrado con un grado de libertad, χ^2_1, y como χ^2 es una chi-cuadrado con n grados de libertad, podemos deducir que $(n-1)\dfrac{\mathcal{S}^2}{\sigma^2}$ sigue una distribución chi-cuadrado con $n-1$ grados de libertad, χ^2_{n-1}. El motivo es que en nuestra muestra tenemos n valores de X_i, pero como debemos calcular $\bar{X}$ a partir de dichos X_i, perdemos un grado de libertad y nos quedaremos con $n-1$ grados de libertad.

7.4.3 Distribución muestral de la media cuando la varianza es desconocida

Vimos anteriormente que al analizar la distribución de la media muestral, $\bar{X}$, obteníamos que tendía a una normal de media μ y de varianza σ^2/n, siendo μ y σ^2 la media y la varianza, respectivamente, de la variable aleatoria X. Sin embargo, en la mayoría de los casos no se conoce, a priori, la varianza σ^2 de la población y tenemos que estimarla a partir de la cuasivarianza muestral $\mathcal{S}^2$. En este caso, en lugar de definir la variable aleatoria $Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ que se comporta como una distribución $N(0,1)$, definiremos la variable aleatoria:

$$T=\frac{\bar{X}-\mu}{\mathcal{S}/\sqrt{n}}$$

Vamos a analizar ahora el comportamiento de este nuevo estadístico.

Como $T=\dfrac{\bar{X}-\mu}{\mathcal{S}/\sqrt{n}}=\dfrac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\mathcal{S}/\sigma}=\dfrac{Z}{\sqrt{\dfrac{1}{n-1}(n-1)\dfrac{\mathcal{S}^2}{\sigma^2}}}=\dfrac{Z}{\sqrt{\dfrac{\chi^2}{n-1}}}$, donde $Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ es una variable normal $N(0,1)$ y $\chi^2=(n-1)\dfrac{\mathcal{S}^2}{\sigma^2}$, según vimos antes, es una chi-cuadrado con $n-1$ grados de libertad, χ^2_{n-1}. Por lo tanto la variable T se comportará como una t de Student con $n-1$ grados de libertad. De nuevo, el motivo por el cual son $n-1$ grados de libertad se explica porque, aunque al tamaño de la muestra es n, perdemos un grado de libertad al estimar la varianza mediante el estadístico $\mathcal{S}^2$.

Cuando el tamaño muestral es muy grande (n tendiendo a infinito), la variable aleatoria T se puede aproximar por una normal $N(0,1)$. En la práctica usaremos la aproximación para $n > 30$.

7.4.4 Distribución muestral del cociente de varianzas

Antes hemos visto que para comparar dos poblaciones independientes estudiabamos la distribución muestral de la diferencia de sus medias. Podríamos hacer lo mismo en el caso de las varianzas para comparar si dos poblaciones independientes tienen la misma varianza o no. Sin embargo, la distribución muestral de la diferencia de varianzas es muy complicada. Resulta más sencillo, como vamos a ver a continuación, definir un estadístico que esté relacionado con el cociente de varianzas.

Supongamos que tenemos dos poblaciones normales o que se puedan aproximar por una normal cuyas varianzas poblacionales son σ_1^2 y σ_2^2, respectivamente. Sean $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$, respectivamente, las cuasivarianzas muestrales de las dos poblaciones medidas para los datos de nuestra muestra aleatoria. A partir de esto, definimos el siguiente estadístico F como:

$$F = \frac{\mathcal{S}_1^2/\sigma_1^2}{\mathcal{S}_2^2/\sigma_2^2}$$

Para cada par de muestras aleatorias de tamaños n_1 y n_2 el valor de este estadístico será diferente. La distribución de probabilidad que seguirá F se denomina *distribución muestral del cociente de varianzas*. Como sabemos que $(n_1-1)\dfrac{\mathcal{S}_1^2}{\sigma_1^2}$ y $(n_2-1)\dfrac{\mathcal{S}_2^2}{\sigma_2^2}$ son variables aleatorias que se comportan como una $\chi^2_{n_1-1}$ y una $\chi^2_{n_2-1}$, respectivamente, nuestro estadístico F será un cociente de distribuciones chi-cuadrado divididas cada una de ellas por sus grado de libertad, lo cual, como ya vimos en capítulos anteriores, se comporta como una distribución F de Fisher con n_1-1 grados de libertad en el numerador y n_2-1 grados de libertad en el denominador, $F_{(n_1-1,n_2-1)}$.

Capítulo 8

Estimación puntual y por intervalos

Características de los estimadores. Estimadores puntuales. Estimación por intervalos. Intervalos para proporciones, medias y varianzas.

8.1 Características de los estimadores

Hemos visto en el capítulo anterior que determinados estadísticos pueden ser útiles para la estimación de los parámetros de una distribución de probabilidad. En esta sección vamos a analizar qué propiedades debe cumplir un buen estimador, de modo que proporcione una estimación lo más precisa posible del parámetro poblacional que queremos conocer.

8.1.1 Propiedades de los estimadores

Un buen estimador debe cumplir lo siguiente:

- *Estimador insesgado o centrado*: decimos que un estimador A de un parámetro poblacional θ es insesgado o centrado si su media coincide con el parámetro poblacional. Es decir, si $E(A) = \mu_A = \theta$ entonces A es un estimador insesgado del parámetro θ.
  Por ejemplo, $\bar{X}$ es un estimador insesgado de la media de una población, μ_X, y $\mathcal{S}^2$ es un estimador insesgado de la varianza poblacional, σ_X^2.

- *Estimador eficiente*: decimos que un estimador A_1 de un parámetro poblacional θ es más eficiente que otro estimador A_2 de θ, si su varianza es menor, o sea, si $\sigma_{A_1}^2 < \sigma_{A_2}^2$.

- *Estimador consistente*: decimos que un estimador A de un parámetro θ es consistente si al crecer el tamaño muestral se aproxima asintóticamente al valor del parámetro poblacional y su varianza se hace nula. Por lo tanto, A es consistente si $\lim_{n\to\infty} A = \theta$ y $\lim_{n\to\infty} \sigma_A^2 = 0$

Un estimador ideal debe ser insesgado y con la máxima eficiencia. Sin embargo, en la práctica no siempre es posible calcular dichos estimadores. En cualquier caso, el requisito mínimo que debe cumplir cualquier estimador es que sea consistente.

8.1.2 Obtención de los estimadores: método de máxima verosimilitud

Cuando desconocemos un parámetro θ de la función de densidad de probabilidad de una población, uno de los objetivos es encontrar un estadístico que nos permita hacer una estimación de dicho parámetro a partir de los valores obtenidos para la variable aleatoria en la muestra, X_1, X_2,..., X_n.
Existen diversos métodos para obtener estimadores de parámetros poblacionales. Uno de ellos es el *método de máxima verosimilitud,* que consiste en construir la distribución de densidad de probabilidad conjunta de la muestra $L = f(X_1, X_2, \ldots, X_n; \theta)$ (denominada función de máxima verosimilitud) y se define el estimador de máxima verosimilitud como el parámetro θ que maximiza dicha función $f(X_1, X_2, \ldots, X_n; \theta)$ con respecto a θ. Por motivos prácticos, se suele utilizar el logaritmo de la función de densidad conjunta, por lo que el método de máxima verosimilitud se reduce a resolver la expresión $\dfrac{d \ln L}{d\theta} = 0$

8.1.3 Procedimientos para estimar un parámetro poblacional

Existen dos procedimientos para estimar los parámetros poblacionales de una distribución de probabilidad:

- El primero es, una vez elegido el estimador que vamos a utilizar, calcular una única estimación del parámetro poblacional a partir de los datos muestrales. En este caso estaremos haciendo una *estimación puntual.*

- Otra opción es, a partir de los datos muestrales, calcular dos valores entre los cuales se considera que, con cierta probabilidad, se encuentra el valor de parámetro poblacional. Dicho procedimiento se denomina *estimación por intervalos de confianza.*

8.2 Estimación puntual

Un estimador puntual de un parámetro poblacional θ es un estadístico A que depende de los n valores de la variable aleatoria, $X_1, X_2, \ldots, X_n$, de nuestra muestra. Una *estimación puntual* es el valor concreto $\hat{\theta}$ que toma el estadístico anterior para la muestra dada.

En el capítulo anterior vimos algunos de los principales estimadores puntuales utilizados para los parámetros poblacionales de las principales distribuciones de probabilidad. Aquí vamos a repasar los de las principales distribuciones de probabilidad:

- *Distribución normal*: una distribución normal se caracteriza por dos parámetros poblacionales, la media μ y la varianza σ^2. Como estimadores puntuales de dichos parámetros poblacionales se usan normalmente la media aritmética $\bar{X}$ para la media poblacional μ y la cuasivarianza muestral $\mathcal{S}^2$ para la varianza poblacional σ^2. Como ya demostramos en el capítulo anterior, ambos son estimadores insesgados, dado que $E(\bar{X}) = \mu$ y $E(\mathcal{S}^2) = \sigma^2$, y además se puede demostrar que ambos tienen una eficiencia máxima.

- *Distribución binomial*: una distribución binomial tiene como único parámetro poblacional la probabilidad de éxito p. Hemos visto que un estimador puntual para dicho parámetro poblacional es la proporción de éxitos P, definida como el cociente entre el número de éxitos y el tamaño muestral. Como ya vimos anteriormente, es un estimador insesgado dado que $E(P) = p$, además se puede demostrar que es de eficiencia máxima dado que su varianza $\sigma_P^2 = \dfrac{p(1-p)}{n}$ es mínima.

- *Distribución de Poisson*: una distribución de Poisson se caracteriza por su parámetro poblacional λ que representa el número medio de sucesos por intervalo seleccionado. Un estimador puntual de dicho parámetro poblacional será la media aritmética $\bar{X}$ de la muestra. Además, ya comprobamos en el capítulo anterior que es un estimador insesgado, ya que $E(\bar{X}) = \lambda$ y tiene la varianza mínima, por lo que también es el de máxima eficiencia.

8.3 Estimación por intervalos de confianza

8.3.1 Definición de intervalo de confianza

Una estimación puntual $\hat{\theta}$ de un parámetro poblacional θ no proporciona un valor exacto de dicho parámetro, ya que depende de los valores $X_1, X_2, \ldots, X_n$ de la variable aleatoria para la muestra que tengamos. Si variamos la muestra, la estimación puntual del parámetro variará, lo cual significa que no es posible conseguir un valor exacto para el parámetro poblacional. Las estimaciones puntuales que obtenemos $\hat{\theta}_1$, $\hat{\theta}_2$, ... para las distintas muestras siguen una distribución de densidad de probabilidad. Por ejemplo, en la figura 8.1 hemos representado la densidad de probabilidad de la media muestral $\bar{X}$. En dicha figura se puede ver que la densidad de probabilidad está centrada en torno al valor de la media poblacional μ. Además se observa que si señalamos el intervalo del 90% de probabilidad centrado en la media poblacional y tenemos 10 estimaciones muestrales, 9 se encontrarán en dicha región y 1 se encontrará fuera.

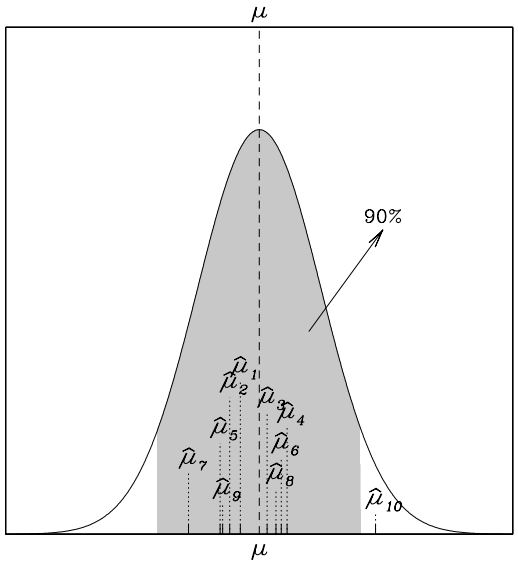


Figura 8.1: *Distribución de densidad de probabilidad para la media muestral $\bar{X}$. Se han representado los valores de 10 estimaciones muestrales, $\hat{\mu}_1$, $\hat{\mu}_2$, ..., $\hat{\mu}_{10}$ y la región del 90% de probabilidad en torno a la media poblacional μ.*

Por este motivo, además de una estimación puntual del parámetro es útil conocer la incertidumbre de dicha estimación para saber lo lejos o cerca que se encuentra nuestra estimación del verdadero parámetro poblacional. Este procedimiento se denomina *estimación por intervalos de confianza*. Siguiendo dicho procedimiento lo que se hace es calcular un intervalo $IC_{(1-\alpha)\cdot 100\%} = [L_{min}, L_{max}]$ en el cual se puede establecer que se encuentra el parámetro poblacional θ con una probabilidad $(1-\alpha)$ dada. Los límites L_{min} y L_{max} del intervalo $IC_{(1-\alpha)\cdot 100\%}$ se denominan límites de confianza y se obtendrán a partir de los valores $X_1, X_2, \ldots, X_n$ que toma la variable aleatoria en la muestra. Al valor $(1-\alpha)$ se le llama nivel de confianza y al intevalo $IC_{(1-\alpha)\cdot 100\%}$ se le denomina intervalo de confianza al $(1-\alpha)\cdot 100\%$.

Según lo anterior, podemos interpretar los intervalos de confianza de dos formas:

- Si θ es el parámetro poblacional que queremos estimar mediante un estimador A dado, un intervalo de confianza $IC_{(1-\alpha)\cdot 100\%} = [L_{min}, L_{max}]$ nos informa de que existe una cierta probabilidad $(1-\alpha)$ de que el verdadero valor de θ se encuentre en dicho intervalo $IC_{(1-\alpha)\cdot 100\%}$, es decir $p(L_{min} < \theta < L_{max}) = 1-\alpha$.

- También podemos interpretar los intervalos de confianza sabiendo que, una vez tengamos una muestra y calculemos el intervalo de confianza $IC_{(1-\alpha)\cdot 100\%}$, no podemos asegurar que el verdadero valor θ del parámetro poblacional se encuentre en dicho intervalo. Lo que sí podemos decir es que un $(1-\alpha)\cdot 100\%$ de los intervalos obtenidos a partir de muestras con las mismas características (el mismo tamaño muestral) contienen el parámetro poblacional θ y un $\alpha \cdot 100\%$ de dichos intervalos no lo contendrán.
Por ejemplo, en la Figura 8.2 se han representado 10 intervalos de confianza al 90% correspondientes a 10 muestras para la estimación de la media poblacional μ. Dichos intervalos están centrados en cada una de las estimaciones puntuales de la media, $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{10}$. Como tenemos 10 muestras y el nivel de confianza es del 90%, 9 de cada 10 intervalos de confianza muestrales contendrán la media poblacional μ y 1 de cada 10 intervalos no contendrá la media poblacional.

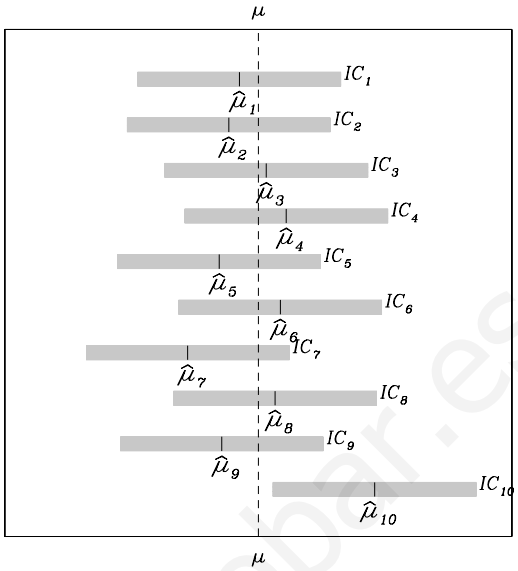


Figura 8.2: *Intervalos de confianza, IC_1, IC_2, ..., IC_{10}, al 90% de 10 muestras para la estimación de la media poblacional μ. Los intervalos de confianza están centrados en las estimaciones puntuales de las muestras, $\hat{\mu}_1$, $\hat{\mu}_2$, ..., $\hat{\mu}_{10}$.*

Observaciones:

- Al aumentar el tamaño de la muestra, la precisión con que conocemos el parámetro poblacional aumentará y, por lo tanto, la longitud del intervalo de confianza para un nivel de confianza dado disminuirá.

- Si aumentamos el nivel de confianza, la longitud del intervalo aumentará, por lo que tendremos menos precisión en el parámetro poblacional.

8.3.2 Procedimiento para el cálculo de los intervalos de confianza

Vamos a describir ahora cómo construir los intervalos de confianza para un parámetro poblacional. Supongamos que queremos estimar un parámetro poblacional θ utilizando un estimador

centrado A, por lo tanto $E(A) = \theta$, y supongamos que tenemos una estimación $\hat{\theta}$ obtenida a partir una muestra. Si consideramos un nivel de confianza al $(1-\alpha)\cdot 100\%$, eso quiere decir que dada la función de densidad de probabilidad de la variable muestral A y tomando un intervalo de probabilidad $IP_{(1-\alpha)\cdot 100\%} = (\theta - \delta_1, \theta + \delta_2)$ centrado en $E(A) = \theta$ que contiene el $(1-\alpha)\cdot 100\%$ de probabilidad, la probabilidad de que $\hat{\theta}$ esté en ese intervalo es $1-\alpha$, esto es $p\left(\theta - \delta_1 < \hat{\theta} < \theta + \delta_2\right) = 1 - \alpha$. A partir de aquí, es fácil demostrar (mediante el cambio de variable $A \longrightarrow \theta - A + \hat{\theta}$) que la expresión $p\left(\theta - \delta_1 < \hat{\theta} < \theta + \delta_2\right) = 1 - \alpha$ equivale a $p\left(\hat{\theta} - \delta_2 < \theta < \hat{\theta} + \delta_1\right) = 1 - \alpha$. Además de esto, conviene saber que generalmente se utilizan intervalos de probabilidad centrados, esto quiere decir que $p\left(\hat{\theta} < \theta - \delta_1\right) = \dfrac{\alpha}{2}$ y $p\left(\hat{\theta} > \theta + \delta_2\right) = \dfrac{\alpha}{2}$.

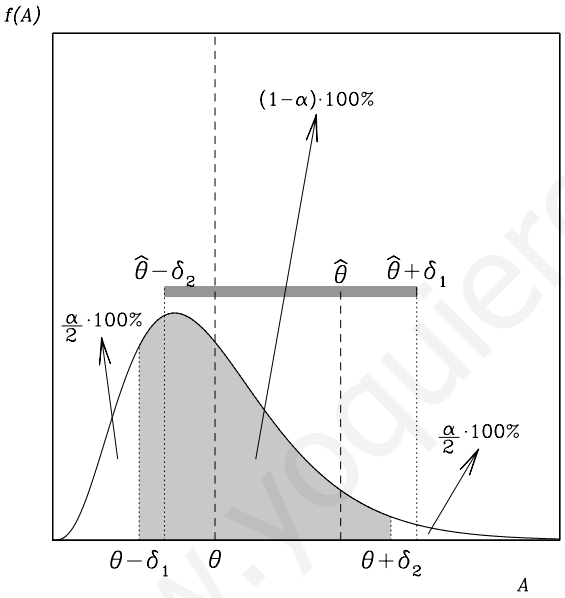


Figura 8.3: *Función de densidad de probabilidad $f(A)$ de un estimador A de un parámetro poblacional θ. En la misma figura se representa la región entre $\theta - \delta_1$ y $\theta + \delta_2$ de $f(A)$ que engloba un $(1-\alpha)\cdot 100\%$ de probabilidad, así como una estimación $\hat{\theta}$ del parámetro junto con su intervalo de confianza al $(1-\alpha)\cdot 100\%$, dado por $IC_{(1-\alpha)\cdot 100\%} = \left(\hat{\theta} - \delta_2, \hat{\theta} + \delta_1\right)$*

En la figura 8.3 se puede ver un ejemplo de una función de densidad $f(A)$ de un parámetro muestral A junto con la región que abarca el $(1-\alpha)\cdot 100\%$ de probabilidad, entre $\theta - \delta_1$ y $\theta + \delta_2$, por lo que $IP_{(1-\alpha)\cdot 100\%} = (\theta - \delta_1, \theta + \delta_2)$. En la misma gráfica se ha representado una estimación muestral $\hat{\theta}$ del parámetro poblacional θ, sabiendo que un $(1-\alpha)\cdot 100\%$ de las estimaciones muestrales se encontrarán en el intervalo $IP_{(1-\alpha)\cdot 100\%} = (\theta - \delta_1, \theta + \delta_2)$. Por este motivo, si consideramos el correspondiente intervalo de confianza al $(1-\alpha)\cdot 100\%$ de nivel de confianza, asociado a la estimación muestral $\hat{\theta}$, tendremos $IC_{(1-\alpha)\cdot 100\%} = \left(\hat{\theta} - \delta_2, \hat{\theta} + \delta_1\right)$, y sabemos que para un $(1-\alpha)\cdot 100\%$ de las estimaciones muestrales, dicho intervalo de confianza contendrá el parámetro poblacional θ.

Observaciones:

- Si la función de densidad de probabilidad de la variable muestral, $f(A)$, es simétrica, entonces, el intervalo de confianza en torno a la estimación muestral $\hat{\theta}$ también será simétrico, por lo que tendremos que $\delta_1 = \delta_2 = \delta$ y, por lo tanto, el intervalo de confianza será de la forma $IC = \left(\hat{\theta} - \delta, \hat{\theta} + \delta\right)$
- Si tenemos información sobre la función de distribución $F(A)$ de la variable muestral y definimos A_α tal que $p\left(A > A_\alpha\right) = 1 - F(A_\alpha) = \alpha$, el intervalo de confianza al $(1-\alpha)\cdot 100\%$ de nivel de confianza para la estimación muestral $\hat{\theta}$ será $IC_{(1-\alpha)\cdot 100\%} = \left(\hat{\theta} - A_{1-\alpha/2}, \hat{\theta} + A_{\alpha/2}\right)$

8.4 Intervalos de confianza para los parámetros de una población

8.4.1 Intervalo de confianza para la media poblacional μ en una población normal

Supongamos que tenemos una población que sigue una distribución normal $N(\mu, \sigma)$ y que consideramos la media aritmética $\bar{X}$ como estimador de la media poblacional μ. Vamos a analizar cómo construir los intervalos de confianza para la media poblacional teniendo una estimación puntual de la media $\hat{\mu}$ obtenida a partir de una muestra. Distinguiremos varios casos:

(a) Varianza poblacional σ^2 conocida:

Si nuestra población sigue una distribución normal $N(\mu, \sigma)$ y conocemos la varianza poblacional σ^2, sabemos que la media muestral $\bar{X}$ seguirá una distribución normal de media $E(\bar{X}) = \mu$ y de varianza $Var(\bar{X}) = \sigma^2_{\bar{X}} = {}^{\sigma^2}/_n$ donde n es el tamaño muestral. Esto quiere decir que $\bar{X} \sim N\left(\mu, {}^{\sigma}/_{\sqrt{n}}\right)$.

Un intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ de probabilidad para la media muestral $\bar{X}$ lo obtendremos tipificando la variable aleatoria $\bar{X}$. De este modo, para la variable tipificada $Z = \dfrac{\bar{X} - \mu}{{}^{\sigma}/_{\sqrt{n}}} \sim N(0,1)$, el intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ será $IP_{(1-\alpha)\cdot 100\%}(Z) = \left(-z_{\alpha/2}, z_{\alpha/2}\right)$, donde $z_{\alpha/2}$ está definido tal que $p\left(Z > z_{\alpha/2}\right) = {}^{\alpha}/_2$. Por lo tanto, para la media muestral $\bar{X}$ el intervalo de probabilidad será $IP_{(1-\alpha)\cdot 100\%}(\bar{X}) = \left(\mu - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$

Así pues, tendremos que si $\hat{\mu}$ es una estimación puntual de μ se cumplirá que:
$p\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \hat{\mu} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$

A partir de lo anterior, podemos obtener fácilmente el intervalo de confianza para la media poblacional μ de una población normal, para una determinada estimación muestral $\hat{\mu}$, obteniéndose que:

$$IC_{(1-\alpha)\cdot 100\%}(\mu) = \left(\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Ejemplo 1:

Calcular el intervalo de confianza para la media de una distribución normal de varianza conocida $\sigma^2 = 16$, si tenemos una muestra de 9 valores dada por $\{4, 13, 8, 12, 8, 15, 14, 7, 8\}$. Usar un nivel de confianza del 95%

Con los datos del enunciado sabemos que la media muestral $\bar{X}$ sigue una distribución normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(\mu, \frac{4}{3}\right)$

Ahora calculamos una estimación puntual para la media utilizando el estimador $\bar{X}$, obteniéndose como estimación puntual $\hat{\mu} = 9.89$. Como queremos utilizar un nivel del confianza de $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$. Sabemos, por lo tanto, que el intervalo de confianza para la media poblacional μ buscado será $IC_{95\%}(\mu) = \left(9.89 - z_{0.025}\frac{4}{3}, 9.89 + z_{0.025}\frac{4}{3}\right)$

Como $z_{0.025} = 1.96$, tendremos que $IC_{95\%}(\mu) = \left(9.89 - 1.96 \cdot \frac{4}{3}, 9.89 + 1.96 \cdot \frac{4}{3}\right)$

$$IC_{95\%}(\mu) = (7.276, 12.502)$$

(b) Varianza poblacional σ^2 desconocida:

Supongamos ahora que sabemos que la población sigue una distribución normal $N(\mu, \sigma)$, pero no conocemos el valor de la varianza poblacional σ^2. Por lo tanto, en primer lugar debemos estimar dicha varianza. Para ello utilizaremos como estimador de σ^2 la cuasivarianza muestral $\mathcal{S}^2$, al tratarse de un estimador centrado.

En este caso, tal y como vimos en el capítulo anterior, sabemos que la variable aleatoria $T = \frac{\bar{X} - \mu}{\mathcal{S}/\sqrt{n}}$ sigue una distribución t de Student con $n - 1$ grados de libertad, siendo n el tamaño muestral, o sea, $T \sim t_{n-1}$. Por lo tanto, un intervalo de probabilidad al $(1 - \alpha) \cdot 100\%$

para la variable T será de la forma $IP_{(1-\alpha)\cdot 100\%}(T) = \left(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}\right)$ siendo $t_{\alpha/2,n-1}$ tal que $p\left(T > t_{\alpha/2,n-1}\right) = \alpha/2$.
A partir de aquí podemos escribir el intervalo de probabilidad para la variable $\bar{X}$, siendo $IP_{(1-\alpha)\cdot 100\%}(\bar{X}) = \left(\mu - t_{\alpha/2,n-1}\dfrac{\mathcal{S}}{\sqrt{n}}, \mu + t_{\alpha/2,n-1}\dfrac{\mathcal{S}}{\sqrt{n}}\right)$

Por lo tanto, el intervalo de confianza para la media poblacional μ de una población normal, en el caso de que la varianza sea desconocida y tengamos una estimación puntual $\hat{\mu}$, y una estimación puntual $\hat{\sigma}^2$ para la varianza obtenida a partir de la cuasivarianza $\mathcal{S}^2$, será:

$$IC_{(1-\alpha)\cdot 100\%}(\mu) = \left(\hat{\mu} - t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

En el caso particular de que el tamaño muestral sea grande ($n > 30$) podemos aproximar la distribución t de Student por una normal $N(0,1)$, por lo que la variable aleatoria $\bar{X}$ seguirá una distribución normal de media μ y de varianza $\dfrac{\mathcal{S}^2}{n}$, así que $\bar{X} \sim N(\mu, \mathcal{S}/\sqrt{n})$ y, por lo tanto, el intervalo de confianza para la media poblacional se podrá escribir como:

$$IC_{(1-\alpha)\cdot 100\%}(\mu) = \left(\hat{\mu} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

siendo $z_{\alpha/2}$ tal que $p\left(Z > z_{\alpha/2}\right) = \alpha/2$, para $Z \sim N(0,1)$.

Ejemplo 2:

Calcular el intervalo de confianza del ejemplo 1 del apartado anterior suponiendo que la varianza es desconocida

Anteriormente ya habíamos obtenido una estimación puntual para la media utilizando el estimador $\bar{X}$, $\hat{\mu} = 9.89$.

Calculamos ahora una estimación para la varianza poblacional utilizando como estimador la cuasivarianza muestral $\mathcal{S}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2$. El resultado que obtenemos para una estimación de la varianza poblacional es $\hat{\sigma}^2 = 13.8611$.

Así pues, la variable $T = \dfrac{\bar{X} - \mu}{\mathcal{S}/\sqrt{n}}$ seguirá una distribución t de Student con 8 grados de libertad, por lo que el intervalo de confianza para la media poblacional μ buscado será:

$$IC_{95\%}(\mu) = \left(9.89 - t_{0.025,8}\frac{\sqrt{13.8611}}{3}, 9.89 + t_{0.025,8}\frac{\sqrt{13.8611}}{3}\right)$$

Como $t_{0.025,8} = 2.306$, tendremos que $IC_{95\%}(\mu) = \left(9.89 - 2.306 \cdot \frac{3.723}{3}, 9.89 + 2.306 \cdot \frac{3.723}{3}\right) \Rightarrow$

$$IC_{95\%}(\mu) = (7.028, 12.752)$$

8.4.2 Intervalo de confianza para la proporción p de una distribución binomial

Supongamos que tenemos una población que sigue una distribución binomial con parámetro p desconocido. Según vimos en el capítulo anterior, la proporción de éxitos en una muestra, P, es un buen estimador del parámetro poblacional p. Además sabemos que si la muestra es grande se puede aproximar por una distribución normal $P \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Dicha aproximación será válida cuando $np > 5$ y $n(1-p) > 5$.

Por lo tanto, el intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ para la variable P será:

$$IP_{(1-\alpha)\cdot 100\%}(P) = \left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

Dado que no conocemos p, para poder calcular el intervalo de confianza para p a partir de una estimación muestral $\hat{p}$, tendremos que tener en cuenta que nuestro intervalo de probabilidad implica que:

$$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow |\hat{p} - p| < z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Rightarrow$$

$$\Rightarrow (\hat{p} - p)^2 < z_{\alpha/2}^2 \frac{p(1-p)}{n} \Rightarrow \left(1 + \frac{z_{\alpha/2}^2}{n}\right) p^2 - \left(2\hat{p} + \frac{z_{\alpha/2}^2}{n}\right) p + \hat{p}^2 < 0 \Rightarrow$$

$$\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}} < p < \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

Para el caso en que tengamos una muestra grande (en la práctica cuando $n > 30$) podremos despreciar los términos $\frac{z_{\alpha/2}^2}{2n}$ y $\frac{z_{\alpha/2}^2}{4n^2}$, por lo que nuestro intervalo de confianza al $(1-\alpha)\cdot 100\%$ de nivel de confianza para el parámetro poblacional p obtenido a partir de una estimación muestral $\hat{p}$ será:

$$IC_{(1-\alpha)\cdot 100\%}(p) = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Se puede ver que si n es grande el intervalo de confianza es el que obtendríamos suponiendo que la variable aleatoria P sigue una distribución normal de media p y de varianza $\frac{\hat{p}(1-\hat{p})}{n}$,

estimada a partir de la estimación muestral $\hat{p}$.

Ejemplo 3:

Un jugador de baloncesto lanza 100 tiros libres y anota 85. Calcular un intervalo de confianza para la probabilidad de aciertos, considerando un nivel de confianza de 0.95.

Podemos estimar la probabilidad de aciertos utilizando el estadístico P que nos da la proporción de aciertos. En este caso, la estimación puntual que tendremos será $\hat{p} = 0.85$. A partir de aquí, dado que el tamaño muestral es grande ($n = 100 > 30$), tendremos que el intervalo de confianza pedido será $IC_{95\%}(p) = \left(\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$, por lo que tendremos:

$$IC_{95\%}(p) = \left(0.85 - 1.96\sqrt{\frac{0.85 \cdot 0.15}{100}}, 0.95 + 1.96\sqrt{\frac{0.85 \cdot 0.15}{100}}\right) \Rightarrow$$

$$IC_{95\%}(p) = (0.78, 0.92)$$

8.4.3 Intervalo de confianza para el parámetro λ de una distribución de Poisson

Si tenemos una distribución de Poisson, ya hemos visto que un estimador puntual para el parámetro λ es la media muestral $\bar{X}$. En el caso de que tengamos una muestra grande, de modo que se pueda aproximar a una distribución normal ($\lambda > 5$), tendremos que un intervalo de probabilidad del $(1-\alpha)\cdot 100\%$ para la media muestral será:

$$IP_{(1-\alpha)\cdot 100\%}(\bar{X}) = \left(\lambda - z_{\alpha/2}\sqrt{\frac{\lambda}{n}}, \lambda + z_{\alpha/2}\sqrt{\frac{\lambda}{n}}\right)$$

Por lo tanto, si $\hat{\lambda}$ es una estimación puntual de λ y tenemos una muestra grande, el intervalo de confianza para λ será:

$$IC_{(1-\alpha)\cdot 100\%}(\lambda) = \left(\hat{\lambda} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}\right)$$

8.4.4 Intervalo de confianza para la varianza σ^2 de una población normal

Supongamos que tenemos una variable aleatoria que sabemos que sigue una distribución normal $X \sim N(\mu, \sigma)$. Anteriormente hemos visto como obtener un intervalo de confianza para la media

poblacional μ y ahora vamos a hacer lo mismo para la varianza poblacional σ^2. Sabemos que un estimador centrado de la varianza poblacional es la cuasivarianza muestral $\mathcal{S}^2$ y, además, vimos en capítulos anteriores que si definimos una variable aleatoria $\theta = (n-1)\frac{\mathcal{S}^2}{\sigma^2}$, dicha variable aleatoria seguirá una distribución chi-cuadrado con $n-1$ grados de libertad, o sea $\theta \sim \chi^2_{n-1}$. Por lo tanto, si definimos $\chi^2_{\alpha/2,n-1}$ tal que $p(\theta > \chi^2_{\alpha/2,n-1}) = \frac{\alpha}{2}$, tendremos que un intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ de probabilidad para la variable aleatoria θ será $IP_{(1-\alpha)\cdot 100\%}\left(\theta = (n-1)\frac{\mathcal{S}^2}{\sigma^2}\right) = \left(\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}\right)$

Por lo que $p\left(\chi^2_{1-\alpha/2,n-1} < (n-1)\frac{\mathcal{S}^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$

Así pues, si tenemos una estimación muestral $\hat{\sigma}^2$ para la varianza poblacional σ^2 obtenida a partir del estadístico $\mathcal{S}^2$, podremos construir un intervalo de confianza al $(1-\alpha)\cdot 100\%$ de nivel de confianza partiendo del intervalo anterior, y obtendremos que:

$$IC_{(1-\alpha)\cdot 100\%}\left(\sigma^2\right) = \left(\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

Ejemplo 4:

Calcular un intervalo de confianza para la varianza para los datos del ejemplo 1, suponiendo que la varianza es desconocida.

Lo primero que hacemos es calcular una estimación de la varianza poblacional utilizando la cuasivarianza muestral $\mathcal{S}^2$, tal y como hicimos en el ejemplo 2, obteniéndose $\hat{\sigma}^2 = 13.8611$. Como tenemos una muestra de tamaño $n = 9$, el intervalo de confianza será $IC_{95\%}(\sigma^2) = \left(\frac{(n-1)\hat{\sigma}^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)\hat{\sigma}^2}{\chi^2_{0.975,n-1}}\right) \Rightarrow IC_{95\%}(\sigma^2) = \left(\frac{8\cdot 13.8611}{\chi^2_{0.025,8}}, \frac{8\cdot 13.8611}{\chi^2_{0.975,8}}\right)$.

Como $\chi^2_{0.025,8} = 17.535$ y $\chi^2_{0.975,8} = 2.18$ tendremos que $IC_{95\%}(\sigma^2) = \left(\frac{8\cdot 13.8611}{17.535}, \frac{8\cdot 13.8611}{2.18}\right) \Rightarrow$

$$IC_{95\%}\left(\sigma^2\right) = (6.324, 50.866)$$

8.4.5 Intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$

Supongamos que tenemos dos variables aleatorias independientes X_1 y X_2 que siguen distribuciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, respectivamente. Vamos a estudiar como determinar un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$. Sabemos que los estadísticos muestrales que nos permiten estimar las medias poblacionales son las medias muestrales $\bar{X}_1$ y $\bar{X}_2$ que

seguirán distribuciones normales de forma que $\bar{X}_1 \sim N\left(\mu_1, {}^{\sigma_1}/_{\sqrt{n_1}}\right)$ y $\bar{X}_2 \sim N\left(\mu_2, {}^{\sigma_2}/_{\sqrt{n_2}}\right)$, siendo n_1 y n_2 los correspondientes tamaños muestrales. Dado que se trata de dos distribuciones normales, la diferencia de estas variables aleatorias independientes también seguirá una distribución normal de media $\mu_1 - \mu_2$ y de varianza $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$, o sea, $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\right)$.

Sabiendo esto vamos a ver como construir los intervalos de confianza para $\mu_1 - \mu_2$.

(a) Varianzas poblaciones σ_1^2 y σ_2^2 conocidas:

Como sabemos que la diferencia de las medias muestrales sigue una distribución normal, un intervalo de probabilidad para la diferencia de medias al $(1-\alpha)\cdot 100\%$ de probabilidad tendremos que $IP_{(1-\alpha)\cdot 100\%}\left(\bar{X}_1 - \bar{X}_2\right) = \left(\mu_1 - \mu_2 - z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}, \mu_1 - \mu_2 + z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\right)$

Por lo tanto, si $\hat{\mu}_1$ y $\hat{\mu}_2$ son estimaciones puntuales de μ_1 y μ_2 podremos construir un intervalo de confianza para la diferencia de medias a partir del intervalo de probabilidad anterior y las estimaciones puntuales. De modo que:

$$IC_{(1-\alpha)\cdot 100\%}\left(\mu_1 - \mu_2\right) = \left(\hat{\mu}_1 - \hat{\mu}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \hat{\mu}_1 - \hat{\mu}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Ejemplo 5:

Consideremos la población descrita en el ejemplo 1 y supongamos que tenemos otra muestra dada por $\{17, 14, 2, 12, 12, 6, 5, 11, 5\}$ con varianza $\sigma^2 = 16$. Calcular un intervalo de confianza para la diferencia de las medias de ambas poblaciones.

Para la primera de las poblaciones calculamos la media en el ejemplo 1, obteniéndose como estimación para la media de dicha población $\hat{\mu}_1 = 9.89$.

Para la segunda población calculamos también una estimación puntual de la media poblacional, utilizando el estadístico media muestral, obteniéndose $\hat{\mu}_2 = 9.33$.

Como $\sigma_1^2 = \sigma_2^2 = 16$ y $n_1 = n_2 = 9$, el intervalo de confianza para la diferencia de medias será $IC_{95\%}\left(\mu_1 - \mu_2\right) = \left(9.89 - 9.33 - z_{0.025}\sqrt{\dfrac{16}{9} + \dfrac{16}{9}}, 9.89 - 9.33 + z_{0.025}\sqrt{\dfrac{16}{9} + \dfrac{16}{9}}\right)$, como

$z_{0.025} = 1.96$, tendremos que $IC_{95\%}(\mu_1 - \mu_2) = \left(0.56 - 1.96\frac{4}{3}\sqrt{2}, 0.56 + 1.96\frac{4}{3}\sqrt{2}\right) \Rightarrow$

$$IC_{95\%}(\mu_1 - \mu_2) = (-3.1, 4.3)$$

(b) Varianzas poblaciones σ_1^2 y σ_2^2 desconocidas y se puede suponer que $\sigma_1^2 \simeq \sigma_2^2$:

Como $\bar{X}_1$ y $\bar{X}_2$ siguen distribuciones normales, su diferencia también seguirá una distribución normal de media $\mu_1 - \mu_2$ y cuya varianza desconocemos, pero la podemos estimar en el caso de que se pueda suponer que las dos poblaciones tienen varianzas similares ($\sigma_1^2 \simeq \sigma_2^2$). En primer lugar estimaremos las varianzas poblacionales de las muestras utilizando la cuasivarianza común, cumpliéndose que $\hat{\sigma}^2 = \hat{\sigma}_1^2 = \hat{\sigma}_2^2$. La cuasivarianza común se calculará como $\mathcal{S}^2 = \dfrac{(n_1-1)\mathcal{S}_1^2 + (n_2-1)\mathcal{S}_2^2}{n_1+n_2-2}$, siendo $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$ las cuasivarianzas muestrales de ambas muestras, dando lugar a una estimación $\hat{\sigma}^2$ para la varianza de X_1 y de X_2. Por lo que la estimación para la varianza poblacional de $\bar{X}_1 - \bar{X}_2$ será $\hat{\sigma}^2_{\bar{X}_1-\bar{X}_2} = \dfrac{\hat{\sigma}^2}{n_1} + \dfrac{\hat{\sigma}^2}{n_2} = \hat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$

En este caso, la variable $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ seguirá una distribución t de Student de $n_1 + n_2 - 2$ grados de libertad. Por lo tanto, un intervalo del $(1-\alpha)\cdot 100\%$ de probabilidad para la variable aleatoria de la diferencia de medias muestrales será:

$$IP_{(1-\alpha)\cdot 100\%}(\bar{X}_1 - \bar{X}_2) = \left(\mu_1 - \mu_2 - t_{\alpha/2,n_1+n_2-2}\,\mathcal{S}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \mu_1 - \mu_2 + t_{\alpha/2,n_1+n_2-2}\,\mathcal{S}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

A partir de dicho intervalo, si $\hat{\mu}_1$ y $\hat{\mu}_2$ son estimaciones muestrales de las medias poblacionales, un intervalo de confianza con un nivel de confianza del $(1-\alpha)\cdot 100\%$ para la diferencia de las medias poblaciones es:

$$IC_{(1-\alpha)\cdot 100\%}(\mu_1 - \mu_2) = \left(\hat{\mu}_1 - \hat{\mu}_2 - t_{\alpha/2,n_1+n_2-2}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \hat{\mu}_1 - \hat{\mu}_2 + t_{\alpha/2,n_1+n_2-2}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

con $\hat{\sigma}^2$ estimado utilizando $\mathcal{S}^2 = \dfrac{(n_1-1)\mathcal{S}_1^2 + (n_2-1)\mathcal{S}_2^2}{n_1+n_2-2}$

Ejemplo 6:

Considerar el siguiente par de muestras y, sabiendo que provienen de distribuciones normales con varianzas similares, establecer un intervalo con un nivel del confianza de 0.95 para la diferencia de medias.

Muestra 1: $\{1.78, 0.52, 5.13, 3.86, 6.29, 2.51, 2.11, 7.66, 6.27, -4.57\}$

Muestra 2: $\{7.47, -0.80, -0.60, 0.03, 4.49, -0.14, -0.99, 0.74, 1.45, 5.38\}$

En primer lugar calculamos las estimaciones para las medias de las dos muestras, $\hat{\mu}_1$ y $\hat{\mu}_2$, utilizando el estimador $\bar{X}$ y las estimaciones de las varianzas, $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$, utilizando la cuasivarianza $\mathcal{S}^2$. Los resultados que se obtienen son $\hat{\mu}_1 = 3.52$ y $\hat{\sigma}_1^2 = 12.69$ y $\hat{\mu}_2 = 1.70$ y $\hat{\sigma}_2^2 = 8.95$

Por lo tanto la estimación para la varianza conjunta será $\hat{\sigma}^2 = \dfrac{9 \cdot 12.69 + 9 \cdot 8.95}{18} = 10.82 \Rightarrow$ $\Rightarrow \hat{\sigma} = 3.289$. Así pues, el intervalo de confianza pedido será:

$$IC_{95\%}\,(\mu_1 - \mu_2) = \left(3.52 - 1.70 - t_{0.025,18}\,3.289\sqrt{\frac{1}{10} + \frac{1}{10}}, 3.52 - 1.70 + t_{0.025,18}\,3.289\sqrt{\frac{1}{10} + \frac{1}{10}}\right).$$

Como $t_{0.025,18} = 2.101$, tendremos que:

$IC_{95\%}\,(\mu_1 - \mu_2) = (1.82 - 2.101 \cdot 1.471, 1.82 + 2.101 \cdot 1.471) \Rightarrow$

$$IC_{95\%}\,(\mu_1 - \mu_2) = (-1.27, 4.91)$$

(c) Varianzas poblaciones σ_1^2 y σ_2^2 desconocidas y $\sigma_1^2 \neq \sigma_2^2$:

Si las varianzas poblacionales no se conocen y no se pueden suponer iguales, se estiman cada una de ellas con las cuasivarianzas muestrales correspondientes, obteniéndose las estimaciones $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ para las varianzas. En este caso, la varianza de diferencia de medias se puede estimar se puede estimar con $\mathcal{S}^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\mathcal{S}_1^2}{n_1} + \dfrac{\mathcal{S}_2^2}{n_2}$ teniéndose que $\hat{\sigma}^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}$ por lo que la variable $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\mathcal{S}_1^2}{n_1} + \dfrac{\mathcal{S}_2^2}{n_2}}}$ seguirá una distribución t de Student con g grados de libertad, siendo g el número natural más próximo a $h = \dfrac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\hat{\sigma}_1^2/n_1\right)^2}{n_1+1} + \frac{\left(\hat{\sigma}_2^2/n_2\right)^2}{n_2+1}} - 2$. Por este motivo, un intervalo del $(1-\alpha)\cdot 100\%$ de probabilidad para la diferencia de medias es:

$$IP_{(1-\alpha)\cdot 100\%}\left(\bar{X}_1 - \bar{X}_2\right) = \left(\mu_1 - \mu_2 - t_{\alpha/2,g}\sqrt{\frac{\mathcal{S}_1^2}{n_1} + \frac{\mathcal{S}_1^2}{n_2}}, \mu_1 - \mu_2 + t_{\alpha/2,g}\sqrt{\frac{\mathcal{S}_1^2}{n_1} + \frac{\mathcal{S}_2^2}{n_2}}\right)$$

Por lo tanto un intervalo de confianza para la diferencia de las medias poblacionales, dadas las estimaciones puntuales $\hat{\mu}_1$ y $\hat{\mu}_2$, es:

$$IC_{(1-\alpha)\cdot 100\%}(\mu_1-\mu_2)=\left(\hat{\mu}_1-\hat{\mu}_2-t_{\alpha/2,g}\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_1^2}{n_2}},\hat{\mu}_1-\hat{\mu}_2+t_{\alpha/2,g}\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}\right)$$

Ejemplo 7:

Repetir el ejemplo 5 suponiendo que las varianzas de ambas poblaciones son desconocidas.

Para la primera de las poblaciones calculamos la media en el ejemplo 1, obteniéndose como estimación para la media de dicha población $\hat{\mu}_1=9.89$. Además calculamos una estimación para la varianza en el ejemplo 2, utilizando la cuasivarianza muestral, obteniéndose $\hat{\sigma}_1^2=13.8611$

Para la segunda población calculamos en el ejemplo anterior una estimación de la media $\hat{\mu}_2=9.33$. Además si estimamos la varianza poblacional para esta población, usando el estimador cuasivarianza muestral, obtenemos $\hat{\sigma}_2^2=25.0$

Como no está claro que las varianzas sean iguales, tendremos que el intervalo de confianza para la diferencia de medias es:

$IC_{95\%}(\mu_1-\mu_2)=\left(9.89-9.33-t_{0.025,g}\sqrt{\frac{13.8611}{9}+\frac{25}{9}},9.89-9.33+t_{0.025,g}\sqrt{\frac{13.8611}{9}+\frac{25}{9}}\right)$ y como $h=\frac{\left(\frac{13.8611}{9}+\frac{25}{9}\right)^2}{\frac{(13.8611/9)^2}{10}+\frac{(25/9)^2}{10}}-2=16.48$, tomaremos $g=16$ y como $t_{0.025,16}=2.12$, por lo tanto $IC_{95\%}(\mu_1-\mu_2)=(0.56-2.12\cdot 2.08,0.56+2.12\cdot 2.08)\Rightarrow$

$$IC_{95\%}(\mu_1-\mu_2)=(-3.85,4.97)$$

8.4.6 Intervalo de confianza para la diferencia de proporciones p_1-p_2

Consideremos ahora dos poblaciones que siguen distribuciones binomiales $Bin(n_1,p_1)$ y $Bin(n_2,p_2)$, respectivamente, cuyas probabilidades p_1 y p_2 son desconocidas. Si queremos hacer una estimación puntual de la diferencia de probabilidades, p_1-p_2, un buen estimador es la diferencia de proporciones P_1-P_2, donde P_1 es la proporción de éxitos de la muestra 1 y P_2 la proporción de éxitos de la muestra 2.

Suponiendo que se cumplen las condiciones de normalidad para ambas distribuciones, esto es, $n_1p_1>5, n_1(1-p_1)>5$ y $n_2p_2>5, n_2(1-p_2)>5$, sabemos que las distribuciones binomiales se puede aproximar por normales de medias p_1 y p_2, respectivamente, y varianzas $\frac{p_1(1-p_1)}{n_1}$

y $\dfrac{p_2(1-p_2)}{n_2}$, respectivamente, por lo que la diferencia de probabilidades seguirá también una distribución normal de media $p_1 - p_2$ y de varianza $\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$. Por lo tanto, un intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ para el estimador $P_1 - P_2$ será:

$$IP_{(1-\alpha)\cdot 100\%}(P_1-P_2) = \left(p_1-p_2 - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, p_1-p_2 + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

A partir de aquí, si tenemos dos estimaciones puntuales $\hat{p}_1$ y $\hat{p}_2$ para las probabilidades poblacionales, bajo el supuesto de tamaños muestrales grandes, se obtiene, como en casos anteriores, el intervalo de confianza para la diferencia de probabilidades poblacionales:

$$IC_{(1-\alpha)\cdot 100\%}(p_1-p_2) = \left(\hat{p}_1-\hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1-\hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

Ejemplo 8:

Considerar el jugador de baloncesto del ejemplo 3 y, sabiendo que un compañero suyo ha acertado 40 tiros libres de 50 intentos, calcular un intervalo para la diferencia de probabilidades de acierto entre estos dos jugadores.

Del primer jugador tenemos una estimación para su probabilidad de encestar dada por $\hat{p}_1 = 0.85$, mientras que para el segundo jugador, dicha estimación es $\hat{p}_2 = 0.8$. Con estos datos y dado que los tamaños muestrales son grandes (se satisface que $np > 5$ y $n(1-p) > 5$ para ambas muestras) tendremos que el intervalo de confianza con 0.95 de nivel de confianza será $IC_{95\%}(p_1 - p_2) = \left(0.85 - 0.8 - z_{0.025}\sqrt{\frac{0.85\cdot 0.15}{100} + \frac{0.8\cdot 0.2}{50}}, 0.85 - 0.8 + z_{0.025}\sqrt{\frac{0.85\cdot 0.15}{100} + \frac{0.8\cdot 0.2}{50}}\right)$, como $z_{0.025} = 1.96$ tendremos que $IC_{95\%}(p_1 - p_2) = (0.05 - 1.96\cdot 0.067, 0.05 + 1.96\cdot 0.067) \Rightarrow$

$$IC_{95\%}(p_1 - p_2) = (-0.08, 0.18)$$

8.4.7 Intervalo de confianza para el cociente de varianzas σ_1^2/σ_2^2

Supongamos que tenemos dos poblaciones normales de varianzas σ_1^2 y σ_2^2, respectivamente. Vamos a construir un intervalo de confianza para el cociente de varianzas, suponiendo que las muestras son de tamaños n_1 y n_2, respectivamente, y los estimadores puntuales para las varianzas son las cuasivarianzas muestrales $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$. Vimos en el capítulo anterior que la variable $F = \dfrac{\mathcal{S}_1^2/\sigma_1^2}{\mathcal{S}_2^2/\sigma_2^2}$ sigue una distribución F de Fisher con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador, $F_{(n_1-1,n_2-1)}$. Definiendo F_{α,n_1-1,n_2-1} tal que $p\left(F > F_{\alpha;n_1-1,n_2-1}\right) = \alpha$ y recordando que $F_{1-\alpha;n_1-1,n_2-1} = 1/F_{\alpha;n_2-1,n_1-1}$ tendremos que

un intervalo del $(1-\alpha)\cdot 100\%$ de probabilidad para el estimador puntual para el cociente de varianzas, $\frac{\mathcal{S}_1^2}{\mathcal{S}_2^2}$, será $IP_{(1-\alpha)\cdot 100\%}\left(\frac{\mathcal{S}_1^2}{\mathcal{S}_2^2}\right) = \left(\frac{\frac{\sigma_1^2}{\sigma_2^2}}{F_{\alpha/2;n_2-1,n_1-1}}, \frac{\sigma_1^2}{\sigma_2^2}F_{\alpha/2;n_1-1,n_2-1}\right)$

Por lo que, si $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ son estimaciones puntuales para las varianzas poblacionales (obtenidas a partir del estadísticos cuasivarianza muestral), tendremos que un intervalo de confianza para las varianzas poblacionales será:

$$IC_{(1-\alpha)\cdot 100\%}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left(\frac{\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}}{F_{\alpha/2;n_1-1,n_2-1}}, \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}F_{\alpha/2;n_2-1,n_1-1}\right)$$

Ejemplo 9:

Considerar las muestras del ejemplo 7 y calcular un intervalo de confianza para el cociente de varianzas, de modo que se pueda analizar si es razonable considerar que ambas varianzas son iguales o no.

Para estas muestras hemos obtenido las estimaciones de las varianzas, $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$, utilizando como estimador la cuasivarianza. Los resultados obtenidos son $\hat{\sigma}_1^2 = 13.8611$ y $\hat{\sigma}_2^2 = 25.0$. Como el tamaño muestral es $n_1 = n_2 = 9$, el intervalo de confianza buscado será:

$IC_{95\%}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left(\frac{13.8611/25.0}{F_{0.025;8,8}}, \frac{13.8611}{25.0}F_{0.025;8,8}\right)$

Como $F_{0.025;8,8} = 4.4333$ tendremos que $IC_{95\%}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = \left(\frac{0.554}{4.4333}, 0.554\cdot 4.4333\right) \Rightarrow$

$$IC_{95\%}\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = (0.125, 2.458)$$

Como el intervalo de confianza para el cociente de las varianzas poblacionales contiene al valor 1, sería razonable considerar que ambas varianzas son iguales $\frac{\sigma_1^2}{\sigma_2^2} = 1$

8.4.8 Intervalo de confianza para datos apareados

En algunos de los apartados anteriores hemos trabajado con dos poblaciones diferentes, pero suponiendo que ambas eran independientes. Sin embargo, no siempre es así, ya que en algunos casos las muestras que extraemos de las poblaciones son dependientes. Supongamos ahora que tenemos dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ de las cuales extraemos dos muestras,

dependientes una de la otra, del mismo tamaño muestral n. Este tipo de situaciones se suele referir a experimentos en los que medimos una determinada característica de una muestra y después de un determinado tratamiento de la muestra volvemos a medir la misma característica. A este tipo de experimentos se les denomina observaciones apareadas.

En este caso lo que nos interesa es construir un intervalo de confianza para la diferencia entre las muestras, de modo que si los valores muestrales que tenemos son $\{x_{1_i}\}_{i=1}^n$ para la población 1 y $\{x_{2_i}\}_{i=1}^n$ para la población 2, definiremos la diferencia entre los datos apareados, $d_i = x_{1_i} - x_{2_i}$ para $i = 1, 2, \ldots, n$. Así construiremos una variable aleatoria $D = X_1 - X_2$, que para una muestra suficientemente grande se puede considerar que sigue una distribución normal de media $\mu_D = \mu_1 - \mu_2$ y de varianza σ_D^2, que cumplirá que $\sigma_D^2 \neq \sigma_1^2 + \sigma_2^2$ ya que se trata de variables dependientes. Las estimaciones puntuales de estos parámetros poblacionales las calcularemos utilizando los estimadores $\bar{D} = \dfrac{1}{n}\sum_{i=1}^{n} d_i$ y $\mathcal{S}_D^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{D}\right)^2$, dando lugar a las estimaciones puntuales $\hat{\mu}_D$ y $\hat{\sigma}_D^2$.

Como la variable aleatoria $\bar{D}$ seguirá una distribución normal de media μ_D y de varianza σ_D^2/n, un intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ para la media de las diferencias $\bar{D}$ será $IP_{(1-\alpha)\cdot 100\%}(\bar{D}) = \left(\mu_D - t_{\alpha/2,n-1}\dfrac{\mathcal{S}_D}{\sqrt{n}}, \mu_D + t_{\alpha/2,n-1}\dfrac{\mathcal{S}_D}{\sqrt{n}}\right)$ que se aproxima por $IP_{(1-\alpha)\cdot 100\%}(\bar{D}) = \left(\mu_D - z_{\alpha/2}\dfrac{\mathcal{S}_D}{\sqrt{n}}, \mu_D + z_{\alpha/2}\dfrac{\mathcal{S}_D}{\sqrt{n}}\right)$ si la muestra es grande ($n > 30$).

Así pues, el intervalo de confianza para la media poblacional de las diferencias, μ_D, se podrá obtener en función de las estimaciones puntuales $\hat{\mu}_D$ y $\hat{\sigma}_D^2$, como:

$$IC_{(1-\alpha)\cdot 100\%}(\mu_D) = \left(\hat{\mu}_D - t_{\alpha/2,n-1}\frac{\hat{\sigma}_D}{\sqrt{n}}, \hat{\mu}_D + t_{\alpha/2,n-1}\frac{\hat{\sigma}_D}{\sqrt{n}}\right)$$

que se aproxima por:

$$IC_{(1-\alpha)\cdot 100\%}(\mu_D) = \left(\hat{\mu}_D - z_{\alpha/2}\frac{\hat{\sigma}_D}{\sqrt{n}}, \hat{\mu}_D + z_{\alpha/2}\frac{\hat{\sigma}_D}{\sqrt{n}}\right)$$

si la muestra es grande.

Ejemplo 10:

Considerar una muestra de 8 pacientes a los que se les administra un medicamento para disminuir la presión arterial. Se miden los siguientes valores de la presión arterial antes y después de administrar el medicamento:

Antes	122	110	95	97	156	120	141	92
Después	121	115	95	112	119	115	104	91

Calcular un intervalo de confianza para la media de la diferencia y comentar el resultado.

En primer lugar calculamos las diferencias individuales, $d_i = x_{d_i} - x_{a_i}$ y, a partir de ahí, una estimación de la media de las diferencias utilizado el estimador $\bar{D}$. Igualmente calculamos una estimación para σ_D^2 utilizando la cuasivarianza muestral.

Teniendo en cuenta que:

Antes	122	110	95	97	156	120	141	92
Después	121	115	95	112	119	115	104	91
Diferencia, d_i	-1	5	0	15	-37	-5	-37	-1

tendremos que $\hat{\mu}_D = -7.625$ y $\hat{\sigma}_D^2 = 364.27 \Rightarrow \hat{\sigma}_D = 19.086$. Por lo tanto, el intervalo de confianza pedido será $IC_{95\%}(\mu_D) = \left(-7.625 - t_{0.025,7}\frac{19.086}{\sqrt{8}}, -7.625 + t_{0.025,7}\frac{19.086}{\sqrt{8}}\right)$. Como $t_{0.025,7} = 2.365$ tendremos que:
$IC_{95\%}(\mu_D) = (-7.625 - 2.365 \cdot 6.748, -7.625 + 2.365 \cdot 6.748) \Rightarrow$

$$IC_{95\%}(\mu_D) = (-23.58, 8.33)$$

Analizando estos datos no podemos asegurar que el medicamento sirva para disminuir la presión arterial, dado que en el intervalo de confianza para la media de la diferencia están incluídos valores positivos, lo cual significaría que no sólo es posible que el medicamento no disminuya la presión arterial, sino que también podría aumentarla.

Capítulo 9

Hipótesis estadísticas I

Contrastes de hipótesis. Tipos de errores. Región crítica. Nivel de significación, valor P y potencia de un contraste. Contrastes relativos a proporciones, medias y varianzas. Tamaños muestrales en los contrastes.

9.1 Contrastes de hipótesis

9.1.1 Introducción

En muchos problemas de investigación se tiene que comprobar, con unos ciertos márgenes de error, si una hipótesis es correcta o no. En ciertos casos estas hipótesis están referidas a los valores de los parámetros de una o varias poblaciones, o a la comparación de una población con un modelo o con otra población. Por ejemplo, queremos saber si la media de una población tiene un determinado valor o no, o si la dispersión de una población es mayor o menor que la de otra, etc.

Un contraste de hipótesis estadístico es un procedimiento que nos permitirá aceptar o rechazar una hipótesis sobre una población, con un determinado error, utilizando los datos de una muestra aleatoria de dicha población. Si la hipótesis se formula sobre un determinado parámetro de la población se dice que el contraste es parámetrico, mientras que si los contrastes se refieren al tipo de distribución se denominan contrastes no paramétricos.

9.1.2 Formulación de un contraste de hipótesis

En la realización de un contraste de hipótesis necesitamos definir varias cosas:

1. *Hipótesis nula:* El primer paso para formular un contraste es establecer la hipótesis

estadística que se quiere aceptar o rechazar. Dicha hipótesis se denomina *hipótesis nula,* H_0, porque suele referirse que no hay diferencias entre el valor verdadero de lo que se quiere contrastar y el valor propuesto. Si θ es el parámetro poblacional que queremos analizar y θ_0 es el valor propuesto, la hipótesis nula será $H_0 : \theta = \theta_0$. Normalmente se suelen formular hipótesis con intención de rechazarlas, tratando de demostrar que las diferencias que se obtienen entre el valor real y el muestral no son debidas simplemente a la fluctuación estadística. Así, si queremos demostrar que una moneda está trucada, nuestra hipótesis nula será que la probabilidad de que salga, por ejemplo, cara es 0.5 (esto es, $H_0 : p = 0.5$) de modo que al realizar un muestreo y analizar el valor muestral obtenido $\hat{p}$ podamos decir si las diferencias obtenidas entre $\hat{p}$ y $p = 0.5$ son explicables dentro de las fluctuaciones estadísticas, en cuyo caso aceptamos la hipótesis nula, o no, rechazándose la hipótesis nula.

2. *Hipótesis alternativa:* Por otro lado, se establece una *hipótesis alternativa*, H_1, que se acepta cuando se rechaza la hipótesis nula. Dicha hipótesis alternativa puede ser:
(a) una negación en sentido estricto de la hipótesis nula, esto es $H_1 : \theta \neq \theta_0$, denominado *contraste bilateral*; o
(b) una negación en sentido amplio, bien porque creamos que el parámetro poblacional es mayor que el valor propuesto, esto es $H_1 : \theta > \theta_0$ (*contraste unilateral derecho*); o bien porque creamos que el parámetro poblacional es menor que el valor propuesto, $H_1 : \theta < \theta_0$ (*contraste unilateral izquierdo*).
En el ejemplo anterior de la moneda la hipótesis alternativa sería: (a) para un contraste bilateral, $H_1 : p \neq 0.5$, y (b) para un contraste unilateral, o bien $H_1 : p < 0.5$ si sospechamos que la probabilidad de salir cara es menor que la de salir cruz, o bien $H_1 : p > 0.5$ si sospechamos lo contrario.

3. *Estadístico de contraste:* Para realizar el contraste de hipótesis se necesita utilizar un *estadístico de prueba o de contraste* o *función de decisión* cuya distribución se supone conocida. En el ejemplo anterior de la moneda, usaríamos como estadístico de prueba la proporción de una distribución binomial, P, que sabemos que sigue, para tamaños muestrales grandes, una distribución normal, $P \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

4. *Nivel de significación:* Una vez definido el estadístico del contraste, se fija un *nivel de significación* α para el contraste, lo que permite obtener: (a) una *región de aceptación* para H_0, que corresponderá a un intervalo con probabilidad $(1-\alpha)$ para el estadístico de contraste; y (b) su correspondiente *región de rechazo*, que será la complementaria en la recta real y que determinará la región de probabilidad α para los valores menos probables del estadístico en el caso de que H_0 fuese cierta.

Una vez definido todo lo anterior, tomaremos una muestra de tamaño n de la población, haremos una estimación muestral del parámetro poblacional y, dependiendo de que la estimación muestral esté en la región de aceptación o en la región de rechazo, aceptaremos o rechazaremos la hipótesis nula.

Ejemplo 1:

Supongamos que queremos saber si una moneda está trucada. Vamos a analizar como realizar un contraste de hipótesis tanto en el supuesto de que la hipótesis alternativa fuese una oposición estricta a la hipótesis nula, como en el caso contrario de una oposición en un sentido amplio.

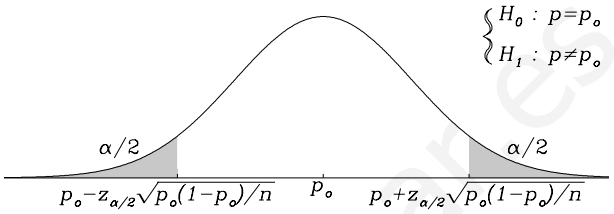


Figura 9.1: *Esquema de un contraste bilateral para la proporción p de una distribución binomial*

- *Contraste bilateral*: Supongamos que sospechamos que una moneda está trucada, pero no sabemos en qué sentido, si en el de que la probabilidad de cara sea mayor que la de cruz o viceversa. En este caso, el contraste de hipótesis se formularía de la siguiente manera, suponiendo que p fuese la probabilidad de que salga cara:

$$\begin{cases} H_0: & p = 0.5 \\ H_1: & p \neq 0.5 \end{cases}$$

Definimos ahora el estadístico del contraste, que en este caso será la proporción de éxitos de la muestra, $P = \dfrac{\#\text{éxitos}}{\#\text{ensayos}}$. Sabemos que dicho estadístico sigue una distribución normal, $P \sim N\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$. Además, si la hipótesis nula es cierta, dicha distribución será $P \sim N\left(p_0, \sqrt{\dfrac{p_0(1-p_0)}{n}}\right)$, con $p_0 = 0.5$.

Posteriormente, fijamos un nivel de significación α y, por lo tanto, la región de aceptación de H_0, RA_α, es un intervalo de probabilidad al $(1-\alpha)\cdot 100\%$ para P, esto es, un intervalo en el cual se encontrará P con probabilidad $(1-\alpha)$ en el caso de que la hipótesis nula

sea cierta. Dicho intervarlo será, según vimos en capítulo anterior:

$$RA_\alpha = IP_{(1-\alpha)\cdot 100\%}(P) = \left(p_0 - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

Por lo tanto, aceptaremos H_0 si nuestro valor muestral $\hat{p}$ está en dicho intervalo, $\hat{p} \in IP_{(1-\alpha)\cdot 100\%}(P) = RA_\alpha$.
La región de rechazo o región crítica será el intervalo complementario al anterior:

$$RC_\alpha = \left(-\infty, p_0 - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}\right] \cup \left[p_0 + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}, \infty\right)$$

de modo que rechazaremos la hipótesis nula H_0 y, por tanto, aceptaremos la alternativa H_1 si nuestro valor muestral $\hat{p}$ está en dicha región, $\hat{p} \in RC_\alpha$. Esto es debido a que si la hipótesis nula es cierta, la probabilidad de que el valor muestral $\hat{p}$ caiga en la región de rechazo RC_α es muy baja (sólo un $\alpha \cdot 100\%$) y, por lo tanto, concluiremos que si hemos obtenido un valor muestral en esa región es debido a que la hipótesis nula es falsa, y no a que se haya producido un suceso de tan baja probabilidad.

- *Contraste unilateral:* De nuevo, bajo la suposición de que la moneda está trucada, vamos a considerar que tenemos sospechas de que la probabilidad de que salga cara es mayor que la de que salga cruz. En este caso, si p es la probabilidad de que salga cara, el contraste se formularía de la siguiente manera:

$$\begin{cases} H_0: & p = 0.5 \\ H_1: & p > 0.5 \end{cases}$$

El estadístico de contraste será de nuevo la proporción de éxitos P y en este caso el

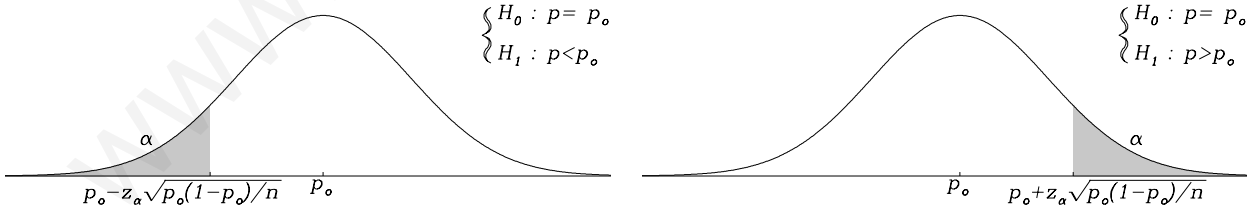


Figura 9.2: *Esquema de un contraste unilateral derecho (a la derecha) y un contraste unilateral izquierdo (a la izquierda) para la proporción p de una distribución binomial*

intervalo de probabilidad para P no estará centrado en $p_0 = 0.5$, sino que será asimétrico:

$$RA_\alpha = IP_{(1-\alpha)\cdot 100\%}(P) = \left(-\infty, p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

de modo que aceptaremos H_0 si el valor muestral $\hat{p}$ se encuentra en dicho intervalo o región de aceptación RA_α. En este caso, la región de rechazo que nos permitirá aceptar la hipótesis alternativa H_1 si el valor muestral $\hat{p}$ se encuentra en dicho intervalo quedará a la derecha (contraste unilateral derecho) y será:

$$RC_\alpha = \left[p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}, \infty\right)$$

Por el contrario, si tuviesemos sospechas de que la probabilidad de obtener cara es menor que la de obtener cruz, el contraste de hipótesis sería unilateral izquierdo y se enunciará de la siguiente manera:

$$\begin{cases} H_0: & p = 0.5 \\ H_1: & p < 0.5 \end{cases}$$

la región de aceptación de la hipótesis nula será, con $p_0 = 0.5$,

$$RA_\alpha = IP_{(1-\alpha)\cdot 100\%}(P) = \left(p_0 - z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}, \infty\right)$$

y la región de rechazo será:

$$RC_\alpha = \left(-\infty, p_0 - z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right]$$

Ejemplo 2:

Supongamos que en el ejemplo anterior hemos tirado la moneda 50 veces y hemos obtenido cara 35 veces.

El valor muestral que hemos obtenido para la probabilidad de cara será $\hat{p} = \dfrac{35}{50} = 0.7$. Este dato podría hacernos sospechar que la moneda está trucada y que la probabilidad de salir cara es mayor que la de salir cruz. Así pues, el contraste de hipótesis lo realizaremos de la siguiente forma:

$$\begin{cases} H_0: & p = 0.5 \\ H_1: & p > 0.5 \end{cases}$$

Eligiendo un nivel de significación $\alpha = 0.05$, tendremos que: la región de rechazo será $RC_{0.05} = \left[p_0 + z_{0.05}\sqrt{\frac{p_0(1-p_0)}{n}}, \infty\right)$. Como $p_0 = 0.5$, $n = 50$ y $z_{0.05} = 1.645$, tendremos que:

$$RC_{0.05} = [0.616, \infty)$$

Como $\hat{p} = 0.7 \in RC_{0.05}$ rechazaremos la hipótesis nula y aceptamos la alternativa, considerando, por lo tanto, que la moneda está trucada en el sentido de que la probabilidad de salir cara es mayor que la de salir cruz. La explicación de por qué tomamos esta decisión es que de ser cierta la hipótesis de que la probabilidad de salir cara es $p_0 = 0.5$ (H_0 cierta), sólo habría un 5% de probabilidad de que nos diese $\hat{p} > 0.616$ (región de rechazo). Por lo tanto, si a nosotros nos ha salido $\hat{p} > 0.616$ (que es un suceso poco probable si H_0 es cierta) es posible que se deba a que H_0 es falsa, por lo que rechazamos la hipótesis nula y aceptamos la alternativa.

9.2 Tipos de errores, nivel de significación y potencia de un contraste

Cuando realizamos un contraste de hipótesis de cara a tomar una decisión sobre un parámetro poblacional, estamos sujetos a incertidumbres debidas al muestreo y al nivel de significación que consideremos. Eso significa que podemos cometer errores, o bien rechazando la hipótesis nula cuando es verdadera o bien aceptándola cuando es falsa. En el primer caso cometeremos un error que se denomina de *tipo I* y en el segundo caso el error cometido será de *tipo II*.

Estos tipos de errores se resumen en la siguiente tabla:

Hipótesis / *Decisión*	H_0 verdadera	H_0 falsa
Se acepta H_0	Decisión correcta	Error de tipo II
Se rechaza H_0	Error de tipo I	Decisión correcta

Evidentemente la probabilidad de cometer un error tipo I está relacionada con el nivel de significación α. Esto es debido a que dicho nivel de significación nos indica que si realizamos un gran número de veces el contraste de hipótesis propuesto, un $\alpha \cdot 100\%$ de veces rechazaremos la hipótesis nula siendo cierta, por lo tanto un $\alpha \cdot 100\%$ de veces cometeremos un error tipo I. Por otro lado α también determina los tamaños de las regiones crítica y de aceptación, un menor valor de α supone un mayor valor de la región de aceptación, lo cual puede llevarnos a cometer mayores errores de tipo II, como veremos a continuación. En general se suelen usar niveles de significación de 0.05 ó de 0.01.

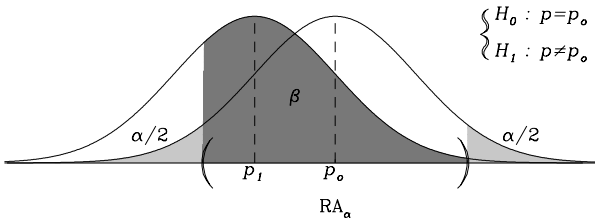


Figura 9.3*: Esquema para el cálculo de los errores de tipo II*

Vamos a analizar ahora la probabilidad de cometer un error de tipo II, denotada normalmente por β. Dicho error es imposible de calcular a no ser que se tenga una hipótesis alternativa específica. Por ejemplo, en el supuesto de la moneda trucada, donde nuestra hipótesis nula es $H_0 : p = 0.5$ y la hipótesis alternativa $H_1 : p \neq 0.5$, podemos preguntarnos por el error de tipo II que cometeríamos si aceptamos la hipótesis nula en el supuesto de que la verdadera probabilidad de que salga cara no sea $p = p_0 = 0.5$ sino $p = p_1 \neq 0.5$. Dicho error β se calculará como la probabilidad de que dado el estadístico muestral P obtengamos un valor muestral en la región de aceptación de H_0 supuesto que el auténtico valor poblacional de dicho parámetro es $p = p_1$. El cálculo de este error de tipo II queda ilustrado en la figura 9.3.

Según lo anterior, está claro que los errores de tipo I y de tipo II se relacionan entre sí. Para una muestra dada, la disminución del error de tipo I, esto es, la disminución del nivel de significación α, supone una mayor región de aceptación y, por lo tanto, un mayor error de tipo II, ya que aumenta β. Para cada caso particular habrá que estudiar cuál de los errores es más importante controlar y fijar las regiones de aceptación y rechazo de modo que se disminuya el error menos deseable. Esto nos lleva a definir un concepto importante en el contraste de hipótesis: la *potencia del contraste*. Dicha potencia del contraste es la probabilidad de rechazar la hipótesis nula H_0 cuando es falsa, por lo tanto, la potencia del contraste será $1 - \beta$ y depende del valor verdadero del parámetro poblacional que se quiere estudiar. La potencia de un contraste nos da una medida de la sensibilidad para detectar diferencias en los valores del parámetro. Si se fija de antemano el nivel de significación, elegiremos siempre el tipo de contraste que presente una potencia mayor para un determinado tamaño muestral.

Con todo esto, podemos repetir la tabla anterior asignando la probabilidad de cada decisión en función de la decisión que tomemos condicionada a que H_0 sea verdadera o falsa:

Hipótesis / *Decisión*	H_0 verdadera	H_0 falsa
Se acepta H_0	Decisión correcta $1 - \alpha = p$(aceptar $H_0 \mid H_0$ verdad)	Error de tipo II $\beta = p$(aceptar $H_0 \mid H_0$ falsa)
Se rechaza H_0	Error de tipo I $\alpha = p$(rechazar $H_0 \mid H_0$ verdad)	Decisión correcta $1 - \beta = p$(rechazar $H_0 \mid H_0$ falsa)

Ejemplo 3:

En el ejemplo anterior en el que hemos tirado la moneda 50 veces, vimos que un contraste de hipótesis unilateral derecho del tipo:

$$\begin{cases} H_0: & p = p_0 = 0.5 \\ H_1: & p > 0.5 \end{cases}$$

con un nivel de significación $\alpha = 0.05$, daba lugar a una región de aceptación $RA_{0.05} = (-\infty, 0.616)$ y una región de rechazo $RC_{0.05} = [0.616, \infty)$.

Si suponemos que el verdadero valor para la probabilidad de obtener cara es $p_1 = 0.75$, el error de tipo II lo obtendremos viendo que probabilidad de la distribución $N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right)$ está en el intervalo $RA_{0.05} = (-\infty, 0.616)$. Si consideramos que $P \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right) = N(0.75, 0.0612)$, tendremos que $\beta = p(P < 0.616) = p\left(Z < \frac{0.616 - 0.75}{0.0612}\right) = p(Z < -2.188)$ siendo $Z \sim N(0,1)$. Por lo tanto, mirando en las tablas, tendremos que $\beta = 0.014$. La potencia de este contraste será $1 - \beta = 0.986$.

Sin embargo, si el verdadero valor de la probabilidad de obtener cara fuese $p_1 = 0.65$, el error de tipo II sería diferente, ya que tendríamos que considerar que $P \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right) = N(0.65, 0.0675)$ y tendríamos $\beta = p(P < 0.616) = p\left(Z < \frac{0.616 - 0.65}{0.0675}\right) = p(Z < -0.504) = 0.307$, en cuyo caso la potencia del contraste sería $1 - \beta = 0.693$

En este mismo ejemplo, si en lugar de tirar la moneda 50 veces la tiramos 100 veces, la región de aceptación y de rechazo para un mismo nivel de significación $\alpha = 0.05$ cambiarán ya que ha cambiado el tamaño muestral, obteniéndose:
$RA_{0.05} = \left(-\infty, p_0 + z_{0.05}\sqrt{\frac{p_0(1-p_0)}{n}}\right) = (-\infty, 0.582)$ y $RC_{0.05} = \left[p_0 + z_{0.05}\sqrt{\frac{p_0(1-p_0)}{n}}, \infty\right) = [0.582, \infty)$. Ahora, en el supuesto de que la verdadera probabilidad de obtener cara fuese $p_1 = 0.65$, el error de tipo II habrá disminuido, ya que supondremos que $P \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n}}\right) = N(0.65, 0.0477)$, por lo tanto $\beta = p(P < 0.582) = p\left(Z < \frac{0.582 - 0.65}{0.0477}\right) = p(Z < -1.420) = 0.078$ y la potencia de este contraste será $1 - \beta = 0.922$

9.3 P-valor de un contraste de hipótesis

Cuando realizamos un contraste decidimos si aceptamos o rechazamos la hipótesis nula H_0 en función de que el valor del parámetro muestral se encuentre en la región de aceptación o en la de rechazo. Sin embargo, es interesante saber, tanto si aceptamos como si rechazamos la hipótesis nula, el grado de acuerdo entre los datos de nuestra muestra y la decisión que hemos tomado. Para ello utilizaremos el *P-valor*, que mide el grado de aceptación de la hipótesis nula para nuestros datos muestrales, de modo que cuanto mayor es el P-valor más se ajustan nuestros datos a la hipótesis nula y cuanto menor es dicho valor más fuerte es el argumento para rechazar la hipótesis nula. El P-valor será una probabilidad que calcularemos a partir del valor muestral del estadístico siguiendo el procedimiento descrito a continuación:

- *Contraste bilateral:* Supongamos que tenemos un contraste bilateral para el parámetro θ, de modo que $\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$, y hemos obtenido $\hat{\theta}$ como la estimación de θ a partir de nuestra muestra usando el estimador A. El P-valor en este caso se calculará como:

$$P\text{-valor} = p\left(|A - \theta_0| > \left|\hat{\theta} - \theta_0\right|\right)$$

  supuesto que H_0 es cierta.

- *Contraste unilateral:* Supongamos que tenemos un contraste unilateral derecho para el parámetro θ, $\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta > \theta_0 \end{cases}$, y $\hat{\theta}$ es el valor muestral obtenido con el estimador A, el P-valor será:

$$P\text{-valor} = p\left(A - \theta_0 > \hat{\theta} - \theta_0\right)$$

  suponiendo que H_0 es cierta.

  En el caso de que el contraste sea unilateral izquierdo, $\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta < \theta_0 \end{cases}$, el P-valor será:

$$P\text{-valor} = p\left(\theta_0 - A > \theta_0 - \hat{\theta}\right)$$

  suponiendo que H_0 es cierta.

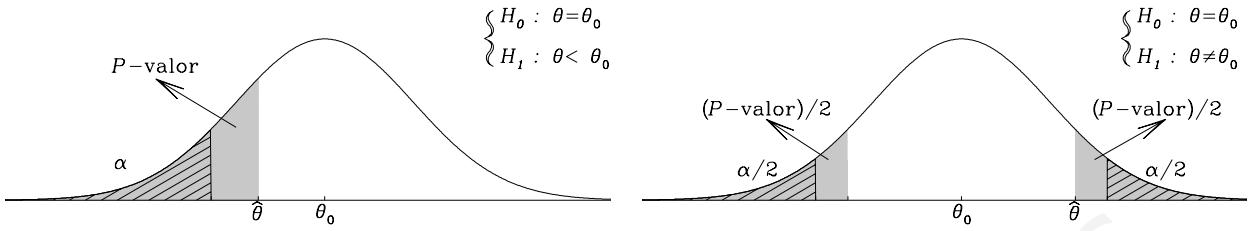


Figura 9.4*: Esquema del cálculo del P-valor para un contraste bilateral (a la derecha) y para un contraste unilateral izquierdo (a la izquierda).*

9.4 Contrastes de hipótesis para una población

9.4.1 Contraste de una proporción de una binomial

Supongamos que queremos contrastar el valor del parámetro poblacional p de una distribución binomial. Para ello utilizaremos el estadístico $P = \dfrac{\#\text{éxitos}}{\#\text{ensayos}}$, que sabemos que si el tamaño muestral n es grande (en la práctica se acepta si $np > 5$ y $n(1-p) > 5$) sigue una distribución normal $P \sim N\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.

Si queremos contrastar la proporción p con un valor teórico p_0, la variable aleatoria tipificada

$$Z = \frac{P - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

seguirá una distribución normal $Z \sim N(0,1)$

Suponiendo que $\hat{p}$ es la estimación muestral que hemos obtenido, el valor de la variable que usaremos en el contraste para comparar con las regiones de aceptación y de rechazo será $z_{muestral} = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Por lo tanto, dado el nivel de significación α, el contraste y las regiones de aceptación y críticas para el estadístico Z serán:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{cases}$

- Región de aceptación: $RA_\alpha = \left(-z_{\alpha/2}, z_{\alpha/2}\right)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
- Región crítica: $RC_\alpha = \left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$

- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & p = p_0 \\ H_1: & p > p_0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, z_\alpha)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [z_\alpha, \infty)$

- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & p = p_0 \\ H_1: & p < p_0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-z_\alpha, \infty)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = (-\infty, -z_\alpha]$

Ejemplo 4:

Supongamos que un jugador de baloncesto tenía hace 3 años un porcentaje de aciertos en tiros libre del 90%. A día de hoy, para tratar de comprobar si dicho porcentaje se mantiene hace 100 lanzamientos y acierta sólo 85, ¿podemos considerar que su porcentaje de aciertos no ha variado?

Usaremos un nivel de significación $\alpha = 0.05$. Vamos a aplicar un contraste bilateral sobre la proporción de una binomial, puesto que no tenemos indicaciones de que el porcentaje de aciertos haya aumentado o disminuído en los 3 años transcurridos. Así pues, tendremos que:

$$\begin{cases} H_0: & p = 0.9 \\ H_1: & p \neq 0.9 \end{cases}$$

El estadístico que utilizaremos será $Z = \dfrac{P - 0.9}{\sqrt{\dfrac{0.9 \cdot 0.1}{100}}} = \dfrac{P - 0.9}{0.03}$ y el valor muestral de Z será

$$z_{muestral} = \frac{P - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.85 - 0.9}{0.03} = -1.667.$$

Como la región de aceptación para $\alpha = 0.05$ es $RA_{0.05} = (-z_{0.025}, z_{0.025}) = (-1.96, 1.96)$, tendremos que $z_{muestral} \in RA_{0.05}$, así que no podemos rechazar la hipótesis nula de que el porcentaje de aciertos del jugador es el mismo que hace 3 años.

En este caso el P-valor es: $P\text{-valor} = 1 - p(-1.667 < Z < 1.667) = 2 \cdot p(Z > 1.667) = 2 \cdot 0.0478 = 0.0956$, lo que significa que si rechazasemos la hipótesis nula nos equivocaríamos un 9.56% de las veces.

9.4.2 Contraste de la media de una población normal

En el caso de querer realizar un contraste sobre el valor desconocido de la media de una variable aleatoria $X \sim N(\mu, \sigma)$ cuyo valor muestral $\hat{\mu}$ se habrá obtenido mediante el estimador $\bar{X}$, tendremos que considerar dos casos posibles:

(a) Varianza poblacional σ^2 conocida:

En este caso sabemos que la variable aleatoria $\bar{X}$ seguirá una distribución normal $\bar{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$, donde n es el tamaño muestral. En este caso, suponiendo que el valor teórico de la media poblacional es μ_0, utilizaremos la variable tipificada

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

que seguirá una distribución $Z \sim N(0,1)$, y el valor muestral que usaremos aceptar o rechazar la hipótesis nula será $z_{muestral} = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}$.

Suponiendo un nivel de significación α, el contraste y los correspondientes intervalos de aceptación y críticos serán:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-z_{\alpha/2}, z_{\alpha/2}\right)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
- *Contraste unilateral derecho:*

- Contraste: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$
- Región de aceptación: $RA_\alpha = (-\infty, z_\alpha)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
- Región crítica: $RC_\alpha = [z_\alpha, \infty)$

- *Contraste unilateral izquierdo:*

  - Contraste: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-z_\alpha, \infty)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = (-\infty, -z_\alpha]$

Ejemplo 5:

Un fabricante de monitores para ordenador asegura que la vida media de sus monitores es 3000 horas, con desviación típica de 48.6. Aceptando como válido el valor de la desviación típica, se quiere contrastar si la vida media es de 3000 horas o menor. Se controla la duración de 45 monitores elegidos al azar de su producción y se obtiene una vida media de 2960 horas. A la vista de los resultados, ¿qué se puede concluir?

En este caso, el contraste de hipótesis que plantearemos será unilateral izquierdo, ya que podemos tener sospechas que la vida media de los monitores es menor que lo que asegura el fabricante, por lo tanto:

$$\begin{cases} H_0: & \mu = 3000 \\ H_1: & \mu < 3000 \end{cases}$$

El estadístico para el contraste será $Z = \dfrac{\bar{X} - 3000}{48.6/\sqrt{45}} = \dfrac{\bar{X} - 3000}{7.245}$ y el valor muestral que toma dicho estadístico en nuestro caso será $z_{muestral} = \dfrac{2960 - 3000}{7.245} = -5.52$.

Como la región de aceptación para $\alpha = 0.05$ es $RA_{0.05} = (-z_{0.05}, \infty) = (-1.645, \infty)$, tendremos que $z_{muestral} \notin RA_{0.05}$ así que rechazaremos la hipótesis nula y aceptaremos que la vida media de los monitores es menor de 3000 horas. De hecho, como el P-valor es: $P\text{-valor} = p(Z < -5.52) < 0.19 \cdot 10^{-7}$, tendremos una probabilidad muy alta de no estar equivocándonos en nuestra decisión, ya que la probabilidad de que sea cierta la hipótesis nula es inferior al $0.19 \cdot 10^{-7}$.

(b) Varianza poblacional σ^2 desconocida:

Aunque sabemos que la variable aleatoria $\bar{X}$ seguirá una distribución normal $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, donde n es el tamaño muestral, como en este caso no conocemos el valor de la varianza, σ^2, en primer lugar tendremos que estimar dicho valor a partir de los datos de la muestra. Para ello utilizaremos como estimador la cuasivarianza muestral, $\mathcal{S}^2$. De este modo, suponiendo que el valor teórico de la media poblacional es μ_0, el estadístico que usaremos para el contraste será:

$$T = \frac{\bar{X} - \mu_0}{\mathcal{S}/\sqrt{n}}$$

que seguirá una distribución t-Student con $n-1$ grados de libertad, $T \sim t_{n-1}$.

Por lo tanto, si $\hat{\mu}$ y $\hat{\sigma}^2$ son los valores muestrales de la media y la varianza obtenidos a partir de los estimadores $\bar{X}$ y $\mathcal{S}^2$, respectivamente, el valor muestral del estadístico T que usaremos para el contraste de hipótesis será $t_{muestral} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}}$.

Suponiendo un nivel de significación α, el contraste y los correspondientes intervalos de aceptación y críticos serán:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}\right)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -t_{\alpha/2,n-1}\right] \cup \left[t_{\alpha/2,n-1}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, t_{\alpha,n-1})$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [t_{\alpha,n-1}, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$

- Región de aceptación: $RA_\alpha = (-t_{\alpha,n-1}, \infty)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
- Región crítica: $RC_\alpha = (-\infty, -t_{\alpha,n-1}]$

En el caso de que la muestra sea grande ($n > 30$) se puede aproximar el comportamiento de la variable aleatoria T a una normal tipificada.

Ejemplo 6:

Supongamos que la vida media de una determinada especie animal sigue una distribución normal y que queremos comprobar que dicha vida media es 12 años. Si tenemos una muestra de la edad de 9 miembros de dicha especie animal, dada por $\{4, 13, 8, 12, 8, 15, 14, 7, 8\}$. ¿Qué conclusión podemos obtener con estos resultados?

En primer lugar estimaremos la media y la varianza poblacionales utilizando los estimadores $\bar{X}$ y $\mathcal{S}^2$, obteniéndose $\hat{\mu} = 9.89$ y $\hat{\sigma}^2 = 13.8611 \Rightarrow \hat{\sigma} = 3.72$.

El contraste de hipótesis que plantearemos es:

$$\begin{cases} H_0: & \mu = 12 \\ H_1: & \mu \neq 12 \end{cases}$$

Sabemos que el estadístico $T = \dfrac{\bar{X} - \mu_0}{\mathcal{S}/\sqrt{n}} = \dfrac{\bar{X} - 12}{\mathcal{S}/3}$ se comportará como una t-Student con $n - 1 = 8$ grados de libertad. La región de aceptación de este contraste con $\alpha = 0.05$ será $RA_{0.05} = (-t_{0.025,8}, t_{0.025,8})$, como $t_{0.025,8} = 2.306$ tendremos que $RA_{0.05} = (-2.306, 2.306)$.

Como el valor muestral de la variable T en nuestro caso es $t_{muestral} = \dfrac{\hat{\mu} - 12}{\hat{\sigma}/3} = \dfrac{9.89 - 12}{3.72/3} = -1.70$, y se cumple que $t_{muestral} \in RA_{0.05}$ aceptaremos la hipótesis nula de que la vida media de esa especie animal es 12 años.

9.4.3 Contraste de la varianza de una población normal

Supongamos que queremos realizar un contraste de hipótesis sobre el valor de la varianza de una población normal $N(\mu, \sigma)$. En este caso conviene recordar que la variable aleatoria

$$\chi^2 = (n-1)\frac{\mathcal{S}^2}{\sigma^2}$$

donde $\mathcal{S}^2$ es la cuasivarianza muestral, σ^2 la varianza poblacional y n el tamaño muestral, se comporta como una distribución chi-cuadrado con $n - 1$ grados de libertad, $\chi^2 \sim \chi^2_{n-1}$.

De este modo, si σ_0^2 es el valor teórico para la varianza que queremos contrastar y $\hat{\sigma}^2$ es la estimación muestral de la varianza obtenida a partir del estimador cuasivarianza muestral $\mathcal{S}^2$, tendremos que el valor muestral que usaremos en el contraste será $\chi^2_{muestral} = (n-1)\dfrac{\hat{\sigma}^2}{\sigma_0^2}$.

Suponiendo un nivel de significación α, el contraste y los correspondientes intervalos de aceptación y críticos serán:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}\right)$, aceptamos H_0 si $\chi^2_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left[0, \chi^2_{1-\alpha/2,n-1}\right] \cup \left[\chi^2_{\alpha/2,n-1}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left[0, \chi^2_{\alpha,n-1}\right)$, aceptamos H_0 si $\chi^2_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left[\chi^2_{\alpha,n-1}, \infty\right)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 < \sigma_0^2 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(\chi^2_{1-\alpha,n-1}, \infty\right)$, aceptamos H_0 si $\chi^2_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left[0, \chi^2_{1-\alpha,n-1}\right]$

Ejemplo 7:

Analizar si en el ejemplo anterior la varianza poblacional puede ser $\sigma_0^2 = 12$.

Hemos calculado una estimación de la varianza poblacional σ^2 utilizando la cuasivarianza muestral $\mathcal{S}^2$, obteniéndose $\hat{\sigma}^2 = 13.8611$.

El contraste de hipótesis que plantearemos es:

$$\begin{cases} H_0: & \sigma^2 = 12 \\ H_1: & \sigma^2 \neq 12 \end{cases}$$

Como el tamaño muestral es $n = 9$, el estadístico para el contraste será $\chi^2 = (9-1)\dfrac{\mathcal{S}^2}{12} = \dfrac{2\mathcal{S}^2}{3}$ y seguirá una distribución χ^2_8. El valor muestral del estadístico será $\chi^2_{muestral} = \dfrac{2\hat{\sigma}^2}{3} = 9.241$.

Como la región de aceptación del contraste, para $\alpha = 0.05$, es $RA_{0.05} = \left(\chi^2_{0.975,8}, \chi^2_{0.025,8}\right) = (2.180, 17.535)$ y $\chi^2_{muestral} \in RA_{0.05}$ aceptaremos la hipótesis nula de que el valor de la varianza poblacional es $\sigma^2 = 12$

9.5 Contrastes de hipótesis para dos poblaciones

9.5.1 Contraste de comparación de dos proporciones

Supongamos ahora que tenemos dos poblaciones que siguen distribuciones binomiales, $Bin(n_1, p_1)$ y $Bin(n_2, p_2)$, y queremos hacer un contraste de hipótesis para comparar los parámetros p_1 y p_2. Si suponemos que las muestras son grandes, los estadísticos P_1 y P_2 tienen distribuciones normales, $P_1 \sim N\left(p_1, \sqrt{\dfrac{p_1(1-p_1)}{n_1}}\right)$ y $P_2 \sim N\left(p_2, \sqrt{\dfrac{p_2(1-p_2)}{n_2}}\right)$, por lo tanto, el estadístico $P_1 - P_2$ también seguirá una distribución normal y será un buen estimador de la diferencia de los parámetros poblacionales $p_1 - p_2$.

Así pues, como $P_1 - P_2 \sim N\left(p_1 - p_2, \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}\right)$, el estadístico

$$Z = \frac{P_1 - P_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

se comporta como una variable normal tipificada. Si queremos contrastar si las proporciones muestrales son iguales, la hipótesis nula será $H_0 : p_1 = p_2$. Por lo tanto estimaremos la varianza poblacional de $\bar{X}_1 - \bar{X}_2$, $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$, considerando que $p_1 = p_2$ y que dicha proporción puede ser estimada con una proporción muestral conjunta $\hat{p} = \dfrac{\#\text{total de éxitos}}{\#\text{total de ensayos}} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$, donde $\hat{p}_1$ y $\hat{p}_2$ son las proporciones muestrales para cada una de las distribuciones. Así, el valor muestral que tendremos que utilizar en nuestro contraste de hipótesis de igualdad de proporciones será $z_{muestral} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_1} + \dfrac{\hat{p}(1-\hat{p})}{n_2}}}$.

Con todo esto tendremos que, dado el nivel de significación α, el contraste y las regiones de aceptación y críticas para el estadístico Z serán:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & p_1 = p_2 \Leftrightarrow p_1 - p_2 = 0 \\ H_1: & p_1 \neq p_2 \Leftrightarrow p_1 - p_2 \neq 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-z_{\alpha/2}, z_{\alpha/2}\right)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & p_1 = p_2 \Leftrightarrow p_1 - p_2 = 0 \\ H_1: & p_1 > p_2 \Leftrightarrow p_1 - p_2 > 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, z_\alpha)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [z_\alpha, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & p_1 = p_2 \Leftrightarrow p_1 - p_2 = 0 \\ H_1: & p_1 < p_2 \Leftrightarrow p_1 - p_2 < 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-z_\alpha, \infty)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = (-\infty, -z_\alpha]$

Ejemplo 8:

Supongamos que queremos comparar dos vacunas para saber cual es más eficaz en la protección contra una determinada enfermedad. La primera de las vacunas se aplica a una muestra de 200 individuos de los cuales sólo 50 desarrollan la enfermedad. La segunda vacuna se aplica a una muestra de 180 individuos, de los cuales, sólo 36 desarrollan la enfermedad. ¿Qué se puede concluir sobre la eficacia de dichas vacunas?

Si consideramos las variables aleatorias que indican el número de individuos que no desarrollan la enfermedad, en ambos casos dichas variables aleatorias siguen distribuciones binomiales. En el primer caso con $n_1 = 200$ y probabilidad muestral $\hat{p}_1 = \dfrac{150}{200} = 0.75$ y en el segundo caso $n_2 = 180$ y probabilidad muestral $\hat{p}_2 = \dfrac{144}{180} = 0.80$.

Ante estos datos podríamos sospechar que la segunda vacuna es más eficaz que la primera, por lo que plantearemos el siguiente contraste unilateral:

$$\begin{cases} H_0: & p_1 = p_2 \Leftrightarrow p_1 - p_2 = 0 \\ H_1: & p_1 < p_2 \Leftrightarrow p_1 - p_2 < 0 \end{cases}$$

Considerando que la hipótesis nula es cierta, la variable aleatoria $Z = \dfrac{P_1 - P_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ seguirá una distribución normal $N(0,1)$.

Como la proporción muestral conjunta es $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \dfrac{150 + 144}{200 + 180} = 0.7737$

El valor muestral del estadístico Z considerando los datos del enunciado será:

$$z_{muestral} = \frac{0.75 - 0.80}{\sqrt{0.7737(1 - 0.7737)\left(\frac{1}{200} + \frac{1}{180}\right)}} = -1.163$$

Si tomamos $\alpha = 0.05$, la región de aceptación es $RA_{0.05} = (-z_{0.05}, \infty) = (-1.645, \infty)$. Como $z_{muestral} \in RA_{0.05}$, aceptaremos la hipótesis nula H_0, lo que significa que ambas vacunas tienen la misma eficacia.

9.5.2 Contraste de las medias de dos poblaciones normales independientes

Supongamos que tenemos dos variables aleatorias independientes X_1 y X_2 que siguen distribuciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, respectivamente. Si queremos comparar las medias de dichas distribuciones tendremos que tener en cuenta que las variables aleatorias $\bar{X}_1$ y $\bar{X}_2$ también siguen distribuciones normales, $N\left(\mu_1, \dfrac{\sigma_1}{\sqrt{n_1}}\right)$ y $N\left(\mu_2, \dfrac{\sigma_2}{\sqrt{n_2}}\right)$ y, por lo tanto, la variable aleatoria $\bar{X}_1 - \bar{X}_2$ sigue una distribución $N\left(\mu_1 - \mu_2, \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\right)$. Para analizar si las medias poblacionales son iguales o no tendremos que tener en cuenta las siguientes posibilidades:

(a) Varianzas poblacionales σ_1^2 y σ_2^2 conocidas:

Si las varianzas poblacionales de las variables aleatorias X_1 y X_2 son conocidas, sabemos que la variable aleatoria

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

sigue una distribución $N(0,1)$. Si $\hat{\mu}_1$ y $\hat{\mu}_2$ son los valores muestrales de las medias y considerando cierta la hipótesis nula de igual de medias poblacionales, el valor muestral para el estadístico Z es $z_{muestral} = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

Para un nivel de significación α, las regiones de aceptación y de rechazo, así como el contraste planteado serán los siguientes:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-z_{\alpha/2}, z_{\alpha/2}\right)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 > \mu_2 \Leftrightarrow \mu_1 - \mu_2 > 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, z_\alpha)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [z_\alpha, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 < \mu_2 \Leftrightarrow \mu_1 - \mu_2 < 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-z_\alpha, \infty)$, aceptamos H_0 si $z_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = (-\infty, -z_\alpha]$

Ejemplo 9:

Consideremos una población normal de varianza conocida $\sigma_1^2 = 16$, para la que hemos obtenido una muestra de 9 valores dada por $\{4, 13, 8, 12, 8, 15, 14, 7, 8\}$, y supongamos que tenemos otra población normal de varianza $\sigma_2^2 = 20$, para la que hemos obtenido la siguiente muestra $\{17, 14, 2, 12, 12, 6, 5, 11, 5\}$. Realizar un contraste de hipótesis sobre la igualdad de medias.

El contraste de hipótesis a realizar será:

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

y el estadístico que utilizaremos para dicho contraste es $Z = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$, que sigue una distribución $N(0,1)$. Bajo el supuesto de que se cumple la hipótesis nula, tendremos que $Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$, por lo que el valor muestral de dicho estadístico será $z_{muestral} = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.

Calculamos los valores muestrales de las medias de las dos poblaciones, obteniéndose $\hat{\mu}_1 = 9.89$ y $\hat{\mu}_2 = 9.33$, por lo tanto $z_{muestral} = \dfrac{9.89 - 9.33}{\sqrt{\dfrac{16}{9} + \dfrac{20}{9}}} = 0.28$

Considerando un nivel de significación $\alpha = 0.05$, la región de aceptación para la hipótesis nula es $RA_{0.05} = (-z_{0.025}, z_{0.025}) = (-1.96, 1.96)$. Como $z_{muestral} \in RA_{0.05}$, aceptaremos la hipótesis nula de igualdad de medias. Además en este caso, como el P-valor es muy alto, $P\text{-valor} = 1 - p(-0.28 < Z < 0.28) = 2 \cdot p(Z > 0.28) = 2 \cdot 0.3897 = 0.7794$, podemos estar seguros de la certeza de la hipótesis nula.

(b) Varianzas poblacionales desconocidas e iguales, $\sigma_1^2 = \sigma_2^2$:

Como podemos aceptar que $\sigma_1^2 = \sigma_2^2 = \sigma^2$, la variable aleatoria $Z = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ seguirá una distribución $N(0,1)$.

Sin embargo, como desconocemos el valor de σ^2 lo estimaremos utilizando la cuasivarianza

muestral conjunta $\mathcal{S}^2 = \dfrac{(n_1-1)\mathcal{S}_1^2 + (n_2-1)\mathcal{S}_2^2}{n_1+n_2-2}$ que sigue una distribución χ^2 con n_1+n_2-2 grados de libertad, obteniéndose la estimación $\hat{\sigma}^2$. Por lo tanto, si definimos la variable aleatoria

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\mathcal{S}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

sabemos que sigue una distribución t-Student de n_1+n_2-2 grados de libertad.

Si $\hat{\mu}_1$ y $\hat{\mu}_2$ son los valores muestrales de las medias y considerando cierta la hipótesis nula de igual de medias poblacionales, el valor muestral para el estadístico T es $t_{muestral} = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.

Para un nivel de significación α, las regiones de aceptación y de rechazo, así como el contraste planteado serán los siguientes:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-t_{\alpha/2,n_1+n_2-2}, t_{\alpha/2,n_1+n_2-2}\right)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -t_{\alpha/2,n_1+n_2-2}\right] \cup \left[t_{\alpha/2,n_1+n_2-2}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 > \mu_2 \Leftrightarrow \mu_1 - \mu_2 > 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, t_{\alpha,n_1+n_2-2})$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [t_{\alpha,n_1+n_2-2}, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 < \mu_2 \Leftrightarrow \mu_1 - \mu_2 < 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-t_{\alpha,n_1+n_2-2}, \infty)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = (-\infty, -t_{\alpha,n_1+n_2-2}]$

Ejemplo 10:

Considerar el siguiente par de muestras y, sabiendo que provienen de distribuciones normales con varianzas similares, realizar un contraste de hipótesis para averiguar si la segunda distribución tiene una media menor.

Muestra 1: $\{1.78, 0.52, 5.13, 3.86, 6.29, 2.51, 2.11, 7.66, 6.27, -4.57\}$

Muestra 2: $\{7.47, -0.80, -0.60, 0.03, 4.49, -0.14, -0.99, 0.74, 1.45, 5.38\}$

El contraste de hipótesis que planteamos es:

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

En primer lugar calculamos las estimaciones para las medias de las dos muestras, $\hat{\mu}_1$ y $\hat{\mu}_2$, utilizando el estimador $\bar{X}$ y las estimaciones de las varianzas, $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$, utilizando la cuasivarianza muestral, $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$. Los resultados que se obtienen son $\hat{\mu}_1 = 3.52$ y $\hat{\sigma}_1^2 = 12.69$ y $\hat{\mu}_2 = 1.70$ y $\hat{\sigma}_2^2 = 8.95$

Por lo tanto la estimación para la varianza conjunta será $\hat{\sigma}^2 = \dfrac{9 \cdot 12.69 + 9 \cdot 8.95}{18} = 10.82 \Rightarrow$ $\Rightarrow \hat{\sigma} = 3.289$ (donde hemos utilizado $\mathcal{S}^2 = \frac{(n_1-1)\mathcal{S}_1^2 + (n_2-1)\mathcal{S}_2^2}{n_1+n_2-2}$).

La variable aleatoria $T = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\mathcal{S}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ seguirá una distribución t-Student con $n_1 + n_2 - 2 = 18$ grados de libertad. Suponiendo que se satisface la hipótesis nula, dicho estadístico se transformará en $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\mathcal{S}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, cuyo valor muestral es $t_{muestral} = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} =$

$$\frac{3.52 - 1.70}{3.289\sqrt{\dfrac{1}{10} + \dfrac{1}{10}}} = 1.237.$$

La región de aceptación, para $\alpha = 0.05$, de H_0 es $RA_{0.05} = (-\infty, t_{0.05,18}) = (-\infty, 1.734)$, por lo que aceptamos la hipótesis nula de igualdad de medias y rechazamos que $\mu_1 > \mu_2$, ya que $t_{muestral} \in RA_{0.05}$

(c) Varianzas poblacionales desconocidas y distintas, $\sigma_1^2 \neq \sigma_2^2$:

Como en este caso no se puede establecer que ambas varianzas poblacionales sean iguales, tendremos, en primer lugar, que estimarlas a partir de los datos muestrales utilizando las cuasivarianzas muestrales, $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$, dando lugar a las estimaciones $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$.

En este caso, la variable aleatoria

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\mathcal{S}_1^2}{n_1} + \dfrac{\mathcal{S}_2^2}{n_2}}}$$

seguirá una distribución t-Student con g grados de libertad, siendo g el entero más próximo al número $h = \dfrac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\hat{\sigma}_1^2/n_1\right)^2}{n_1+1} + \frac{\left(\hat{\sigma}_2^2/n_2\right)^2}{n_2+1}} - 2$. Por lo tanto, si $\hat{\mu}_1$ y $\hat{\mu}_2$ son los valores de las medias muestrales, bajo el supuesto de que se satisface la hipótesis nula, esto es, $\mu_1 = \mu_2$, el valor muestral para el estadístico T será $t_{muestral} = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$.

Para un nivel de significación α, las regiones de aceptación y de rechazo, así como el contraste planteado serán los siguientes:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-t_{\alpha/2,g}, t_{\alpha/2,g}\right)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -t_{\alpha/2,g}\right] \cup \left[t_{\alpha/2,g}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 > \mu_2 \Leftrightarrow \mu_1 - \mu_2 > 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, t_{\alpha,g})$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [t_{\alpha,g}, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1: & \mu_1 < \mu_2 \Leftrightarrow \mu_1 - \mu_2 < 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-t_{\alpha,g}, \infty)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$

– Región crítica: $RC_\alpha = (-\infty, -t_{\alpha,g}]$

Ejemplo 11:

Repetir el ejemplo 9 suponiendo que las varianzas poblacionales son desconocidas (y no se pueden considerar iguales).

El contraste de hipótesis a realizar sigue siendo:

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

Para la primera de las poblaciones calculamos la media muestral es $\hat{\mu}_1 = 9.89$ y la estimación de la varianza poblacional, obtenida utilizando la cuasivarianza muestral, es $\hat{\sigma}_1^2 = 13.8611$

Para la segunda población la estimación de la media es $\hat{\mu}_2 = 9.33$ y la de la varianza, obtenida usando el estimador cuasivarianza muestral, es $\hat{\sigma}_2^2 = 25.0$

Sabemos que el estadístico $T = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\mathcal{S}_1^2}{n_1} + \dfrac{\mathcal{S}_2^2}{n_2}}}$ sigue una distribución t-Student con g grados de libertad, siendo g el entero más próximo a $h = \dfrac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\hat{\sigma}_1^2/n_1\right)^2}{n_1+1} + \frac{\left(\hat{\sigma}_2^2/n_2\right)^2}{n_2+1}} - 2 = \dfrac{\left(\frac{13.8611}{9} + \frac{25}{9}\right)^2}{\frac{(13.8611/9)^2}{10} + \frac{(25/9)^2}{10}} - 2 = 16.48$, por lo que tomaremos $g = 16$. El valor muestral de dicho estadístico será, bajo el supuesto de que se satisface la hipótesis nula de igualdad de medias poblacionales, $t_{muestral} = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}} = \dfrac{9.89 - 9.33}{\sqrt{\frac{13.8611}{9} + \frac{25}{9}}} = 0.269$

Como la región de aceptación, para nivel de significación $\alpha = 0.05$, de dicho contraste es $RA_{0.05} = (-t_{0.025,16}, t_{0.025,16}) = (-2.12, 2.12)$ y $t_{muestral} \in RA_{0.05}$, aceptaremos la hipótesis nula.

9.5.3 Contraste de hipótesis de igualdad de medias para datos apareados

Supongamos que tenemos un experimento de observaciones apareadas, es decir, tenemos dos muestras no independientes de tamaño n de dos poblaciones normales. Según vimos en el tema anterior, este problema se resuelve definiendo una nueva variable aleatoria D que sea la diferencia entre cada par de observaciones apareadas. Para una muestra suficientemente grande se puede considerar que D sigue una distribución normal de media $\mu_D = \mu_1 - \mu_2$ y

de varianza σ_D^2, que cumplirá que $\sigma_D^2 \neq \sigma_1^2 + \sigma_2^2$ ya que se trata de variables dependientes. Las estimaciones puntuales de estos parámetros poblacionales las calcularemos utilizando los estimadores $\bar{D} = \dfrac{1}{n}\sum_{i=1}^{n} d_i$ y $\mathcal{S}_D^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{D}\right)^2$, siendo $d_i = x_{1_i} - x_{2_i}$ para $i = 1, 2, \ldots, n$, y dando lugar a las estimaciones puntuales $\hat{\mu}_D$ y $\hat{\sigma}_D^2$. El estadístico

$$T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}}$$

seguirá una distribución t-Student de $n-1$ grados de libertad. Como en este caso el contraste de hipótesis consiste en probar si se satisface la hipótesis nula $H_0 : \mu_D = 0$ o no, el valor muestral del estadístico T, bajo la suposición de que H_0 es cierta, será $t_{muestral} = \dfrac{\hat{\mu}_D}{\hat{\sigma}_D/\sqrt{n}}$.

Por lo que, para un nivel de significación de α, el contraste de hipótesis y las regiones de aceptación y rechazo son:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0 : & \mu_D = 0 \\ H_1 : & \mu_D \neq 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}\right)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left(-\infty, -t_{\alpha/2,n-1}\right] \cup \left[t_{\alpha/2,n-1}, \infty\right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0 : & \mu_D = 0 \\ H_1 : & \mu_D > 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-\infty, t_{\alpha,n-1})$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [t_{\alpha,n-1}, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0 : & \mu_D = 0 \\ H_1 : & \mu_D < 0 \end{cases}$
  - Región de aceptación: $RA_\alpha = (-t_{\alpha,n-1}, \infty)$, aceptamos H_0 si $t_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = (-\infty, -t_{\alpha,n-1}]$

En el caso de que la muestra sea grande ($n > 30$) se puede aproximar el comportamiento de la variable aleatoria T a una normal tipificada.

Ejemplo 12:

Se aplica un prodimiento para aumentar el rendimiento en 10 fábricas muy diferentes, consistente en no dejar tomarse el bocadillo a media mañana. Los rendimientos (en ciertas unidades, como toneladas/día) antes y después son:

Antes: X_1	13	22	4	10	63	18	34	6	19	43
Después: X_2	15	22	2	15	65	17	30	12	20	42

Contrastar si aumenta la producción el no permitir el bocadillo de media mañana.

En este caso se trata de un contraste unilateral derecho sobre la media μ_D de la diferencia $D = X_2 - X_1$

$$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D > 0 \end{cases}$$

Como las diferencias d_i son $\{2, 0, -2, 5, 2, -1, -4, 6, 1, -1\}$, tendremos que $\hat{\mu}_D = 0.8$ y, además, la estimación de la desviación típica será $\hat{\sigma}_D = 3.08$, obtenida a partir de la cuasidesviación típica de la muestra.

Como el estadístico $T = \dfrac{\bar{D} - \mu_D}{\mathcal{S}_D/\sqrt{n}}$ sigue una distribución t-Student con $n - 1 = 9$ grados de libertad, bajo el supuesto de que es cierta la hipótesis nula, el valor muestral de dicho estadístico es $t_{muestral} = \dfrac{0.8}{3.08/\sqrt{10}} = 0.821$.

Para $\alpha = 0.05$ la región de aceptación de H_0 es $RA_{0.05} = (-\infty, t_{0.05,9}) = (-\infty, 1.833)$. Con estos datos concluiremos que no se puede rechazar la hipótesis nula, por lo que no queda probado que no permitir el bocadillo de media mañana aumente la producción.

9.5.4 Contraste de hipótesis para la comparación de varianzas de poblaciones normales

Hemos visto en apartados anteriores que, en determinadas ocasiones, podemos estar interesados en saber si dos poblaciones normales tienen la misma varianza o no.

Supongamos que σ_1^2 y σ_2^2 son las varianzas poblacionales de las dos poblaciones y las variables aleatorias $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$ son las cuasivarianzas muestrales. Vimos en capítulos anteriores que la variable aleatoria

$$F = \frac{\mathcal{S}_1^2/\sigma_1^2}{\mathcal{S}_2^2/\sigma_2^2}$$

sigue una distribución F de Fisher con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador, $F_{(n_1-1,n_2-1)}$, siendo n_1 y n_2 los tamaños de las muestras de las dos poblaciones.

Si queremos contrastar si las varianzas poblacionales de las dos poblaciones son iguales, la hipótesis nula será $H_0: \sigma_1^2 = \sigma_2^2$. Suponiendo que dicha hipótesis es cierta, el estadístico F se reescribirá como $F = \dfrac{\mathcal{S}_1^2}{\mathcal{S}_2^2}$, cuyo valor muestral será $f_{muestral} = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$, siendo $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ las estimaciones de la varianza poblacional obtenidas a partir del estadístico cuasivarianza muestral.

Para un nivel de significación de α, el contraste de hipótesis y las regiones de aceptación y rechazo son:

- *Contraste bilateral:*
  - Contraste: $\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \Leftrightarrow \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \Leftrightarrow \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$
  - Región de aceptación: $RA_\alpha = \left(\dfrac{1}{F_{\alpha/2;n_2-1,n_1-1}}, F_{\alpha/2;n_1-1,n_2-1} \right)$, aceptamos H_0 si $f_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = \left[0, \dfrac{1}{F_{\alpha/2;n_2-1,n_1-1}} \right] \cup \left[F_{\alpha/2;n_1-1,n_2-1}, \infty \right)$
- *Contraste unilateral derecho:*
  - Contraste: $\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \Leftrightarrow \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \sigma_1^2 > \sigma_2^2 \Leftrightarrow \dfrac{\sigma_1^2}{\sigma_2^2} > 1 \end{cases}$
  - Región de aceptación: $RA_\alpha = [0, F_{\alpha;n_1-1,n_2-1})$, aceptamos H_0 si $f_{muestral} \in RA_\alpha$
  - Región crítica: $RC_\alpha = [F_{\alpha;n_1-1,n_2-1}, \infty)$
- *Contraste unilateral izquierdo:*
  - Contraste: $\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \Leftrightarrow \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \sigma_1^2 < \sigma_2^2 \Leftrightarrow \dfrac{\sigma_1^2}{\sigma_2^2} < 1 \end{cases}$

– Región de aceptación: $RA_\alpha = \left(\frac{1}{F_{\alpha;n_2-1,n_1-1}}, \infty\right)$, aceptamos H_0 si $f_{muestral} \in RA_\alpha$

– Región crítica: $RC_\alpha = \left[0, \frac{1}{F_{\alpha;n_2-1,n_1-1}}\right]$

Ejemplo 13:

Contrastar si las varianzas de las poblaciones del ejemplo 11 son iguales.

En el ejemplo 11 teníamos dos muestras de poblaciones normales, para las cuales obtuvimos como estimaciones de las varianzas muestrales los valores $\hat{\sigma}_1^2 = 13.8611$ y $\hat{\sigma}_2^2 = 25.0$, obtenidas con el estimador cuasivarianza muestral.

Como en este caso el contraste de hipótesis es:

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

Como $n_1 = n_2 = 9$, la región de aceptación de H_0 es, para $\alpha = 0.05$, $RA_{0.05} = \left(\frac{1}{F_{0.025;8,8}}, F_{0.025;8,8}\right) = \left(\frac{1}{4.433}, 4.433\right) = (0.226, 4.433)$. Como el valor muestral del estadístico de contraste es $f_{muestral} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{13.8611}{25.0} = 0.554 \in RA_{0.05}$, aceptaremos la hipótesis nula de igualdad de varianzas.

Capítulo 10

Hipótesis estadísticas II

Estadística no paramétrica. La prueba χ^2. Ajuste de una distribución observada a una distribución teórica. Pruebas de independencia y de homogeneidad.

10.1 Introducción

En el capítulo anterior hemos estudiado contrastes de hipótesis que nos permiten hacer suposiciones sobre los parámetros poblacionales de una determinada distribución de probabilidad, que en general es una distribución normal. Para poder hacer dichas suposiciones es necesario considerar que nuestra muestra es aleatoria y proviene de una población con una distribución dada.

En este capítulo vamos a presentar diversos contrastes de hipótesis que nos permitirán comprobar si la distribución que suponemos para una población es consistente con la muestra, si dos muestras diferentes pueden considerarse homogéneas o si las observaciones de dos o más parámetros de una población son independientes. Estos contrastes que no hacen referencia a los parámetros poblacionales sino al tipo de distribución se denominan *contrastes no paramétricos.*

En primer lugar, analizaremos como comprobar si los datos de una muestra siguen una determinada distribución de probabilidad, son los denominados *contrastes de bondad de ajuste.* Dentro de los contrastes de bondad de ajuste, nos centraremos en las *pruebas* χ^2.

Posteriormente veremos otros contrastes no paramétricos (basados también en pruebas χ^2) que nos permitirán comparar dos muestras para analizar si son homogéneas (*contrastes de homogeneidad*), o si dos o más características de una población son independientes o no (*contrastes de independencia*).

10.2 Pruebas χ^2 de bondad de ajuste

En determinadas ocasiones necesitamos saber si una muestra dada se ajusta a un modelo teórico o no. Debemos para ello diseñar un contraste de que nos permita determinar si nuestra muestra sigue la distribución de probabilidad teórica considerada.

Para comprobar si nuestra muestra se puede describir con una distribución de probabilidad dada formularemos el siguiente contraste de probabilidad:

$$\begin{cases} H_0: & \text{La muestra sigue la distribución de probabilidad teórica elegida} \\ H_1: & \text{La muestra no sigue la distribución de probabilidad teórica elegida} \end{cases}$$

Para ello, primero tenemos que definir un estadístico que nos permita evaluar el contraste. La forma de hacerlo es darse cuenta que al tratarse de una muestra aleatoria siempre va a haber desviaciones con respecto a la población teórica. Tendremos que comprobar si dichas deviaciones son compatibles con fluctuaciones debidas al azar o si son debidas a que la muestra no sigue la distribución de densidad supuesta.

La prueba que describiremos a continuación, denominada *prueba χ^2 de bondad de ajuste*, consiste en comparar las frecuencias observadas en la muestra con las frecuencias esperadas si la muestra siguiese la distribución de probabilidad teórica. Dichas frecuencias son o bien el número de veces que se repite en nuestra muestra cada valor dado de la variable aleatoria o bien el número de datos de nuestra muestra que se encuentran cada intervalo de la variable aleatoria (supuesto que hemos dividido recorrido de la variable en intervalos).

Supongamos que nuestra variable aleatoria X puede tomar los valores X_1, X_2, ..., X_k, bien porque es una variable discreta con dicho recorrido, o bien porque hemos agrupado en intervalos centrados en dichos valores nuestra variable aleatoria por ser continua. Definimos O_i como las frecuencias observadas en nuestra muestra para el valor X_i de la variable y E_i la frecuencia esperada, que se puede calcular como $E_i = np_i$, siendo n el tamaño muestral y p_i la probabilidad para la variable X_i.

Valor de la variable	X_1	X_2	...	X_k	*Total*
Frecuencia observada	$O_1 = n_1$	$O_2 = n_2$	...	$O_k = n_k$	$\sum_{i=1}^{k} n_i = n$
Frecuencia esperada	$E_1 = np_1$	$E_2 = np_2$	...	$E_k = np_k$	$\sum_{i=1}^{k} np_i = n$

Se puede demostrar que si definimos el siguiente estadístico, denominado *estadístico de*

Pearson,

$$\mathcal{Q} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

dicho estadístico, si $E_i = np_i > 5$, se comporta de forma aproximada como una distribución chi-cuadrado con $k-1$ grados de libertad, χ^2_{k-1}.

La razón *intuitiva* que explica por qué podemos asumir dicho comportamiento la describimos a continuación. Definamos una variable aleatoria O_i que indica el número de veces que aparece X_i en nuestra muestra, por lo que si la muestra es suficientemente grande se puede aproximar por una distribución de Poisson de parámetro $\lambda_i = E_i = np_i$, así que $O_i \sim Poi(np_i)$. Si $\lambda_i = np_i > 5$ podemos aproximar dicha distribución de Poisson por una distribución normal, por lo que $O_i \sim N\left(\lambda_i, \sqrt{\lambda_i}\right) = N\left(np_i, \sqrt{np_i}\right)$, así que la variable aleatoria $Z_i = \dfrac{O_i - np_i}{\sqrt{np_i}}$ seguirá una distribución $N(0,1)$. Si sumamos los cuadrados de k variables normales tipificadas independientes se comportará como una distribución chi-cuadrado con k grados de libertad, sin embargo en nuestro caso no tenemos k variables linealmente independientes, ya que tenemos la restricción $\sum_{i=1}^{k} O_i = n$, por lo que $\mathcal{Q} = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \dfrac{(O_i - np_i)^2}{np_i}$ sigue una distribución chi-cuadrado con $k-1$ grados de libertad, $\mathcal{Q} \sim \chi^2_{k-1}$.

Por todo lo expresado anteriormente, para poder aplicar el método anterior tendremos que exigir que la muestra sea suficientemente grande (en la práctica cuando $n > 30$) y que las frecuencias esperadas para cada X_i sean mayores que 5, $np_i > 5$. Si esto último no se cumple, podremos agrupar diferentes valores de X_i en un único intervalo para que se cumpla la condición, lo que reducirá los grados de libertad.

Analicemos ahora el significado del estadístico $\mathcal{Q}$. Como en los distintos sumandos de dicho estadístico aparece en el numerador el cuadrado de la diferencia entre la frecuencia observada y la esperada, cuando menor sea el valor del estadístico $\mathcal{Q}$ menor será la diferencia entre el valor esperado y el observado para las frecuencias, así que dichas diferencias podrían ser debidas a fluctuaciones estadísticas. Sin embargo, si las frecuencias observadas son muy diferentes de las esperadas, el estadístico $\mathcal{Q}$ será muy grande y significa que es poco probable que dichas diferencias sean debidas a fluctuaciones aleatorias debidas al muestreo y se deberán a que la distribución poblacional de la muestra no se corresponde con la distribución teórica que habíamos supuesto en nuestro contraste.

Por todo lo explicado anteriormente podemos reescribir el contraste de hipótesis para analizar la bondad del ajuste como:

$$\begin{cases} H_0: & O_i = np_i, \forall i = 1, \ldots, k \\ H_1: & O_i \neq np_i, \text{ para algún } i \end{cases}$$

En este caso, para un nivel de significación α las regiones de aceptación y de rechazo serán, respectivamente:

$$RA_\alpha = \left[0, \chi^2_{\alpha,k-1}\right)$$
$$RC_\alpha = \left[\chi^2_{\alpha,k-1}, \infty\right)$$

estando la región de rechazo siempre a la derecha en este tipo de contrastes de hipótesis. Así, aceptaremos la hipótesis nula para valores pequeños del estadístico $\mathcal{Q}$ (pequeñas diferencias entre lo observado y lo esperado) y lo rechazaremos para valores grandes de $\mathcal{Q}$ (grandes diferencias entre lo observado y lo esperado).

Una consideración importante es que si, para calcular las frecuencias esperadas para la distribución, tenemos que estimar previamente p parámetros de la población, los grados de libertad de la distribución chi-cuadrado quedan reducidos por el número p de parámetros que hayamos estimados. Así, por ejemplo, si queremos analizar si una muestra sigue una distribución normal de la cual desconocemos su media y su varianza, previamente debemos estimarlas a partir de la muestra, de modo que si tenemos k intervalos para el cálculo de la frecuencia, el estadístico $\mathcal{Q} = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$ seguirá una distribución $\chi^2_{k-1-2} = \chi^2_{k-3}$ debido a los dos parámetros, media y varianza, que hemos estimado.

Ejemplo 1:

Queremos saber si un cierto dado está trucado, para ello lo lanzamos 600 veces obteniéndose los siguientes resultados:

X_i	1	2	3	4	5	6
O_i	92	85	102	94	117	110

El contraste de hipótesis que plantearemos será:

$$\begin{cases} H_0: & \text{la población sigue una distribución uniforme} \\ H_1: & \text{la población no sigue una distribución uniforme} \end{cases}$$

En este caso, al lanzar el dado 600 veces, si la población sigue una distribución uniforme, el valor esperado para el número de veces que sale cada cara del dado será $np_i = 600 \cdot \frac{1}{6} = 100$.

El número de grados de libertad en este caso es $k - 1 = 6 - 1 = 5$. Calculamos el valor muestral del estadístico $\mathcal{Q}_{muestral} = \sum_{i=1}^{6} \frac{(O_i - 100)^2}{100} = 7.18$. Tomando un nivel de significación de $\alpha = 0.05$, como $\chi^2_{0.05,5} = 11.070$ y se cumple que $\mathcal{Q}_{muestral} < \chi^2_{0.05,5}$ concluiremos que no

se puede rechazar la hipótesis nula H_0 y que las fluctuaciones con respecto a las frecuencias observadas son debidas al azar.

Ejemplo 2:

Una fábrica de componentes electrónicos decide analizar la eficiencia en su cadena de producción y mide, durante 40 días, el número de componentes fallidos que se producen al día, obteniéndose los siguientes resultados:

$X_i = \#$ *componentes fallidos*	0	1	2	3	4	≥ 5
$O_i =$ *frecuencia (días)*	3	6	12	11	1	7

Analizar si el número de errores diarios en la cadena de producción sigue una distribución de Poisson de parámetro $\lambda = 2$.

Vamos a realizar un contraste de hipótesis para analizar si los datos muestrales se pueden ajustar a una distribución de Poisson con parámetro $\lambda = 2$, por lo tanto:

$$\begin{cases} H_0: & \text{la población sigue una distribución de Poisson con } \lambda = 2 \\ H_1: & \text{la población no sigue una distribución de Poisson con } \lambda = 2 \end{cases}$$

En el caso de que se cumpla la hipótesis nula, la función de densidad de probabilidad será $f(x) = \dfrac{2^x e^{-2}}{x!}$ y, por lo tanto, el número días en los que esperamos x_i errores será $E_i = nf(x_i) = 40f(x_i)$, así que tendremos:

X_i	0	1	2	3	4	≥ 5
O_i	3	6	12	11	1	7
E_i	5.4	10.8	10.8	7.2	3.6	2.2

Como el valor esperado de los dos últimos intervalos es menor que 5, los agruparemos en uno sólo:

X_i	0	1	2	3	≥ 4
O_i	3	6	12	11	8
E_i	5.4	10.8	10.8	7.2	5.8

Calculamos ahora el estadístico $\mathcal{Q} = \displaystyle\sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i}$ que seguirá una distribución χ^2_4. Como tenemos que $\mathcal{Q}_{muestral} = 6.17$ y, considerando un nivel de significación $\alpha = 0.05$, $\chi^2_{0.05,4} =$

$9.488 > \mathcal{Q}_{muestral}$, aceptaremos la hipótesis nula de que la distribución de errores diarios en la cadena de producción sigue una ley de Poisson con $\lambda = 2$.

Ejemplo 3:

La distribución de frecuencias de los diámetros normales en cm, es decir, los diámetros de los árboles a 1.30 cm del suelo, de 100 alcornoques elegidos al azar es:

Diámetro (en cm)	$(15, 20]$	$(20, 25]$	$(25, 30]$	$(30, 35]$	$(35, 40]$	$(40, 45]$
# alcornoques	3	15	28	39	11	4

Analizar si podemos suponer que el diámetro de los árboles siguen una distribución normal.

En primer lugar tenemos que estimar la media y la varianza de la población, para lo que usaremos los estimadores $\bar{X}$ y $\mathcal{S}^2$. Para ello, construímos previamente la siguiente tabla con las frecuencias y los valores de las variables:

Diámetro	x_i	$O_i = n_i$	$n_i x_i$	$n_i x_i^2$
$(15, 20]$	17.5	3	52.5	918.75
$(20, 25]$	22.5	15	337.5	7593.75
$(25, 30]$	27.5	28	770.0	21175.00
$(30, 35]$	32.5	39	1267.5	41193.75
$(35, 40]$	37.5	11	412.5	15468.75
$(40, 45]$	42.5	4	170.0	7225.00
		$\sum_{i=1}^{6} n_i = 100$	$\sum_{i=1}^{6} n_i x_i = 3010.0$	$\sum_{i=1}^{6} n_i x_i^2 = 93575.00$

Por lo tanto, $\hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{6} n_i x_i = 30.1$ y $\hat{\sigma}^2 = \dfrac{n}{n-1}\left(\dfrac{1}{n}\sum_{i=1}^{6} n_i x_i^2 - \hat{\mu}^2\right) = 30.04 \Rightarrow \hat{\sigma} = 5.48$

Ahora estimamos las frecuencias esperadas suponiendo que el diámetro de los alcornoques sigue una distribución $N(30.1, 5.48)$

Diámetro	p_i	$E_i = np_i$	$O_i = n_i$
≤ 20	0.0327	3.27	3
$(20, 25]$	0.1433	14.33	15
$(25, 30]$	0.3167	31.67	28
$(30, 35]$	0.3217	32.17	39
$(35, 40]$	0.1502	15.02	11
> 40	0.0354	3.54	4

Como el primer y el último intervalo tienen valores esperados de las frecuencias menores que 5 los agrupamos para obtener:

Diámetro	p_i	$E_i = np_i$	$O_i = n_i$
≤ 25	0.1760	17.60	18
$(25-30]$	0.3167	31.67	28
$(30-35]$	0.3217	32.17	39
> 35	0.1856	18.56	15

A partir de estos datos calculamos $\mathcal{Q}_{muestral} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 2.57$. Como tenemos $k = 4$ intervalos y hemos estimado $p = 2$ parámetros a partir de la muestra, el número de grados de libertad es $k - 1 - p = 4 - 1 - 2 = 1$. Considerando un nivel de significación $\alpha = 0.05$, como $\chi^2_{0.05,1} = 3.84 > \mathcal{Q}_{muestral} = 2.57$, aceptaremos la hipótesis nula de que la distribución del diámetro del tronco sigue una distribución normal.

10.3 Pruebas χ^2 de independencia de dos variables

En muchas ocasiones se presentan problemas en los que aparecen dos o más características de los elementos de una población y es necesario saber si son dependientes entre sí o independientes, por ejemplo, el peso y la estatura de una población de individuos. Vamos a explicar ahora como plantear un test de hipótesis, basado en la distribución χ^2, que nos permita contrastar la independencia de dos variables.

Supongamos que tenemos una muestra de tamaño n de una población para la cual hemos medido dos caracteres representados por las variables aleatorias X e Y, que pueden tomar los valores $x_1, x_2, \ldots, x_m$ la primera de ellas y $y_1, y_2, \ldots, y_k$ la segunda. Denotaremos $O_{ij} = n_{ij}$ la frecuencia observada de la muestra de elementos que tienen conjuntamente $X = x_i$ e $Y = y_j$. Con estas frecuencias observadas podemos construir una tabla de contingencia que represente todas las frecuencias observadas para cualquier par de valores posibles (x_i, y_j), con $i = 1, \ldots m$ y $j = 1, \ldots, k$, de modo que tendremos:

X \ Y	y_1	y_2	$\cdots$	y_k	
x_1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1*}=\sum_{j=1}^{k} n_{1j}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2*}=\sum_{j=1}^{k} n_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	$n_{m*}=\sum_{j=1}^{k} n_{mj}$
	$n_{*1}=\sum_{i=1}^{m} n_{i1}$	$n_{*2}=\sum_{i=1}^{m} n_{i2}$	$\cdots$	$n_{*k}=\sum_{i=1}^{m} n_{ik}$	$n=\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}$

Si suponemos que ambas variables, X e Y, son independientes se cumplirá que la probabilidad conjunta se puede expresar como producto de las probabilidades marginales, $p(X = x_i, Y = y_j) = p_{ij} = p(X = x_i)p(Y = y_j) = p_{i*}\, p_{*j}$. Por lo tanto, el contraste de hipótesis que plantearemos será:

$$\begin{cases} H_0: & X \text{ e } Y \text{ independientes} \Rightarrow p_{ij} = p_{i*}\, p_{*j},\ \forall i = 1, \ldots, m, \forall j = 1, \ldots, k \\ H_1: & X \text{ e } Y \text{ dependientes} \Rightarrow p_{ij} \neq p_{i*}\, p_{*j} \text{ para algún par } (i, j) \end{cases}$$

Teniendo esto en cuenta y dado que las probabilidades marginales se pueden estimar a partir de las frecuencias marginales observadas, de modo que $\hat{p}_{i*} = \dfrac{n_{i*}}{n}$ y $\hat{p}_{*j} = \dfrac{n_{*j}}{n}$, si se satisface la hipótesis nula de independencia de las variables, las frecuencias esperadas serán:

$$E_{ij} = n\hat{p}_{ij} = n\frac{n_{i*}}{n}\frac{n_{*j}}{n} = \frac{n_{i*}\, n_{*j}}{n}$$

Para el contraste de hipótesis de independencia de variables se usa el estadístico de Pearson que vimos anterioremente, $\mathcal{Q}= \displaystyle\sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i*}\, n_{*j}}{n}\right)^2}{\frac{n_{i*}\, n_{*j}}{n}}$, que tomará valores cercanos a cero si se satisface la hipótesis nula de independencia en las variables, ya que en ese caso la frecuencia esperada y la observada serán parecidas. En el caso de que H_0 sea cierta, el estadístico χ^2 sigue una distribución chi-cuadrado con $\nu = (m-1)(k-1)$ grados de libertad. El motivo de que sean estos los grados de libertad es porque, si bien tenemos $m \times k$ frecuencias posibles, hay que tener en cuenta que la suma de las frecuencias por filas y por columnas deben

dar las frecuencias marginales, así que para cada fila y cada columna sólo habrá que calcular $k-1$ y $m-1$ valores, respectivamente. Por este motivo, el número de grados de libertad es $\nu=(m-1)(k-1)$. De este modo, aceptaremos la hipótesis nula con un nivel de significación α si $\mathcal{Q}_{muestral}<\chi^2_{\alpha,\nu}$ y la rechazaremos si $\mathcal{Q}_{muestral}>\chi^2_{\alpha,\nu}$, con $\nu=(m-1)(k-1)$

Al igual que sucedía cuando hacíamos pruebas de bondad de ajuste, tendremos que tener en cuenta que el estadístico de Pearson $\mathcal{Q}$ sigue una distribución χ^2 si las frecuencias esperadas son superiores a 5. En caso contrario, habrá que agrupar varias clases para conseguir que todas las frecuencias esperadas sean mayores o iguales a 5.

Por último, conviene señalar que dado que las frecuencias observadas son discretas, el estadístico $\mathcal{Q}$ tomará valores discretos, pero su distribución la estamos aproximando a una χ^2, que es una función continua. Esta aproximación puede introducir cierto error en el análisis de independencia de las variables, aunque, en general, sólo tendrá importancia cuando se tengan tablas de contingencia con pocas clases y con frecuencias esperadas pequeñas. Para solucionar esto se suele aplicar una corrección de continuidad, *corrección de continuidad de Yates*, que consiste en disminuir las diferencias entre las frecuencias observadas y las esperadas en una cantidad 0.5. Así el estadístico de Pearson se definiría, aplicando la corrección de Yates, como $\mathcal{Q}=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}}$. En la práctica, dicha corrección sólo se aplica a tablas de contingencia de dimensión 2×2 y cuando las frecuencias esperadas están entre 5 y 10.

Ejemplo 4:

Consideremos un estudio en el cual se quiere analizar si existe relación entre el sexo de una persona y su color de pelo. Para ello se analiza una muestra de $n=100$ personas, obteniéndose los siguientes resultados:

Sexo \ *Color de pelo*	Calvo	Rubio	Pelirrojo	Moreno
Hombre	5	10	5	20
Mujer	0	18	12	30

En primer lugar calculamos las frecuencias marginales y, a partir de ahí, las frecuencias esperadas si se satisface la hipótesis nula de independencia de las variables, obteniéndose la siguiente tabla:

Observado / *Esperado*	Calvo	Rubio	Pelirrojo	Moreno	*Marginales*
Hombre	5 / 2	10 / 11.2	5 / 6.8	20 / 20	40
Mujer	0 / 3	18 / 16.8	12 / 10.2	30 / 30	60
Marginales	5	28	17	50	100

Como las frecuencias esperadas para la primera de la clase *Calvo* de la variable aleatoria *Color de pelo* son menores que 5, agruparemos dicha clase con otra, por ejemplo con *Rubio*, obteniéndose la siguiente tabla:

Observado / *Esperado*	Calvo o Rubio	Pelirrojo	Moreno	*Marginales*
Hombre	15 / 13.2	5 / 6.8	20 / 20	40
Mujer	18 / 19.8	12 / 10.2	30 / 30	60
Marginales	33	17	50	100

El valor muestral del estadístico de Pearson será:

$$\mathcal{Q}_{muestral} = \frac{(15-13.2)^2}{13.2} + \frac{(5-6.8)^2}{6.8} + \frac{(20-20)^2}{20} + \frac{(18-19.8)^2}{19.8} + \frac{(12-10.2)^2}{10.2} + \frac{(30-30)^2}{30} = 1.20$$

Dicha variable aleatoria seguirá una distribución χ^2 con $(3-1)(2-1) = 2$ grados de libertad. Como $\chi^2_{0.05,2} = 5.991$ y $\mathcal{Q}_{muestral} = 1.20 < \chi^2_{0.05,2} = 5.991$, concluiremos que se satisface la hipótesis nula de independencia entre el color del pelo y el sexo de la persona.

10.4 Pruebas χ^2 de homogeneidad

Un problema similar al anterior es el contraste de homogeneidad de varias muestras. En el contraste de independencia se medían dos características de la misma muestra y en el contraste de homogeneidad lo que se hace es elegir k muestras y se trata de comprobar si todas pueden pertenecer a una misma población o, al menos, si la función de distribución de la variable observada es la misma en todas las muestras.

Consideremos k muestras de tamaños $n_1, n_2, \ldots, n_k$, respectivamente, para las que medimos una variable aleatoria X que puede tomar los valores $x_1, x_2, \ldots, x_m$. Si denominamos $O_{ij} = n_{ij}$ la frecuencia del valor x_i observada en la muestra j tendremos la siguiente tabla de frecuencias:

Muestras / *Clases*	Muestra 1	Muestra 2	$\cdots$	Muestra k	*Total*
x_1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1*} = \sum_{j=1}^{k} n_{1j}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2*} = \sum_{j=1}^{k} n_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	$n_{m*} = \sum_{j=1}^{k} n_{mj}$
Tamaños muestrales	n_1	n_2	$\cdots$	n_k	$n = \sum_{j=1}^{k} n_j = \sum_{i=1}^{m} n_{i*}$

Si las muestras son homogéneas se cumplirá que la probabilidad de obtener un determinado valor x_i será la misma para todas las muestras, por lo tanto se cumplirá que $p_{i1} = p_{i2} = \cdots = p_{ik} = p_{i*}$, siendo $p_{ij} = p(X = x_i|\text{muestra } j)$. Además dicha probabilidad la podremos estimar utilizando los datos muestrales, de modo que $\hat{p}_{i*} = \dfrac{n_{i*}}{n}$, de modo que el valor estimado de la frecuencia de la variable x_i en la muestra j será:

$$E_{ij} = n_j \hat{p}_{i*} = \frac{n_{i*} n_j}{n}$$

El contraste de hipótesis que plantearemos para evaluar la homogeneidad de las muestras será:

$$\begin{cases} H_0: & \text{Las } k \text{ muestras son homogéneas} \Rightarrow p_{i1} = p_{i2} = \cdots = p_{ik} = p_{i*},\ \forall i = 1, \ldots m \\ H_1: & \text{Las } k \text{ muestras no son homogéneas} \Rightarrow p_{ij} \neq p_{i*} \text{ para alguna muestra } j \text{ y para algún } x_i \end{cases}$$

El estadístico que usaremos en este contraste será, al igual que en los casos anteriores, el estadístico de Pearson $\mathcal{Q} = \sum_{i=1}^{m} \sum_{j=1}^{k} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{m} \sum_{j=1}^{k} \dfrac{\left(n_{ij} - \frac{n_{i*} n_j}{n}\right)^2}{\frac{n_{i*} n_j}{n}}$, que seguirá una distribución chi-cuadrado con $\nu = (m-1)(k-1)$ grados de libertad. La forma de calcular los grados de libertad es la misma que en el caso del contraste de independencia, ya que aunque

inicialmente tenemos una tabla de frecuencias $m \times k$, como en cada columna la suma de las frecuencias debe dar el tamaño muestral y en cada fila la suma de las frecuencias debe ser la frecuencia del correspondiente valor muestral, sólo tendremos $m-1$ valores independientes por columna y $k-1$ valores independientes por fila, así pues el número de grados de libertad será $\nu = (m-1)(k-1)$.

De modo que aceptaremos la hipótesis nula de muestras homogéneas a un nivel de significación α si $\mathcal{Q}_{muestral} < \chi^2_{\alpha,\nu}$ con $\nu = (m-1)(k-1)$, y la rechazaremos en caso contrario.

Al igual que en casos anteriores, si el valor estimado para la frecuencia es menor que 5, agruparemos varios valores muestrales en una misma clase para asegurar que la frecuencia estimada sea mayor o igual que 5.

Ejemplo 5:

Consideremos las siguientes frecuencias en las notas de 4 grupos de primero de la asignatura de estadística:

Grupos / *Notas*	A	B	C	D	*Total*
Nt-Sb	14	5	13	5	37
Ap	26	31	23	10	90
SS	29	30	25	26	110
Tamaño muestral	69	66	61	41	237

Estudiar la homogeniedad de las calificaciones al comparar los distintos grupos.

En primer lugar calculamos los valores estimados de las frecuencias utilizando $E_{ij} = \dfrac{n_{i*}\, n_j}{n}$, siendo n_{i*} la frecuencia total para cada calificación y n_j el tamaño muestral del grupo j. Por lo tanto,

Observado / *Estimado*	A	B	C	D	*Total*
Nt-Sb	14 / 10.8	5 / 10.3	13 / 9.5	5 / 6.4	37
Ap	26 / 26.2	31 / 25.1	23 / 15.6	10 / 15.6	90
SS	29 / 32.0	30 / 30.6	25 / 19.0	26 / 19.0	110
Tamaño muestral	69	66	61	41	237

Calculamos ahora el valor muestral del estadístico de Pearson $\mathcal{Q}_{muestral} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} =$ 11.93.

Dicho estadístico seguirá una distribución χ^2 con $\nu = (3-1)(4-1) = 6$ grados de libertad y como $\chi^2_{0.05,6} = 12.592 > \mathcal{Q}_{muestral}$, aceptaremos la hipótesis nula.

10.5 Otros estadísticos utilizados en contrastes no paramétricos

La prueba χ^2 es utilizada en muchos contrastes no parámetricos tal y como hemos visto en apartados anteriores, sin embargo, no es la única existente. Vamos ahora a enumerar brevemente otros contrastes no paramétricos que pueden resultar de utilidad.

10.5.1 Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov (o test K-S) es un contraste no paramétrico para determinar la bondad de ajuste de una muestra a una distribución teórica. Es válida sólo para variables continuas y tiene la ventaja, frente a la prueba χ^2 de bondad de ajuste, de que se puede aplicar a muestras pequeñas y que su potencia (probabilidad de rechazar H_0 siendo falsa) es mayor que la de la prueba χ^2.

El estadístico que se usa en este contraste es el máximo de la diferencia entre la función de distribución de probabilidad teórica y la de la muestra:

$$D_n = \max_{1\leq i\leq n} |F_{muestra}(x_i) - F_0(x_i)|$$

siendo $F_{muestra}(x_i)$ el valor de función de distribución muestral en el punto x_i de la muestra y $F_0(x_i)$ el correspondiente para el valor teórico de la función de distribución.

Dicho estadístico D_n sigue una distribución deducida por Smirnov en el caso de que la hipótesis nula sea cierta, cuyos valores críticos $D_{n,\alpha}$ están tabulados y dependen del tamaño muestral n. De este modo, aceptaremos la hipótesis nula a un nivel de significación α si $D_n < D_{n,\alpha}$, lo que quiere decir que las diferencias entre el valor teórico y el muestral en la función de distribución son debidas al azar.

10.5.2 Prueba de los rangos con signo de Wilcoxon para dos muestras apareadas

La prueba de los rangos con signo es una alternativa no paramétrica a la prueba de la t de Student para datos apareados. La ventaja es que no exige requisitos para poder aplicarlo.

La hipótesis nula que se plantea es que las medianas de las dos poblaciones de las que proceden las muestras son iguales.

La prueba se basa en calcular las diferencias entre dos muestras apareadas y ordenar de menor a mayor el valor absoluto de dichas diferencias. Si la hipótesis nula es cierta la suma de las posiciones (también llamadas rangos) que ocupan las diferencias negativas debe ser igual a la suma de los rangos de las diferencias positivas.

10.5.3 Prueba U de Mann-Whitney

El contraste de Mann-Whitney sirve para analizar si dos muestras independientes proceden de la misma población. Es una prueba no paramétrica alternativa al contraste de igualdad de medias de dos muestras independientes.

10.5.4 Prueba W de Shapiro-Wilk

Este prueba permite contrastar si una muestra procede de una población normal. La ventaja de este contraste es que se puede aplicar aunque la muestra sea pequeña. Se basa en la comparación de los cuantiles muestrales con los teóricos para una distribución normal.

Capítulo 11

Introducción al Análisis Multivariante

Principios de Análisis de la Varianza. Prueba de Fisher (LSD). Prueba de Bartlett.

11.1 Introducción al análisis multivariante

11.1.1 Conceptos previos

En los capítulos anteriores hemos estudiado los contrastes de hipótesis que permiten comparar dos poblaciones. En particular se analizó el contraste de hipótesis para comparar las medias de dos poblaciones normales, tanto en el caso de que las varianzas poblacionales fuesen iguales, como en el caso de que no lo fuesen. Sin embargo, en determinadas ocasiones interesa comparar si más de dos poblaciones normales, con la misma varianza poblacional, tienen la misma media poblacional o no. Por ejemplo, si se quiere estudiar el efecto de distintos tipos de abono en la producción agrícola de un determinado producto, la eficacia de distintos tratamientos en la curación de una enfermedad o el efecto de distintos tipos de dieta en la pérdida de peso lo que interesa es analizar si hay diferencias que se puedan considerar significativas entre las distintas muestras correspondientes a los diferentes tratamientos aplicados o si dichas diferencias son las esperables en un muestreo aleatorio.

Por lo tanto, el contraste de hipótesis que queremos plantear es:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 = \cdots \\ H_1: & \mu_i \neq \mu_j \text{ para algún par } i, j \end{cases}$$

El motivo por el cual no resulta apropiado presentar un contraste de igualdad de medias dos a dos se explica a continuación. Supongamos que fijamos un nivel de significación α para cada contraste de dos muestras, esto es, una probabilidad de aceptar la hipótesis nula de igualdad de dos medias en el caso de ser cierta de $1 - \alpha$. Si tenemos k muestras habría que realizar

$\binom{k}{2} = \frac{k(k-1)}{2}$ contrastes, por lo que la probabilidad de aceptar $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$ siendo cierta es $(1-\alpha)^{\frac{k(k-1)}{2}}$, suponiendo que las muestras son independientes. Así que rechazaríamos H_0 siendo cierta, esto es, cometeríamos un error de tipo I con una probabilidad $1-(1-\alpha)^{\frac{k(k-1)}{2}}$, que en general es bastante grande. Por ejemplo, si tuviesemos 4 muestras y un nivel de significación $\alpha = 0.05$ para cada contraste de dos muestras, el nivel de significación final de comparación de las cuatro muestras dos a dos sería $1-(1-0.05)^{\frac{4\cdot 3}{2}} = 0.265$, que es demasiado grande para un error de tipo I.

Por este motivo, es necesario plantear un procedimiento que permita contrastar la igualdad de medias de k poblaciones de forma global. Dicho procedimiento se conoce con el nombre de *Análisis de la Varianza* o *ANOVA* (que proviene del inglés, ANalysis Of the VAriance).

11.1.2 Analisis de la Varianza con un factor de variación: descripción.

En esta sección vamos a describir el procedimiento que permite analizar la homogeneidad de varias poblaciones para estudiar si sus medias son iguales, suponiendo que sólo hay un factor que diferencia una población de otra. Como hemos visto anteriormente, un ejemplo podría ser la producción de un determinado producto agrícola en función del tipo de abono que se aplique. También hay otros procedimientos, que no vamos a tratar en este capítulo, que permiten estudiar la igualdad de medias de varias poblaciones cuando hay varios factores de variación, por ejemplo, que en la producción de un determinado producto agrícola se varíe no sólo el abono aplicado, sino también las condiciones de humedad.

El procedimiento de análisis de las k poblaciones con un factor de variación se explica a continuación. Supongamos que tenemos k poblaciones independientes y estudiamos en todos ellos una característica común en la que lo único que diferencia una población de otra es un factor. Suponiendo que la variable aleatoria que estudia esa característica sigue una distribución normal con la misma varianza en todas las poblaciones, lo que diferenciará una población de otra será que tengan distintas medias. Así pues, el contraste que tendremos que plantear es si las medias de dichas poblaciones son iguales o no.

Si suponemos que todas las poblaciones tienen la misma varianza, la estimación de la varianza que obtengamos para cada una de las poblaciones será una buena estimación de dicha varianza común. Además, si suponemos que todas las poblaciones tienen la misma media (esto es, no hay diferencias entre las poblaciones) y estimamos la varianza de la población total (obtenida uniendo todas las poblaciones), ésta también será una buena estimación de la varianza común, tal y como se puede ver en el panel izquierdo de la figura 11.1. Sin embargo, si las distintas poblaciones tiene distinta media (esto es, hay diferencias entre las poblaciones), la varianza estimada a partir de la población total (obtenida uniendo todas las poblaciones) será mayor que la varianza común de las poblaciones individuales, tal y como se puede ver

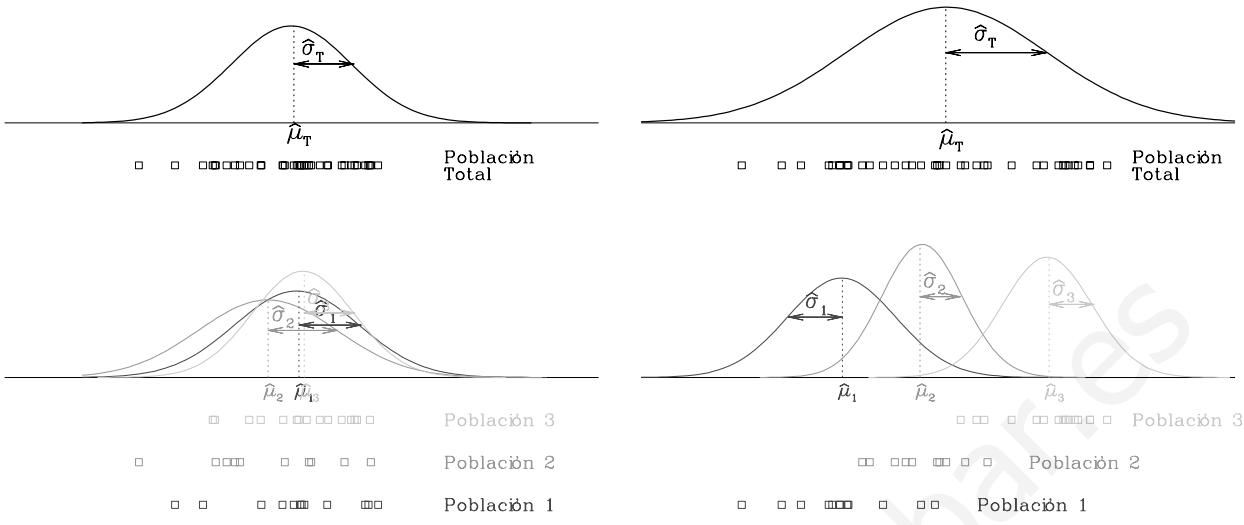


Figura 11.1: *Funciones de densidad de probabilidad estimada para un conjunto de poblaciones en las cuales la variable aleatoria tiene la misma varianza y la misma media en todas las poblaciones (a la izquierda) y para un conjunto de poblaciones en las cuales la variable aleatoria tiene la misma varianza pero diferente media (a la derecha)*

en el panel derecho de la figura 11.1, ya que al haber una dispersión grande de las medias de cada población con respecto a la media de la población total, se *ensancha* la distribución de la población total.

Siguiendo este planteamiento, podemos ver que analizar si varias poblaciones con la misma varianza tienen la misma media es equivalente a comparar la varianza común de la población total con la varianza de las medias de cada población con respecto a la media total. Por este motivo, a este tipo de estudio de igualdad de medias entre varias poblaciones se le denomina Análisis de la Varianza.

Vamos a desarrollar en la siguiente sección el formalismo que nos va a permitir realizar el contraste de igualdad de medias entre varias poblaciones con un factor de variación.

11.2 Principios del análisis de la varianza

Consideremos que tenemos k poblaciones o grupos independientes y supongamos que en dichas poblaciones observamos una característica definida por una variable aleatoria X. En cada grupo la variable X tendrá asociada un valor esperado y una varianza, de modo que μ_i y σ_i^2 serían dichos valores para el grupo i, con $i = 1, \ldots, k$. Supongamos que si existen diferencias de la

variable aleatoria entre los k grupos (también denominados tratamientos) pueden ser debidas a un factor de variación entre los grupos (por ejemplo, utilización de distinto tratamiento médico entre varios grupos para la curación de una enfermedad, uso de diferente abono para la producción agrícola, etc). Por lo tanto, para comprobar dicha hipótesis tendremos que plantear un contraste de hipótesis de homogeneidad entre los grupos de la forma:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: & \mu_i \neq \mu_j \text{ para algún par } i, j \end{cases}$$

Si como consecuencia del contraste rechazamos la hipótesis nula y aceptamos la alternativa, querrá decir que el factor de variación será el que haya originado una diferencia entre los grupos.

Consideremos ahora que en los k grupos extraemos k muestras (una para cada grupo) de tamaños n_1, n_2, ..., n_k, respectivamente. Si representamos por x_{ij} el valor $i-$esimo de la variable aleatoria X en la muestra j, podemos resumir los datos muestrales en la siguiente tabla:

Grupo	*Datos muestrales*	*Tamaño muestral*	*Sumas muestrales*	*Medias muestrales*
1	$x_{11}, x_{21}, \ldots, x_{n_1 1}$	n_1	$T_1 = \sum_{i=1}^{n_1} x_{i1}$	$\hat{\mu}_1 = \frac{T_1}{n_1}$
2	$x_{12}, x_{22}, \ldots, x_{n_2 2}$	n_2	$T_2 = \sum_{i=1}^{n_2} x_{i2}$	$\hat{\mu}_2 = \frac{T_2}{n_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1k}, x_{2k}, \ldots, x_{n_k k}$	n_k	$T_k = \sum_{i=1}^{n_k} x_{ik}$	$\hat{\mu}_k = \frac{T_k}{n_k}$
Total	$\{x_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots k\}$	$n = \sum_{j=1}^{k} n_j$	$T = \sum_{j=1}^{k} T_j$	$\hat{\mu} = \frac{T}{n}$

donde en dicha tabla hemos incluído las sumas de los valores muestrales de la variable X en cada grupo, $T_j = \sum_{i=1}^{n_j} x_{ij}$ para $j = 1, \ldots, k$, y las estimaciones de la media de X en cada grupo, $\hat{\mu}_j = \frac{T_j}{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$ para $j = 1, \ldots, k$, como se puede ver, dichas estimaciones de la media se han obtenido utilizando la media aritmética, $\bar{X}$. También hemos incluído en la tabla estas

cantidades para la muestra total que agrupa los datos de las k poblaciones, $T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$ y $\hat{\mu} = \frac{T}{n} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$ con $n = \sum_{j=1}^{k} n_j$.

El procedimiento de análisis de la varianza, ANOVA, que vamos a explicar aquí se basa en que:

- La variable aleatoria X sigue distribuciones normales para cada uno de los grupo (hipótesis de *normalidad*). Por ello, debemos contrastar inicialmente si las muestras obtenidas para cada población siguen una distribución normal (utilizando alguno de los test de normalidad descritos en el capítulo anterior).

- Las varianzas de la variable aleatoria X en los distintos grupos, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, son iguales, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ (hipótesis de *homocedasticidad*). En la última sección de este capítulo explicaremos cómo contrastar si varias poblaciones normales tienen la misma varianza o no.

Una vez que comprobamos que se cumplen ambos supuestos, el procedimiento de la ANOVA se basa en que la *variación total*, SST, que observamos en el conjunto de los datos muestrales, x_{ij}, o sea, las variaciones de los datos muestrales con respecto a la media de la población total, se pueden separar en dos tipos de variaciones:

- *Variación dentro de los grupos* o *variación debida al error*, SSE: son las variaciones de los elementos de cada muestra con respecto a la media de dicha muestra, que son debidas a las características del proceso aleatorio del muestreo, esto es, debidas al azar.

- *Variación entre los grupos* o *variación entre los tratamientos*, SSG: son las variaciones de los valores medios de la variable X en un grupo dado, $\bar{X}_j$ (cuya estimación puntual es $\hat{\mu}_j$), con respecto a la media total, $\bar{X}$ (cuya estimación puntual es $\hat{\mu}$), que pueden ser debidas o bien a efectos aleatorios o bien a que existen diferencias entre los grupos debidas al factor de variación considerado.

Por lo tanto, por un lado definiremos la variación total de los datos muestrales con respecto a la media global, $\bar{X}$, que representaremos por $SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}\right)^2$, y por otro lado definiremos la variación total dentro de los grupos de los datos muestrales de cada grupo con respecto a la media muestral de dicho grupo, $SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_j\right)^2$, y la variación entre

los grupos de las medias muestrales de cada grupo con respecto a la media muestral total pesadas con el tamaño muestral, $SSG = \sum_{j=1}^{k} n_j \left(\bar{X}_j - \bar{X}\right)^2$. Con estas definiciones, es fácil demostrar que:

$$SST = SSE + SSG$$

ya que $SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(X_{ij} - \bar{X}\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(\left(X_{ij} - \bar{X}_j\right) + \left(\bar{X}_j - \bar{X}\right)\right)^2 =$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(\left(X_{ij} - \bar{X}_j\right)^2 + \left(\bar{X}_j - \bar{X}\right)^2 + 2\left(X_{ij} - \bar{X}_j\right)\left(\bar{X}_j - \bar{X}\right)\right) =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(X_{ij} - \bar{X}_j\right)^2 + \sum_{j=1}^{k} n_j \left(\bar{X}_j - \bar{X}\right)^2 + 2\sum_{j=1}^{k}\left(\bar{X}_j - \bar{X}\right)\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_j\right) = SSE + SSG$$

Estas variaciones que hemos calculado pueden ser útiles, como veremos a continuación, a la hora de hacer estimaciones sobre la varianza de la población total, σ^2, bajo el supuesto de que se satisface la hipótesis nula H_0 de igualdad de medias.

Por un lado, sabemos que un estimador de la varianza de X en el grupo j será la cuasi-varianza muestral, $\mathcal{S}_j^2 = \dfrac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_j\right)^2$, que dará lugar a las estimaciones puntuales $\hat{\sigma}_j^2 = \dfrac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(x_{ij} - \hat{\mu}_j\right)^2$. Si se satisface la hipótesis nula se cumplirá que $\sigma^2 = \sigma_1^2 = \cdots = \sigma_k^2$ ya que las medias poblacionales de los distintos grupos son iguales y, por tanto, cada uno de ellos será un muestreo de una misma población total. Así pues, podremos estimar σ^2 haciendo una media ponderada de las estimaciones de las varianzas de cada grupo pesadas con el número de grados de libertad de cada muestra utilizando el estimador *cuadrado medio del error* definido por:

$$MSE = \frac{\sum_{j=1}^{k}\left(n_j - 1\right)\mathcal{S}_j^2}{\sum_{j=1}^{k}\left(n_j - 1\right)} = \frac{SSE}{n-k}$$

Este estimador permite calcular una estimación puntual para la varianza común a todos los grupos, $\hat{\sigma}_{MSE}^2$. El valor esperado de este estimador, considerando que $\sigma^2 = \sigma_i^2$ para $i = 1, ..., k$, es $E(MSE) = \dfrac{E\left(\sum_{j=1}^{k}(n_j - 1)\mathcal{S}_j^2\right)}{n-k} = \dfrac{\sum_{j=1}^{k}(n_j - 1)E\left(\mathcal{S}_j^2\right)}{n-k} = \dfrac{\sum_{j=1}^{k}(n_j - 1)\sigma_j^2}{n-k} = \sigma^2$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}\right)^2$$

Por otro lado, como $\mathcal{S}^2 = \dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}\right)^2}{n-1} = \dfrac{SST}{n-1}$ es también un estimador de la varianza de la población total, σ^2, y se cumple que $SST = SSE + SSG$, tendremos que $\mathcal{S}^2 = \dfrac{SSE+SSG}{n-1}$, cuyo valor esperado es $E(\mathcal{S}^2) = \sigma^2$. Por lo tanto, como $(n-k+k-1)\,\mathcal{S}^2 = (n-k)\,\mathcal{S}^2 + (k-1)\,\mathcal{S}^2 = SSE + SSG$ se puede deducir que $(k-1)\,\mathcal{S}^2 = SSG$. Vamos a comprobarlo calculando el valor esperado de la expresión $(n-1)\mathcal{S}^2 = SSE + SSG$. Bajo el supuesto de la hipótesis nula sabemos que la varianza total, σ^2, es igual a la varianza de cada tratamiento, σ_i^2 con $i = 1, ..., k$, y como el valor esperado de la varianza de cada tratamiento es $\sigma^2 = E(MSE) = \dfrac{E(SSE)}{n-k}$, tendremos que $E(SSE+SSG) = E\left((n-k)\mathcal{S}^2 + (k-1)\mathcal{S}^2\right)$ $\Rightarrow E(SSE) + E(SSG) = (n-k)E(\mathcal{S}^2) + (k-1)E(\mathcal{S}^2) \Rightarrow (n-k)\sigma^2 + E(SSG) = (n-k)\sigma^2 + (k-1)E(\mathcal{S}^2) \Rightarrow \sigma^2 = \dfrac{E(SSG)}{k-1}$. Así que podemos definir otro estimador para la varianza de la población total denominado *cuadrado medio de los tratamientos* y definido por

$$MSG = \frac{\sum_{j=1}^{k} n_j\left(\bar{X}_j-\bar{X}\right)^2}{k-1} = \frac{SSG}{k-1}$$

que permite obtener una estimación para la varianza de la población total, $\hat{\sigma}^2_{MSG}$, y cuyo valor esperado si es cierta la hipótesis nula será σ^2.

Bajo el supuesto de que es cierta la hipótesis nula de nuestro contraste la varianza de la población total obtenida a partir del cuadrado medio del error, $\sigma^2_{MSE} = E(MSE) = \dfrac{E(SSE)}{n-k}$, y la varianza de la población total obtenida a partir del cuadrado medio de los tratamientos, $\sigma^2_{MSG} = E(MSG) = \dfrac{E(SSG)}{k-1}$, serán iguales. Sin embargo, si la hipótesis nula no es cierta y, por tanto, las medias de los distintos tratamientos son diferentes, dichas medias muestrales presentarán una variación mayor que la que esperaríamos si no hubiese variación entre los tratamientos, por lo que $SSG = \sum_{j=1}^{k} n_j\left(\bar{X}_j-\bar{X}\right)^2$ será mayor en el caso de que la hipótesis nula sea falsa. Como consecuencia de esto, si la hipótesis nula es falsa se cumplirá $\sigma^2_{MSG} > \sigma^2_{MSE}$.

Así pues, reescribiremos el contraste de hipótesis inicial de la siguiente forma:

$$\begin{cases} H_0: & \sigma^2_{MSG} = \sigma^2_{MSE} \Rightarrow \dfrac{\sigma^2_{MSG}}{\sigma^2_{MSE}} = 1 \\ \\ H_1: & \sigma^2_{MSG} > \sigma^2_{MSE} \Rightarrow \dfrac{\sigma^2_{MSG}}{\sigma^2_{MSE}} > 1 \end{cases}$$

Para poder analizar este contraste definiremos el estadístico

$$F = \frac{MSG/\sigma^2_{MSG}}{MSE/\sigma^2_{MSE}}$$

que seguirá una distribución F de Fisher con $k-1$ grados de libertad en el numerador (los correspondientes a MSG) y $n-k$ grados de libertad en el denominador (los correspondientes a MSE), por lo tanto $F \sim F_{(k-1,n-k)}$. Bajo el supuesto de que es cierta la hipótesis nula, dicho estadístico F se escribirá como $F = \dfrac{MSG}{MSE}$, que para una muestra concreta tomará el valor $f_{muestral} = \dfrac{\hat{\sigma}^2_{MSG}}{\hat{\sigma}^2_{MSE}}$.

Dado que se trata de un contraste unilateral derecho, la región de aceptación para dicho contraste para un nivel de significación α será $RA_\alpha = [0, F_{\alpha;k-1,n-k})$, de modo que aceptaremos la hipótesis nula de igualdad de medias entre los k distintos tratamientos si $f_{muestral} < F_{\alpha;k-1,n-k}$.

Para calcular de forma sencilla el estadístico F, podemos tener en cuenta que $MSG = \dfrac{SSG}{k-1}$ con $SSG = \displaystyle\sum_{j=1}^{k} n_j \left(\bar{X}_j - \bar{X}\right)^2$ y $MSE = \dfrac{SSE}{n-k}$ con $SSE = \displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(X_{ij} - \bar{X}_j\right)^2$, lo de que da lugar a las estimaciones puntuales $\hat{\sigma}^2_{MSG}$ y $\hat{\sigma}^2_{MSE}$, respectivamente.

Como los estimadores $\bar{X}$ y $\bar{X}_j$ permiten obtener las estimaciones puntuales de las medias $\hat{\mu} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} = \frac{T}{n}$ y $\hat{\mu}_j = \dfrac{1}{n_j}\displaystyle\sum_{i=1}^{n_j} x_{ij} = \frac{T_j}{n_j}$, respectivamente, si denominamos $\widehat{SSG}$ y $\widehat{SSE}$ a las estimaciones puntuales de las variaciones obtenidas a partir de SSG y SSE, tendremos que:

$$\widehat{SSG} = \sum_{j=1}^{k} n_j \left(\hat{\mu}_j - \hat{\mu}\right)^2 = \sum_{j=1}^{k} n_j \left(\hat{\mu}_j^2 - 2\hat{\mu}_j\hat{\mu} + \hat{\mu}^2\right) = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j} - 2T_j\frac{T}{n} + n_j\frac{T^2}{n^2}\right) = \sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{T^2}{n}$$

y por otro lado:

$$\widehat{SSE} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \hat{\mu}_j\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}^2 - 2x_{ij}\frac{T_j}{n_j} + \frac{T_j^2}{n_j^2}\right) = \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{ij}^2 - \frac{T_j^2}{n_j}\right).$$

Así pues,

$$\hat{\sigma}^2_{MSG} = \frac{\widehat{SSG}}{k-1} = \frac{1}{k-1}\left(\sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{T^2}{n}\right)$$

y

$$\hat{\sigma}^2_{MSE} = \frac{\widehat{SSE}}{n-k} = \frac{1}{n-k}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k}\frac{T_j^2}{n_j}\right)$$

Ejemplo 1:

Se quiere analizar el efecto de tres fertilizantes, A, B y C, en la producción de uva de una cierta variedad. Para ello se suministra dicho fertilizante a varias parcelas del mismo tamaño. El fertilizante de tipo A se aplica en 15 parcelas, el de tipo B en 12 parcelas y el de tipo C en 10 parcelas. Posteriormente se miden las toneladas de producción de uva en cada parcela obteniéndose los siguientes resultados

Fertilizante A	8.2	8.7	8.8	9.5	7.0	7.7	6.7	9.0	7.1	9.2	8.3	8.7	8.9	8.6	9.2
Fertilizante B	4.1	4.5	3.2	5.9	4.2	4.3	4.4	6.2	4.5	2.4	3.6	5.2			
Fertilizante C	5.2	5.8	5.4	4.6	3.9	4.4	3.8	4.7	6.2	5.3					

Analizar si existen diferencias significativas en la producción de uva para los distintos tratamientos. Suponer que las varianzas de cada uno de los tratamientos son iguales.

El contraste de hipótesis que plantearemos será:

$$\begin{cases} H_0: & \mu_A = \mu_B = \mu_C \\ H_1: & \mu_A \neq \mu_B \text{ ó } \mu_A \neq \mu_C \text{ ó } \mu_B \neq \mu_C \end{cases}$$

que, suponiendo que $\sigma_A^2 = \sigma_B^2 = \sigma_C^2$, será equivalente a:

$$\begin{cases} H_0: & \dfrac{\sigma^2_{MSG}}{\sigma^2_{MSE}} = 1 \\ H_1: & \dfrac{\sigma^2_{MSG}}{\sigma^2_{MSE}} > 1 \end{cases}$$

siendo $\hat{\sigma}^2_{MSG} = \dfrac{\widehat{SSG}}{k-1} = \dfrac{1}{k-1}\left(\displaystyle\sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{T^2}{n}\right)$ con $k = 3$ y

$\hat{\sigma}^2_{MSE} = \dfrac{\widehat{SSE}}{n-k} = \dfrac{1}{n-k}\left(\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k}\frac{T_j^2}{n_j}\right)$ con $n = 15 + 12 + 10 = 37$.

Teniendo esto en cuenta construímos las siguiente tabla:

	Fertilizante A	Fertilizante B	Fertilizante C	$TOTAL$
$T_j = \sum_{i=1}^{n_j} x_{ij}$	125.6	52.5	49.3	227.4
$\sum_{i=1}^{n_j} x_{ij}^2$	1062.24	242.05	248.63	1552.92
n_j	15	12	10	37

Por lo tanto tendremos que

$\widehat{SSG} = \frac{125.6^2}{15} + \frac{52.5^2}{12} + \frac{49.3^2}{10} - \frac{227.4^2}{37} = 126.839$

$\widehat{SSE} = 1553.37 - \frac{125.6^2}{15} - \frac{52.5^2}{12} - \frac{49.3^2}{10} = 28.4928$

Así que $\hat{\sigma}^2_{MSG} = \frac{126.839}{3-1} = 63.4195$ y $\hat{\sigma}^2_{MSE} = \frac{28.4928}{37-3} = 0.838$, por lo que $f_{muestral} = \frac{\hat{\sigma}^2_{MSG}}{\hat{\sigma}^2_{MSE}} = 75.680$. Estos resultados se pueden resumir en la siguiente tabla:

	Variaciones	*g.l.*	*Cuadrados medios*	*Estadístico F*
Tratamientos (entre grupos)	126.839	2	63.4195	75.680
Errores (intra grupos)	28.4928	34	0.838	
Totales	155.3318	36		

Como $F \sim F_{(2,34)}$ y $F_{0.05;2,34} = 3.2759$ y $f_{muestral} = 75.680 > F_{0.05;2,34} = 3.2759$ podemos rechazar la hipótesis nula de igualdad de medias. Así pues, hay diferencias entre las producciones de uva dependiendo del fertilizante utilizado.

11.3 Prueba de Fisher (LSD)

Cuando realizamos un contraste de hipótesis como el descrito en el apartado anterior para analizar si una variable aleatoria definida para varios tratamientos tienen la misma media o no, puede ocurrir que rechacemos la hipótesis nula y, como consecuencia de ello, aceptemos que las medias poblacionales de la variable aleatoria para los distintos tratamientos no son iguales. En este caso, tendremos que plantearnos dónde se encuentran dichas diferencias entre las medias, esto es, para qué par de tratamientos hay diferencias significativas en la media de la variable aleatoria. Para poder analizar esto vamos a describir la *prueba LSD* (Least Significant Difference) de Fisher. Dicha prueba se basa en las dos condiciones que impusimos al comienzo

del contraste de la ANOVA, que son la normalidad de la variable para cada tratamiento y la igualdad de varianzas entre los distintos tratamientos.

Para poder analizar qué par (ya sea uno sólo o varios) de tratamientos es el responsable del resultado del contraste de la ANOVA que originó el rechazo de igualdad de medias, tendremos que comparar todos los posibles pares de tratamientos, de modo que si tenemos k tratamientos, habrá que realizar $\frac{k(k-1)}{2}$ contrastes de hipótesis. Los contrastes de hipótesis que tenemos que hacer son:

$$\begin{cases} H_0: & \mu_i = \mu_j \Rightarrow \mu_i - \mu_j = 0 \\ H_1: & \mu_i \neq \mu_j \Rightarrow \mu_i - \mu_j \neq 0 \end{cases}$$

con $i = 1, \ldots, k$; $j = i+1, \ldots, k$.

El estadístico que utilizaremos para el contraste se basa en que tenemos que contrastar una diferencia de medias de variables que siguen una distribución normal, ya que $X_i \sim N(\mu_i, \sigma_i)$ y $X_j \sim N(\mu_j, \sigma_j)$. Además las varianzas poblacionales son desconocidas, pero también sabemos que son iguales, $\sigma_i^2 = \sigma_j^2 = \sigma^2$, y hemos calculado una estimación puntual previamente, utilizando el estadístico MSE, obteniéndose el valor $\hat{\sigma}^2_{MSE} = \dfrac{1}{n-k}\displaystyle\sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{ij}^2 - \frac{T_j^2}{n_j}\right)$. Teniendo en cuenta los contrastes de hipótesis estudiados en temas anteriores para comparar las medias de dos poblaciones normales de varianza desconocida pero igual, el estadístico que utilizaremos para los distintos contrastes será

$$T_{ij} = \frac{\bar{X}_i - \bar{X}_j - (\mu_i - \mu_j)}{\sqrt{MSE\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}}$$

que seguirá una distribución t de Student de $n-k$ grados de libertad. Si aceptamos la hipótesis nula de igualdad de medias, el estadístico del constraste será $T_{ij} = \dfrac{\bar{X}_i - \bar{X}_j}{\sqrt{MSE\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}}$, cuyo valor muestral será $t_{ij,muestral} = \dfrac{\hat{\mu}_i - \hat{\mu}_j}{\sqrt{\hat{\sigma}^2_{MSE}\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}}$.

Como tenemos un contraste bilateral para un estadístico que sigue una distribución de t de Student con $n-k$ grados de libertad, la región de aceptación del estadístico T_{ij} para un nivel de significación α será $RA_\alpha = \left(-t_{\alpha/2,n-k}, t_{\alpha/2,n-k}\right)$, de modo que aceptaremos la hipótesis nula de igualdad de medias entre los tratamientos i y j si $t_{ij,muestral} \in RA_\alpha$ y la rechazaremos en

caso contrario.

Ejemplo 2:

Analizar en el ejemplo 1, cual o cuales de los fertilizantes presenta una diferencia significativa con respecto a los otros, comparando las medias de producción de uva dos a dos, utilizando el método LSD de Fisher.

Según calculamos anteriormente, tenemos que $\hat{\sigma}^2_{MSE} = 0.838$ y

	Fertilizante A	Fertilizante B	Fertilizante C
$T_j = \sum_{i=1}^{n_j} x_{ij}$	125.6	52.5	49.3
n_j	15	12	10
$\hat{\mu}_j$	8.373	4.375	4.93

Como la región de aceptación, $RA_\alpha = \left(-t_{\alpha/2,n-k}, t_{\alpha/2,n-k}\right)$, para $\alpha = 0.05$ es $RA_{0.05} = (-t_{0.025,34}, t_{0.025,34}) = (-2.0323, 2.0323)$, si calculamos los valores muestrales del estadístico $T_{ij} = \dfrac{\bar{X}_i - \bar{X}_j}{\sqrt{MSE\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}}$ para cada par de fertilizantes, podemos analizar las diferencias entre los distintos tratamientos, obteniéndose:

- $t_{12,muestral} = \dfrac{8.373 - 4.375}{\sqrt{0.838\left(\dfrac{1}{15} + \dfrac{1}{12}\right)}} = 11.277$. Por lo tanto $t_{12,muestral} \notin RA_{0.05}$ y rechazaremos que el efecto del fertilizante A y del fertilizante B sea el mismo.
- $t_{13,muestral} = \dfrac{8.373 - 4.93}{\sqrt{0.838\left(\dfrac{1}{15} + \dfrac{1}{10}\right)}} = 9.213$. Por lo tanto $t_{13,muestral} \notin RA_{0.05}$ y rechazaremos que el efecto del fertilizante A y del fertilizante C sea el mismo.
- $t_{23,muestral} = \dfrac{4.375 - 4.93}{\sqrt{0.838\left(\dfrac{1}{12} + \dfrac{1}{10}\right)}} = -1.262$. Por lo tanto $t_{23,muestral} \in RA_{0.05}$ y aceptaremos que el efecto del fertilizante B y del fertilizante C es el mismo.

Así pues, concluiremos que el fertilizante A es estadísticamente más efectivo (su producción media es mayor) que los otros dos, mientras que los fertilizantes B y C tienen el mismo efecto.

11.4 Prueba de Bartlett

En los apartados anteriores hemos analizado varias poblaciones, para las que hemos supuesto que seguían distribuciones normales con la misma varianza. Por lo tanto, como paso previo a los análisis anteriores debemos comprobar si cada una de las variables sigue una distribución normal (por ejemplo, utilizando un contraste χ^2, o un test de Kolmogorov-Smirnov, o una prueba W de Shapiro-Wilk) y además contrastando que se cumple la homocedasticidad (todas las varianzas pueden considerarse iguales). Para realizar este último contraste de forma global sobre las k poblaciones, vamos a describir la *prueba de Bartlett*. Dicha prueba permite contrastar de forma global si varias variables aleatorias que siguen distribuciones normales tienen la misma varianza.

Supongamos que tenemos k variables aleatorias independientes que siguen distribuciones normales, $X_j \sim N(\mu_j, \sigma_j)$ con $j = 1, \ldots, k$, cuyos tamaños muestrales son n_j, con $j = 1, \ldots, k$, respectivamente. Si queremos analizar si se cumple homocedasticidad entre todas las variables, el contraste de hipótesis que tenemos que plantear es:

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \\ H_1: & \sigma_i^2 \neq \sigma_j^2 \text{ para algún par } i, j \end{cases}$$

Consideremos el estimador cuasivarianza muestral, $\mathcal{S}_j^2 = \dfrac{1}{n_j - 1} \displaystyle\sum_{j=1}^{n_j} \left(X_{ij} - \bar{X}_j\right)^2$ con $j = 1, \cdots, k$, para obtener estimaciones puntuales, $\hat{\sigma}_j^2$, de las varianzas de cada población. Podemos igualmente considerar el estimador de la varianza común, obtenido como media ponderada de las cuasivarianzas muestrales, $MSE = \dfrac{\displaystyle\sum_{j=1}^{j} (n_j - 1)\, \mathcal{S}_j^2}{\displaystyle\sum_{j=1}^{k} (n_j - 1)} = \dfrac{SSE}{n-k}$ con $SSE = \displaystyle\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{ij} - \bar{X}_j\right)^2$, cuya estimación puntual de la varianza común denotaremos por $\hat{\sigma}_{MSE}^2$.

Teniendo esto en cuenta, definiremos el estadístico $V = (n-k) \ln MSE - \displaystyle\sum_{j=1}^{k} (n_j - 1) \ln \mathcal{S}_j^2$, que será cercano a cero si las varianzas son todas iguales, esto es, si la hipótesis nula es cierta. Además, Bartlett demostró que si se define $\chi_B^2 = \dfrac{V}{C}$ con $C = 1 + \dfrac{1}{3(k-1)} \left(\displaystyle\sum_{j=1}^{k} \frac{1}{n_j - 1} - \frac{1}{n-k} \right)$, dicho estadístico sigue una distribución χ^2 con $k-1$ grados de libertad. Como se trata de un contraste unilateral, dado que a mayor valor de χ_B^2 más se alejará V de cero y, por tanto, mayores serán las diferencias entre las varianzas, la región de aceptación de la hipótesis nula de

homocedasticidad entre las variables para el estadístico χ^2_B con un nivel de aceptación α será $RA_\alpha = \left[0, \chi^2_{\alpha,k-1}\right)$. De este modo, aceptaremos que las varianzas son iguales si $\chi^2_{B,muestral} \in RA_\alpha$ y la rechazaremos en caso contrario.

Ejemplo 3:

Considerar los datos del ejemplo 1 y analizar si se cumple la hipótesis de homocedasticidad.

Según los cálculos realizados en el ejemplo 1 tenemos que

	Fertilizante A	Fertilizante B	Fertilizante C
$T_j = \sum_{i=1}^{n_j} x_{ij}$	125.6	52.5	49.3
$\sum_{i=1}^{n_j} x_{ij}^2$	1062.24	242.05	248.63
n_j	15	12	10

Por lo tanto, podemos estimar las varianzas individuales considerando el estimador cuasi-varianza muestral $\mathcal{S}_j^2 = \dfrac{1}{n_j - 1} \sum_{j=1}^{n_j} \left(X_{ij} - \bar{X}_j\right)^2 = \dfrac{1}{n_j - 1} \left(\sum_{j=1}^{n_j} X_{ij}^2 - \dfrac{\left(\sum_{j=1}^{n_j} X_{ij}\right)^2}{n_j} \right)$, por lo que obtenemos las estimaciones puntuales $\hat{\sigma}_j^2 = \dfrac{1}{n_j - 1} \left(\sum_{i=1}^{n_j} x_{ij}^2 - \dfrac{T_j^2}{n_j} \right)$, así que podemos completar la tabla anterior escribiendo las estimaciones puntuales para la varianza poblacional de cada tratamiento.

	Fertilizante A	Fertilizante B	Fertilizante C
$T_j = \sum_{i=1}^{n_j} x_{ij}$	125.6	52.5	49.3
$\sum_{i=1}^{n_j} x_{ij}^2$	1062.24	242.05	248.63
n_j	15	12	10
$\hat{\sigma}_j^2$	0.7535	1.1239	0.6201

Además, sabemos que $\hat{\sigma}^2_{MSE} = 0.838$ ya que lo calculamos anteriormente, por lo que, como el estadístico de Bartlett se define como $\chi^2_B = \frac{V}{C}$ con $V = (n-k)\ln MSE - \sum_{j=1}^{k}(n_j - 1)\ln \mathcal{S}_j^2$ y $C = 1 + \frac{1}{3(k-1)}\left(\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{n-k}\right)$, tendremos que $V_{muestral} = 34\ln 0.838 - 14\ln 0.7535 - 11\ln 1.1239 - 9\ln 0.6201 = 0.9693$ y $C_{muestral} = 1 + \frac{1}{3\cdot 2}\left(\frac{1}{14} + \frac{1}{11} + \frac{1}{9} - \frac{1}{34}\right) = 1.0407$, obteniéndose que $\chi^2_{B,muestral} = \frac{0.9693}{1.0407} = 0.9314$.

Como la variable aleatoria χ^2_B sigue una distribución χ^2 con $k-1 = 2$ grados de libertad, la región de aceptación de la hipotesis nula para la variable χ^2_B con un nivel de significación $\alpha = 0.05$ es $RA_{0.05} = \left[0, \chi^2_{0.05,2}\right) = [0, 5.99146)$. Como $\chi^2_{B,muestral} = 0.9314 \in RA_{0.05}$ aceptaremos la hipótesis de homocedasticidad.

Igualmente, si contrastásemos cada uno de los tratamientos para comprobar si los datos muestrales pueden provenir de una distribución normal, podríamos comprobar que es cierta la hipótesis nula de normalidad para todas las variables. De este modo, los análisis realizados en los ejemplos 1 y 2 de igualdad de medias entre los tratamientos son válidos porque se satisface las hipótesis de igualdad de varianzas entre los tratamientos y de normalidad de las variables.

Capítulo 12

Introducción a la regresión
Regresión Lineal Simple. Estimación y contrastes sobre los coeficientes de regresión.

12.1 Introducción

En este tema vamos a estudiar como realizar un contraste para analizar si dos variables aleatorias X e Y están relacionadas entre sí y, en caso de estarlo, en qué sentido es dicha relación, esto es, si al aumentar los valores de X aumentan también los de Y o si, por el contrario, disminuyen. Otro de los objetivos de este análisis es poder predecir el valor de Y para un valor concreto de la variable X.

Dadas las dos variables X e Y, se suele denominar variable independiente (o explicativa) a la que resulta más fácil medir o controlar y, normalmente, suele corresponder a la variable X. A la otra variable, generalmente denotada por Y, se la denomina variable dependiente (o respuesta). La forma general de proceder para obtener una muestra de dichas variables aleatorias se describe a continuación. En primer lugar se eligen n valores de la variable independiente, $x_1, x_2, \ldots, x_n$. Como dichos valores son elegidos por el investigador y se supone que tienen un error de medida despreciable, no se consideran variables aleatorias y se toman como constantes. Posteriormente se mide para cada uno de los valores x_i un valor de la variable Y. Dicha variable dependiente Y tiene una distribución de probabilidad asociada, por lo que el valor obtenido, y_i, para un x_i dado variará de un muestreo a otro siguiendo la distribución de probabilidad de Y, que tendrá su correspondiente media, $\mu_{Y|X=x_i}$, y varianza, $\sigma^2_{Y|X=x_i}$, que dependen del valor de X que se considere. Por lo tanto, obtenemos finalmente n pares de datos (x_i, y_i), lo que da lugar a una muestra de tamaño n de la variable aleatoria bidimensional (X, Y).

La hipótesis que vamos a plantear es que la media, $\mu_{Y|X=x_i}$, de la distribución de la variable aleatoria Y está relacionada con el valor concreto x_i de la variable X, de modo que podemos escribir $\mu_{Y|X=x} = f(x)$ y hablaremos de regresión de Y sobre X, que nos permitirá hacer

predicciones de Y para un valor concreto de X. En este capítulo se va considerar que dicha relación es lineal, esto es, $\mu_{Y|X=x_i} = \alpha + \beta x_i$, lo que se conoce como modelo de regresión lineal simple. Nuestro objetivo será, en este caso, estimar los parámetros poblacionales α y β a partir de nuestra muestra $\{(x_i, y_i), i = 1, \ldots, n\}$. De forma general se puede considerar que la relación entre X e Y es lineal (como en nuestro caso), cuadrática, cúbica, logarítmica, exponencial, etc.

12.2 Regresión Lineal Simple

12.2.1 Introducción

Tal y como hemos mencionado en el apartado anterior, vamos considerar una muestra de tamaño n de una variable aleatoria bidimensional (X, Y), donde X es la variable independiente e Y la variable dependiente. Los valores muestrales de la variable independiente, $x_1, x_2, \ldots, x_n$, no se consideran variables aleatorias ya que se considera que tienen un error de medida despreciable y son elegidos por el investigador. Sin embargo, para un valor concreto de $X = x_i$ el valor muestral de Y, que será y_i, variará siguiendo una distribución de media $\mu_{Y|X=x_i}$ y de varianza $\sigma^2_{Y|X=x_i}$. En el caso de considerar un modelo de regresión lineal simple lo que estaremos asumiendo es que existe una relación lineal entre la media $\mu_{Y|X=x}$ y el valor concreto que toma la variable $X = x$, de modo que tendremos $\mu_{Y|X=x} = \alpha + \beta x$, donde α y β son dos parámetros a determinar. La recta $\mu_{Y|X=x} = \alpha + \beta x$ se denomina recta de regresión de Y sobre X y representa la recta que mejor se ajusta a los n valores muestrales (x_i, y_i), $i = 1, \ldots, n$. Gráficamente, α es la ordenada en el origen de la recta de regresión y β su pendiente.

12.2.2 Cálculo de la recta de regresión de Y sobre X

Para poder calcular la recta de regresión $\mu_{Y|X=x} = \alpha + \beta x$ asociada a una muestra de n pares (x_i, y_i), $i = 1, \ldots, n$, vamos a suponer que se cumplen las siguientes condiciones:

- Los valores de X son elegidos por el investigador y tiene error de medida despreciable, por lo que no se consideran valores aleatorios.
- Normalidad: Para cada valor x_i la variable aleatoria $Y_i = Y|_{X=x_i}$ sigue una distribución normal de media $\mu_{Y|X=x_i}$ y varianza $\sigma^2_{Y|X=x_i} = \sigma^2_i$.
- Linealidad: Existe una relación lineal entre la media de la distribución $\mu_{Y|X=x}$ y el valor concreto que toma la variable $X = x_i$, de modo que tendremos $\mu_{Y|X=x_i} = \alpha + \beta x_i$
- Independencia: Las variables aleatorias Y_i son independientes entre sí, $\forall i = 1, \ldots, n$

- Homocedasticidad: Las varianzas de todas las variables aleatorias Y_i son iguales, por lo tanto, $\sigma_i^2 = \sigma^2$ para $i = 1, \ldots, n$

Según lo considerado anteriormente, como las variables aleatorias Y_i siguen distribuciones normales de media $\alpha + \beta x_i$ y de varianza σ^2, esto es, $Y_i \sim N(\alpha + \beta x_i, \sigma)$, podemos definir la variable aleatoria $\delta_i = Y_i - (\alpha + \beta x_i)$, denominada residuo, que seguirá una distribución normal de media 0 y varianza σ^2, por lo tanto, $\delta_i \sim N(0, \sigma)$.

Para cada valor x_i el valor muestral de la variable δ_i representa la diferencia entre el valor de y_i que se observa y el correspondiente valor de la recta de regresión $\alpha + \beta x_i$, por lo tanto, es la distancia vertical entre la recta de regresión y el valor observado y_i para un valor dado x_i. Como necesitamos estimar los parámetros α y β se puede demostrar, utilizando el método de máxima verosimilitud, que dichas estimaciones se obtienen al minimizar las distancias muestrales δ_i mencionadas anteriormente. Si denominamos a y b a los estimadores puntuales de α y β, respectivamente, y consideraremos la expresión $SSE = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$ que representa la suma del cuadrado de las distancias verticales entre los datos muestrales de la variable Y y la recta de ajuste (suma de los cuadrados de los residuos), y buscaremos los valores de $\hat{\alpha}$ y $\hat{\beta}$ que minimizan SSE. Dichos valores de $\hat{\alpha}$ y $\hat{\beta}$ serán las estimaciones puntuales de los parámetros de la recta de regresión. Por lo tanto, tendremos que resolver el sistema de ecuaciones:

$$\left.\begin{array}{l} \left.\dfrac{\partial \sum_{i=1}^{n} \delta_i^2}{\partial a}\right|_{muestra} = 0 \\ \left.\dfrac{\partial \sum_{i=1}^{n} \delta_i^2}{\partial b}\right|_{muestra} = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} \sum_{i=1}^{n} \hat{\delta}_i = 0 \\ \sum_{i=1}^{n} x_i \hat{\delta}_i = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) = 0 \\ \sum_{i=1}^{n} \left(y_i x_i - \hat{\alpha} x_i - \hat{\beta} x_i^2\right) = 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \left\{ \begin{array}{l} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \dfrac{1}{n}\sum_{i=1}^{n} y_i x_i = \hat{\alpha}\bar{x} + \hat{\beta}\dfrac{1}{n}\sum_{i=1}^{n} x_i^2 \end{array} \right.$$

donde $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ y $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$.

Definiendo $s_{XY} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$, que representa la covarianza muestral de X e Y, y recordando que $s_X^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ es la varianza muestral de X (no la cuasivarianza muestral), tendremos que

$$\begin{cases} \hat{\beta} = \dfrac{s_{XY}}{s_X^2} \\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \end{cases}$$

Si además queremos estimar σ^2, que es la varianza común a las variables aleatorias Y_i, como sabemos que $\delta_i \sim N(0,\sigma)$, podremos calcular dicha estimación utilizando los valores muestrales de δ_i y calculando una estimación puntual de la varianza de δ_i, que coincidirá con la estimación puntual de la varianza de Y_i. Como a partir de los datos muestrales hemos calculado estimaciones puntuales de α y de β, tendremos $n-2$ grados de libertad, por lo que el estimador de σ^2 será $\mathcal{S}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n} \delta_i^2$ y la estimación puntual de la varianza es

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \left(\hat{\alpha} + \hat{\beta}x_i\right)\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i\right)^2 =$$

$$= \frac{1}{n-2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 + \hat{\beta}^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) - 2\hat{\beta}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)\right) =$$

$$= \frac{n}{n-2}\left(s_Y^2 + \hat{\beta}^2 s_X^2 - 2\hat{\beta}s_{XY}\right) = \frac{n}{n-2}\left(s_Y^2 + \hat{\beta}\frac{s_{XY}}{s_X^2}s_X^2 - 2\hat{\beta}s_{XY}\right) \Rightarrow$$

$$\Rightarrow \hat{\sigma}^2 = \frac{n}{n-2}\left(s_Y^2 - \hat{\beta}s_{XY}\right) = \frac{n}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right)$$

donde $s_Y^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$ es la varianza muestral de Y.

Teniendo esto en cuenta y suponiendo que es cierta nuestra hipótesis de existe una relación lineal entre las variables X e Y, se puede demostrar que:

- El estadístico $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ sigue una distribución normal de media $\alpha + \beta\bar{x}$ y de varianza $\dfrac{\sigma^2}{n}$, por lo tanto $\bar{Y} \sim N\left(\alpha + \beta\bar{x}, \dfrac{\sigma}{\sqrt{n}}\right)$

- El estadístico $b = \dfrac{S_{XY}}{s_X^2}$ sigue una distribución normal de media β y de varianza $\dfrac{\sigma^2}{n\,s_X^2}$,

así que $b \sim N\left(\beta, \frac{\sigma}{s_X\sqrt{n}}\right)$. Como se puede ver, la varianza del estadístico b, que permite estimar la pendiente de la recta, es más pequeña y, por lo tanto, la estimación de la pendiente es más precisa, cuando la dispersión de la variable X (dada por s_X^2) es más grande. Eso significa que cuanto mayor sea el rango de X en el cual se han tomado medidas, más precisa será la estimación de la pendiente de la recta de regresión.

- El estadístico $a = \bar{Y} - b\bar{x}$ sigue una distribución normal de media α y de varianza $\frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_X^2}\right)$, por lo que $a \sim N\left(\alpha, \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{\bar{x}^2}{s_X^2}}\right)$

- El estadístico $\frac{(n-2)\,\mathcal{S}^2}{\sigma^2}$ sigue una distribución chi-cuadrado con $n-2$ grados de libertad, o sea, $\frac{(n-2)\,\mathcal{S}^2}{\sigma^2} \sim \chi^2_{n-2}$

12.2.3 Recta de regresión de X sobre Y

Es importante resaltar que en los cálculos realizados en la sección anterior hemos considerado que X es la variable independiente e Y la dependiente, lo que ha dado lugar a la recta de ajuste $y - \bar{y} = \hat{\beta}_{Y_X}(x - \bar{x})$ con $\hat{\beta}_{Y_X} = \frac{s_{XY}}{s_X^2}$. Sin embargo, si hubieramos considerado que Y es la variable independiente y X la variable dependiente, la recta de regresión obtenida, denominada recta de regresión de X sobre Y, sería $x - \bar{x} = \hat{\beta}_{X_Y}(y - \bar{y})$ con $\hat{\beta}_{X_Y} = \frac{s_{XY}}{s_Y^2}$.

La recta de regresión de Y sobre X y la recta de regresión de X sobre Y son diferentes, ya que, en general, no se cumplirá que $\frac{s_{XY}}{s_X^2} = \frac{s_Y^2}{s_{XY}}$. Sin embargo el signo de sus pendientes (dado por s_{XY}) es el mismo, esto es, o bien ambas son crecientes o bien ambas son decrecientes. Como veremos más adelante, cuanto menor sea el ángulo entre estas dos pendientes mayor será la relación lineal entre X e Y.

Ejemplo 1:

Un investigador quiere averiguar si existe una correlación lineal entre la edad en años, X, de un tipo de árboles y el diámetro en centímetros de su tronco, Y. Para ello obtiene una muestra con los siguientes datos:

Edad	18	13	12	10	15	8	12	9	8	7	14	4	6	10	11	16	15
Diámetro	14.7	12.7	10.9	12.5	13.5	12.6	14.7	10.5	9.6	9.7	15.3	8.5	11.2	15.7	10.8	14.3	16.0

Calcular la recta de regresión del diámetro sobre la edad y la de la edad sobre el diámetro.

Si denominamos X a la variable edad (en años) e Y al diámetro (en cm), calcularemos primero la recta de regresión de Y sobre X, esto es, $y = \alpha_{Y_X} + \beta_{Y_X} x$. Las estimaciones de los parámetros α_{Y_X} y β_{Y_X} son $\hat{\beta}_{Y_X} = \dfrac{s_{XY}}{s_X^2}$ y $\hat{\alpha}_{Y_X} = \bar{y} - \hat{\beta}_{Y_X}\bar{x}$.

Teniendo en cuenta que $T_x = \sum_{i=1}^{17} x_i = 188$, $T_y = \sum_{i=1}^{17} y_i = 213.2$, $SSX = \sum_{i=1}^{17} x_i^2 = 2314$, $SSY = \sum_{i=1}^{17} y_i^2 = 2761.44$, $SXY = \sum_{i=1}^{17} x_i y_i = 2464.4$ y $n = 17$, tendremos que $s_{XY} = \dfrac{1}{n}\left(SXY - \dfrac{1}{n}T_x T_y\right) = 6.274$, $s_X^2 = \dfrac{1}{n}\left(SSX - \dfrac{1}{n}T_x^2\right) = 13.82$, $\bar{x} = \dfrac{1}{n}T_x = 11.06$ y $\bar{y} = \dfrac{1}{n}T_y = 12.54$.

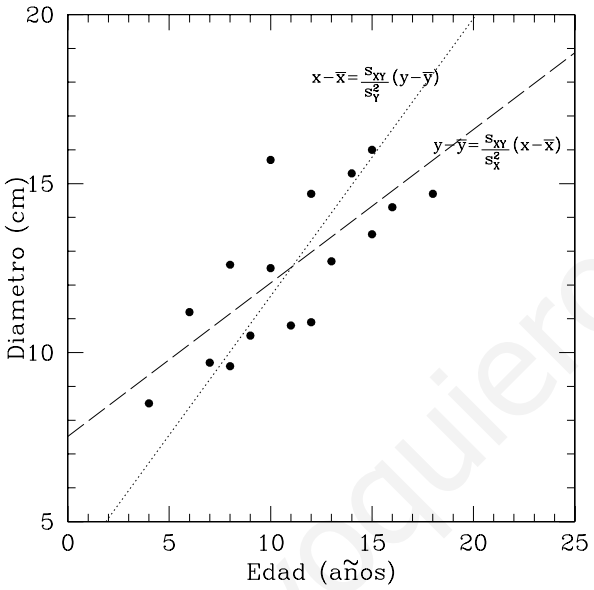


Figura 12.1*: Rectas de regresión de Y sobre X (línea de trazos) y de X sobre Y (línea de puntos) para los datos del Ejemplo 1.*

Con estos datos tendremos que $\hat{\beta}_{Y_X} = 0.45398$ y $\hat{\alpha}_{Y_X} = 7.52068$, por lo que la recta de regresión de Y sobre X será $y = 7.52068 + 0.45398\,x$, y corresponde a la recta de trazos de la Figura 12.1

Calculamos ahora la recta de regresión de X sobre Y, que será de la forma $x = \alpha_{X_Y} + \beta_{X_Y} y$. Las estimaciones de los parámetros α_{X_Y} y β_{X_Y} son $\hat{\beta}_{X_Y} = \dfrac{s_{XY}}{s_Y^2}$ y $\hat{\alpha}_{X_Y} = \bar{x} - \hat{\beta}_{X_Y}\bar{y}$.

Sabiendo que $s_{XY} = 6.274$ y $s_Y^2 = \dfrac{1}{n}\left(SSY - \dfrac{1}{n}T_y^2\right) = 5.157$ tenemos que $\hat{\beta}_{X_Y} = 1.21672$ y $\hat{\alpha}_{X_Y} = -4.20024$, por lo que la recta de regresión de X sobre Y será $x = -4.20024 + 1.21672\,y$, que está dibujada con una línea de puntos en la Figura 12.1.

12.2.4 Correlación lineal: Coeficientes de correlación y de determinación lineal

Después de haber calculado la recta de regresión entre dos variables podemos plantearnos cuál es la *correlación lineal* entre dichas variables, esto es, el grado de asociación o dependencia de

una de las variables con respecto a la otra o, lo que es lo mismo, el grado de dispersión de los datos muestrales con respecto a la recta de regresión. Conocer la correlación lineal entre ambas variables nos permite evaluar la utilidad de la recta de regresión a la hora de hacer predicciones de una variable a partir de la otra. Además diremos que existe una correlación positiva cuando al aumentar el valor de una de las variables la otra también aumente (en cuyo caso la pendiente de la recta de regresión será positiva), mientras que la correlación será negativa cuando el aumento de una de las variables suponga la disminución de la otra (la recta de regresión tendrá pendiente negativa).

Supongamos ahora que tenemos dos variables aleatorias X e Y, donde X es la variable independiente e Y la dependiente, para las cuales hemos calculado la recta de regresión de Y sobre X, esto es, $y - \bar{y} = \dfrac{s_{XY}}{s_X^2}(x - \bar{x})$. Una manera de cuantificar el ajuste de los datos muestrales a dicha recta de regresión es mediante el *coeficiente de determinación lineal,* r^2*,* que define la diferencia entre la varianza de los datos muestrales y_i y la varianza de los residuos $\hat{\delta}_i = y_i - \left(\bar{y} + \dfrac{s_{XY}}{s_X^2}(x_i - \bar{x})\right)$ (o sea, las desviaciones entre los valores y_i y la recta de regresión) dividida por la varianza de los datos y_i. Por lo tanto, la estimación puntual de dicho coeficiente de determinación lineal es $r^2 = \dfrac{s_Y^2 - s_\delta^2}{s_Y^2}$. Se puede ver que si el ajuste de los datos muestrales a la recta de regresión es perfecto la varianza de los residuos será $s_\delta^2 = 0$, por lo que el coeficiente de determinación valdrá $r^2 = 1$. Por el contrario, si no existe correlación entre las variables X e Y, la varianza de los datos muestrales y la varianza de los residuos son iguales, por lo que $r^2 = 0$.

Para poder calcular el coeficiente de determinación lineal calcularemos, por un lado, la varianza de los datos muestrales y_i, dada por $s_Y^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2 - \bar{y}^2$, y por otro lado la varianza muestral de δ_i, definida como $s_\delta^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{\delta}_i^2$ (dado que la media de los δ_i es nula), por lo que $s_\delta^2 = s_Y^2 + \left(\dfrac{s_{XY}}{s_X^2}\right)^2 s_X^2 - 2\dfrac{s_{XY}^2}{s_X^2} = s_Y^2 - \dfrac{s_{XY}^2}{s_X^2}$. Así pues, tendremos que:

$$r^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2}$$

Se puede ver que dicho coeficiente de determinación lineal está relacionado con la covarianza, s_{XY}, entre las variables X e Y, que mide el grado de independencia de dichas variables. Cuando la covarianza es 0 las variables son independientes, lo que se traduce en que el coeficiente de determinación es también 0, y significa que no hay correlación entre X e Y.

El coeficiente de determinación r^2 que hemos calculado es la estimación puntual del parámetro poblacional $\rho^2 = \dfrac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$ que definimos en capítulos anteriores.

Igualmente hemos visto con anterioridad que si la recta de regresión de X sobre Y coincide con la de Y sobre X se cumplirá que $\dfrac{s_{XY}}{s_X^2} = \dfrac{s_Y^2}{s_{XY}}$, lo que es equivalente a que $r^2 = 1$ y significa que existe un ajuste perfecto entre los datos muestrales y la recta de regresión.

Además del coeficiente de determinación lineal, se suele definir el coeficiente de *correlación lineal* r como:

$$r = \frac{s_{XY}}{s_X s_Y}$$

que tomará valores entre -1 y 1, y que es la estimación puntual de la correlación poblacional $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$. El coeficiente de correlación lineal cumple que su cuadrado es el coeficiente de determinación lineal, de modo que valdrá: (a) 0 cuando no haya correlación lineal entre las variables; (b) 1 cuando la correlación sea perfecta y positiva; (c) -1 cuando la correlación sea perfecta y negativa; y (d) otros valores dependiendo del grado de ajuste entre los datos muestrales y la recta de regresión.

Si el coeficiente de correlación es bajo, eso no quiere decir que no exista correlación entre las variables. Lo único que significa es que dicha correlación no es lineal (no se ajusta a una recta), pero es posible que exista una correlación de otro tipo (cuadrática, cúbica, ...).

Ejemplo 2:

Calcular una estimación de los coeficientes de correlación y de determinación de las variables X e Y del ejemplo 1.

Como una estimación del coeficiente de correlación de las variables X e Y se obtiene como $r = \dfrac{s_{XY}}{s_X s_Y}$ y, tal y como hemos calculado anteriormente, $s_{XY} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = 6.274$, $s_X^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = 13.82$ y $s_Y^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = 5.157$, tendremos que el coeficiente de correlación es $r = \dfrac{6.274}{\sqrt{13.82 \cdot 5.157}} = 0.743$. Este valor significa que entre las variables X e Y existe una fuerte correlación directa, cuando la variable X crece también crece la variable Y.

Por otro lado, como el coeficiente de determinación lineal de las variables es el cuadrado del coeficiente de correlación, tendremos que $r^2 = 0.743^2 = 0.552$, lo cual significa que un 55.2% de la variación de Y se puede explicar por la variación de X.

12.3 Intervalos de confianza sobre los coeficientes de regresión y sobre las predicciones

Anteriormente, en la sección 12.2.2, hemos descrito como calcular una estimación puntual de los parámetros poblacionales, la pendiente β y la ordenada en el origen α, de una recta de regresión. Sin embargo, es importante no sólo conocer estimaciones puntuales de los parámetros poblacionales, sino también establecer intervalos de confianza para dichos parámetros y para las predicciones realizadas mediante la recta de regresión. En esta sección vamos a explicar como realizar estos cálculos.

12.3.1 Intervalo de confianza sobre la pendiente poblacional β

Según hemos visto en la sección 12.2.2 el estadístico $b = \dfrac{S_{XY}}{s_X^2}$, que sirve de estimador de la pendiente de la recta de regresión de Y sobre X dada por $y = \alpha + \beta x$, sigue una distribución normal de media β y de varianza $\dfrac{\sigma^2}{n\, s_X^2}$, donde σ^2 es la varianza de cada una de las variables aleatorias Y_i (y que es la misma para cada Y_i independientemente del valor de X) y s_X^2 es la varianza muestral de la variable aleatoria X, de modo que $b \sim N\left(\beta, \dfrac{\sigma}{s_X\sqrt{n}}\right)$.

Sin embargo, como no conocemos σ^2 tendremos que estimarlo y para ello utilizaremos la expresión $\mathcal{S}^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}\left(y_i - (a+bx_i)\right)^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\delta_i^2$. Por lo tanto, la variable aleatoria $T = \dfrac{b-\beta}{\mathcal{S}/(s_X\sqrt{n})}$ sigue una distribución t de Student con $n-2$ grados de libertad.

La estimación puntual de σ^2 vimos que es $\hat{\sigma}^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}\hat{\delta}_i^2 \Rightarrow \hat{\sigma}^2 = \dfrac{n}{n-2}\left(s_Y^2 - \dfrac{s_{XY}^2}{s_X^2}\right)$. Por lo tanto, una estimación de la varianza de b será $\hat{\sigma}_b^2 = \dfrac{1}{n-2}\left(\dfrac{s_X^2 s_Y^2 - s_{XY}^2}{s_X^4}\right)$ y el valor muestral del estadístico T será $t = \dfrac{\hat{\beta}-\beta}{\hat{\sigma}/(s_X\sqrt{n})} = \dfrac{\frac{s_{XY}}{s_X^2} - \beta}{\sqrt{\frac{1}{n-2}\left(\frac{s_X^2 s_Y^2 - s_{XY}^2}{s_X^4}\right)}}$.

Tomando un nivel de significación α podemos calcular un intervalo de confianza para el parámetro β, que, teniendo en cuenta las estimaciones puntuales de dicho parámetro, $\hat{\beta} = \dfrac{s_{XY}}{s_X^2}$,

y de la varianza, $\hat{\sigma}^2 = \dfrac{n}{n-2}\left(s_Y^2 - \dfrac{s_{XY}^2}{s_X^2}\right)$, vendrá dado por:

$$IC_{(1-\alpha)\cdot 100\%}(\beta) = \left(\hat{\beta} - t_{\alpha/2,n-2}\frac{\hat{\sigma}}{s_X\sqrt{n}}, \hat{\beta} + t_{\alpha/2,n-2}\frac{\hat{\sigma}}{s_X\sqrt{n}}\right) \Rightarrow$$

$$IC_{(1-\alpha)\cdot 100\%}(\beta) = \left(\frac{s_{XY}}{s_X^2} - t_{\alpha/2,n-2}\sqrt{\frac{1}{n-2}\left(\frac{s_X^2 s_Y^2 - s_{XY}^2}{s_X^4}\right)}, \frac{s_{XY}}{s_X^2} + t_{\alpha/2,n-2}\sqrt{\frac{1}{n-2}\left(\frac{s_X^2 s_Y^2 - s_{XY}^2}{s_X^4}\right)}\right)$$

Ejemplo 3:

Calcular un intervalo de confianza del 95% para la pendiente de la recta de regresión de Y sobre X del ejemplo 1.

En el ejemplo 1 ya obtuvimos una estimación puntual de la pendiente de la recta de regresión de Y sobre X, dada por $\hat{\beta} = \dfrac{s_{XY}}{s_X^2} = 0.45398$.

Además podemos obtener una estimación de la varianza de los residuos, esto es, de la varianza de las variables Y_i, dada por $\hat{\sigma}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\hat{\delta}_i^2 = \dfrac{n}{n-2}\left(s_Y^2 - \dfrac{s_{XY}^2}{s_X^2}\right)$. Como $n = 17$, $s_{XY} = 6.274$, $s_X^2 = 13.82$ y $s_Y^2 = 5.157$, tendremos que $\hat{\sigma}^2 = \dfrac{17}{15}\left(5.157 - \dfrac{6.274^2}{13.82}\right) = 2.617$.

Así pues, un intervalo de confianza del 95% para la pendiente de la recta será:

$$IC_{95\%}(\beta) = \left(\hat{\beta} - t_{0.025,15}\frac{\hat{\sigma}}{s_X\sqrt{17}}, \hat{\beta} + t_{0.025,15}\frac{\hat{\sigma}}{s_X\sqrt{17}}\right)$$

Como $t_{0.025,15} = 2.13145$, tendremos:

$$IC_{95\%}(\beta) = \left(0.45398 - 2.13145\sqrt{\frac{2.617}{17\cdot 13.82}}, 0.45398 + 2.13145\sqrt{\frac{2.617}{17\cdot 13.82}}\right) = (0.229, 0.679)$$

12.3.2 Intervalo de confianza para la ordenada en el origen α

Al igual que hemos calculado en la sección anterior un intervalo de confianza para el parámetro poblacional β, vamos a hacerlo ahora para el parámetro poblacional α (no confundir dicho parámetro con el nivel de significación, si bien se usan para ambas cantidades el símbolo "α").

Según vimos en la sección 12.2.2, el estadístico $a = \bar{Y} - b\bar{x}$ que permite estimar la ordenada en el origen de la recta de regresión de Y sobre X sigue una distribución normal de media α y de varianza $\dfrac{\sigma^2}{n}\left(1 + \dfrac{\bar{x}^2}{s_X^2}\right)$, por lo que $a \sim N\left(\alpha, \dfrac{\sigma}{\sqrt{n}}\sqrt{1 + \dfrac{\bar{x}^2}{s_X^2}}\right)$. Sin embargo, al

igual que pasaba en la sección anterior cuando calculamos un intervalo de confianza para β, por el hecho de no conocer el valor poblacional de la varianza σ^2 de las variables Y_i, tendremos que utilizar una estimación $\hat{\sigma}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\hat{\delta}_i^2 = \dfrac{n}{n-2}\left(s_Y^2 - \dfrac{s_{XY}^2}{s_X^2}\right)$, obtenida a partir del estimador $\mathcal{S}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\delta_i^2$. Por este motivo, tendremos que definir una variable aleatoria $T = \dfrac{a-\alpha}{\dfrac{\mathcal{S}}{\sqrt{n}}\sqrt{1+\dfrac{\bar{x}^2}{s_X^2}}}$ sigue una distribución t de Student con $n-2$ grados de libertad. Por lo tanto, una estimación puntual de la varianza de a será $\hat{\sigma}_a^2 = \dfrac{1}{n-2}\left(\dfrac{s_X^2 s_Y^2 - s_{XY}^2}{s_X^2}\left(1+\dfrac{\bar{x}^2}{s_X^2}\right)\right)$ y como la estimación puntual de α es $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, el valor muestral del estadístico T será

$$t = \frac{\hat{\alpha}-\alpha}{\dfrac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\dfrac{\bar{x}^2}{s_X^2}}} = \frac{\bar{y} - \frac{s_{XY}}{s_X^2}\bar{x} - \alpha}{\sqrt{\frac{1}{n-2}\left(\frac{s_X^2 s_Y^2 - s_{XY}^2}{s_X^2}\right)\left(1+\frac{\bar{x}^2}{s_X^2}\right)}}$$

Un intervalo de confianza del $(1-\alpha)\cdot 100\%$ para el parámetro α, teniendo en cuenta las estimaciones puntuales de dicho parámetro $\hat{\alpha}$ y de la varianza $\hat{\sigma}^2$, vendrá dado por:

$$IC_{(1-\alpha)\cdot 100\%}(\alpha) = \left(\hat{\alpha} - t_{\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}, \hat{\alpha} + t_{\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}\right)$$

Ejemplo 4:

Calcular un intervalo de confianza del 95% para la ordenada en el origen de la recta de regresión de Y sobre X del ejemplo 1.

En el ejemplo 1 ya obtuvimos una estimación puntual de la ordenada en el origen de la recta de regresión de Y sobre X, dada por $\hat{\alpha} = \bar{y} - \dfrac{s_{XY}}{s_X^2}\bar{x} = 7.52068$.

Además en el ejemplo 3 hemos obtenido una estimación de la varianza de los errores, $\hat{\sigma}^2 = 2.617 \Rightarrow \hat{\sigma} = 1.618$.

Así pues, un intervarlo de confianza del 95% para la ordenada en el origen de la recta será, sabiendo que $n = 17$, $IC_{95\%}(\alpha) = \left(\hat{\alpha} - t_{0.025,15}\dfrac{\hat{\sigma}}{\sqrt{17}}\sqrt{1+\dfrac{\bar{x}^2}{s_X^2}}, \hat{\alpha} + t_{0.025,15}\dfrac{\hat{\sigma}}{\sqrt{17}}\sqrt{1+\dfrac{\bar{x}^2}{s_X^2}}\right)$. Como $t_{0.025,15} = 2.13145$, $s_X^2 = 13.82$ y $\bar{x} = 11.06$ tendremos:

$$IC_{95\%}(\alpha) = \left(7.52068 - 2.13145\sqrt{\frac{2.617}{17}\left(1+\frac{11.06^2}{13.82}\right)}, 7.52068 + 2.13145\sqrt{\frac{2.617}{17}\left(1+\frac{11.06^2}{13.82}\right)}\right) \Rightarrow$$

$$IC_{95\%}(\alpha) = (4.956, 10.145)$$

12.3.3 Intervalo de confianza para la estimación del valor medio de Y correspondiente a $X = x_0$

Uno de los objetivos de analizar la correlación entre dos variables X (independiente) e Y (dependiente) y obtener la recta de regresión de Y sobre X es poder hacer predicciones sobre los valores de la variable dependiente para un determinado valor de la variable independiente.

En esta sección vamos a ver cómo podemos obtener un intervalo de confianza para la media de la variable dependiente para un valor dado de la variable independiente $X = x_0$, que no suele coincidir con los valores x_i utilizados para calcular la recta de regresión.

Una estimación puntual de $\mu_{Y|X=x_0}$, que es la media de Y para un valor dado $X = x_0$, la podemos obtener utilizando como estimador la recta de regresión, $Y|_{X=x_0} = a + bx_0$, de modo que la estimación puntual obtenida es $\hat{y}_{x_0} = \hat{\alpha} + \hat{\beta}x_0$. A partir de aquí vamos a tratar de obtener un intervalo de confianza para el valor poblacional $\mu_{Y|X=x_0}$.

En primer lugar sabemos que $Y|_{X=x_0} = a + bx_0 = a + b\bar{x} + b\,(x_0 - \bar{x}) = \bar{Y} + b\,(x_0 - \bar{x})$. Como $\bar{Y}$ y b con variables aleatorias independientes que siguen distribuciones normales dadas por $\bar{Y} \sim N\left(\alpha + \beta\bar{x}, \dfrac{\sigma}{\sqrt{n}}\right)$ y $b \sim N\left(\beta, \dfrac{\sigma}{s_X\sqrt{n}}\right)$, podemos asegurar que $Y|_{X=x_0}$ sigue una distribución normal de media $\mu_{Y|X=x_0} = \alpha + \beta x_0$ y de varianza $\sigma^2_{Y|X=x_0} = \dfrac{\sigma^2}{n}\left(1 + \dfrac{(x_0 - \bar{x})^2}{s_X^2}\right)$, esto es, $Y|_{X=x_0} \sim N\left(\mu_{Y|X=x_0}, \sigma\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{n\,s_X^2}}\right)$. Esto significa que si definimos la variable aleatoria $Z = \dfrac{Y|_{X=x_0} - \mu_{Y|X=x_0}}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{n\,s_X^2}}}$, seguirá una distribución $N(0,1)$. Sin embargo, como no conocemos el parámetro poblacional σ^2 tendremos que estimarlo utilizando el estimador $\mathcal{S}^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}\delta_i^2$, así pues, deberemos usar el estadístico

$$T = \frac{Y|_{X=x_0} - \mu_{Y|X=x_0}}{\mathcal{S}\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{n\,s_X^2}}}$$

que sigue una distribución t de Student con $n-2$ grados de libertad. Teniendo en cuenta este estimador podemos obtener un intervalo de confianza del $(1-\alpha)\cdot 100\%$ para $\mu_{Y|X=x_0}$, que vendrá dado por:

$$IC_{(1-\alpha)\cdot 100\%}(\mu_{Y|X=x_0}) = \left(\hat{y}_{x_0} - t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{n\,s_X^2}}, \hat{y}_{x_0} + t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{n\,s_X^2}}\right)$$

Ejemplo 5:

Teniendo en cuenta los datos del ejemplo 1, calculad un intervalo de confianza del 95% para el diámetro medio de los árboles con 20 años de vida.

La recta de regresión estimada en el ejemplo 1 es $y = 7.52068 + 0.45398\,x$, por lo que una estimación puntual para el diámetro medio de los árboles de 20 años será $\hat{y}_{20} = 7.52068 + 0.45398\cdot 20 = 16.60$ cm.

Si ahora queremos obtener un intervalo de confianza para el valor poblacional del diámetro medio de los árboles de 20 años de vida, tendremos que utilizar los valores $n = 17$, $\bar{x} = 11.06$, $s_X^2 = 13.82$, $t_{0.025,15} = 2.13145$ y $\hat{\sigma} = 1.618$, obtenidos en los ejemplos anteriores, para calcular

$$IC_{95\%}(\mu_{Y|X=20}) = \left(\hat{y}_{20} - t_{0.025,15}\hat{\sigma}\sqrt{\frac{1}{17}+\frac{(20-\bar{x})^2}{17\,s_X^2}}, \hat{y}_{20} + t_{0.025,15}\hat{\sigma}\sqrt{\frac{1}{17}+\frac{(20-\bar{x})^2}{17\,s_X^2}}\right) =$$

$$\left(16.60 - 2.13145\cdot 1.618\sqrt{\frac{1}{17}+\frac{(20-11.06)^2}{17\cdot 13.82}}, 16.60 + 2.13145\cdot 1.618\sqrt{\frac{1}{17}+\frac{(20-11.06)^2}{17\cdot 13.82}}\right) \Rightarrow$$

$$IC_{95\%}(\mu_{Y|X=20}) = (14.42, 18.78)$$

12.3.4 Intervalo de confianza para un valor individual de Y correspondiente a $X = x_0$

Supongamos ahora que queremos obtener un intervalo de confianza del $(1-\alpha)\cdot 100\%$ para un valor individual, y_{x_0}, de la variable dependiente Y correspondiente a un valor de la variable independiente $X = x_0$. Sabemos que los valores individuales y_{x_0} obtenidos para $X = x_0$ se puede escribir como el valor obtenido mediante la recta de regresión poblacional, $y = \alpha + \beta x$, más un residuo δ_{x_0} que sigue una distribución normal centrada en 0 y de varianza σ^2 (tal y como asumimos al principio de este capítulo al considerar que la varianza de los residuos de Y con respecto a la recta de regresión son independientes de X). De este modo, la variable aleatoria $Y_{x_0} = \alpha + \beta x_0 + \delta_{x_0}$ seguirá una distribución normal de media $\alpha + \beta x_0$ y de varianza

σ^2. Sin embargo, no conocemos la recta de regresión poblacional, así que para las estimaciones sobre y_{x_0} tendremos que utilizar las estimaciones sobre la recta de regresión. Según vimos en el apartado anterior, la variable aleatoria obtenida a partir de la recta de regresión estimada es $Y|_{X=x_0} = a + bx_0$, que cumple que $Y|_{X=x_0} \sim N\left(\alpha + \beta x_0, \sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}}\right)$. Por este motivo, si definimos la variable aleatoria $\Delta_{x_0} = Y_{x_0} - Y|_{X=x_0} = (\alpha - a) + (\beta - b)\,x_0 + \delta_{x_0}$, seguirá una distribución normal de media 0 y de varianza $\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}\right)$, así pues $\Delta_{x_0} \sim N\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}}\right)$. La variable aleatoria $Z = \dfrac{\Delta_{x_0}}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}}}$ sigue una distribución $N(0,1)$. Sin embargo, como no conocemos el parámetro σ^2 tendremos que estimarlo utilizando el estimador $\mathcal{S}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\delta_i^2$, y tendremos que definir el estadístico:

$$T = \frac{Y_{x_0} - Y|_{X=x_0}}{\mathcal{S}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}}}$$

que sigue una distribución t de Student con $n-2$ grados de libertad. Teniendo en cuenta este estimador podemos obtener un intervalo de confianza del $(1-\alpha)\cdot 100\%$ para un valor individual de y_{x_0} de la variable Y_{x_0} que vendrá dado por:

$$IC_{(1-\alpha)\cdot 100\%}(y_{x_0}) = \left(\hat{y}_{x_0} - t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}}, \hat{y}_{x_0} + t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{n\,s_X^2}}\right)$$

donde $\hat{y}_{x_0} = \hat{\alpha} + \hat{\beta}x_0$ es la estimación puntual dada por la recta de regresión estimada y $\hat{\sigma}$ es la estimación puntual de la varianza de los residuos σ.

Ejemplo 6:

Teniendo en cuenta los datos del ejemplo 1, calculad un intervalo de confianza del 95% para el diámetro de un árbol con 20 años de vida.

En este ejemplo tenemos que calcular un intervalo de confianza para un valor individual, no para la media, como en el caso anterior.

El procedimiento es similar, lo único que tendremos que cambiar el valor de la varianza de la estimación individual. Por lo tanto, utilizaremos la recta de regresión estimada en el ejemplo

1, $y = 7.52068 + 0.45398\,x$, para obtener una estimación puntual para el diámetro de un árbol de 20 años, que será $\hat{y}_{20} = 7.52068 + 0.45398 \cdot 20 = 16.60$ cm.

Utilizando los valores $n = 17$, $\bar{x} = 11.06$, $s_X^2 = 13.82$, $t_{0.025,15} = 2.13145$ y $\hat{\sigma} = 1.618$, obtenidos en los ejemplos anteriores, calcularemos el intervalo de confianza al 95% para dicho valor del diámetro de un árbol de 20 años:

$$IC_{95\%}(y_{20}) = \left(\hat{y}_{20} - t_{0.025,15}\hat{\sigma}\sqrt{1 + \frac{1}{17} + \frac{(20-\bar{x})^2}{17\,s_X^2}}, \hat{y}_{20} + t_{0.025,15}\hat{\sigma}\sqrt{1 + \frac{1}{17} + \frac{(20-\bar{x})^2}{17\,s_X^2}} \right) =$$

$$\left(16.60 - 2.13145 \cdot 1.618\sqrt{1 + \frac{1}{17} + \frac{(20-11.06)^2}{17 \cdot 13.82}}, 16.60 + 2.13145 \cdot 1.618\sqrt{1 + \frac{1}{17} + \frac{(20-11.06)^2}{17 \cdot 13.82}} \right) \Rightarrow$$

$$IC_{95\%}(y_{20}) = (12.52, 20.68)$$

12.4 Contrastes de hipótesis sobre la regresión

Una vez que calculamos una recta de regresión de una variable Y sobre otra X, es necesario contrastar si la relación encontrada es significativa o no desde el punto de vista estadístico.

12.4.1 Contraste de hipótesis para el parámetro β

(a) Contraste sobre la correlación lineal ($\beta = 0$):

Una forma de contrastar la relación lineal entre las variables X e Y dada por la recta de regresión $Y = \alpha + \beta X$ es comprobar si se puede aceptar que el valor de β sea distinto de 0, ya que si $\beta = 0$ tendremos que $Y = \alpha$ independientemente del valor de X, lo que significa que no hay correlación entre las variables.

Para ello se planteará un contraste de hipótesis de la siguiente forma:

$$\begin{cases} H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \end{cases}$$

Como sabemos que el estadístico $b = \dfrac{S_{XY}}{s_X^2}$ es un estimador de la pendiente de la recta y que además $b \sim N\left(\beta, \dfrac{\sigma}{s_X\sqrt{n}}\right)$, la variable $Z = \dfrac{b-\beta}{\sigma/(s_X\sqrt{n})}$ sigue una distribución $N(0,1)$. Sin embargo, como no conocemos la varianza σ^2 tendremos que estimarla utilizando $\mathcal{S}^2 = \dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}\delta_i^2$.

Por este motivo tendremos que utilizar en nuestro contraste de hipótesis el estadístico $T = \dfrac{b-\beta}{\mathcal{S}/(s_X\sqrt{n})}$, que, asumiendo que es cierta la hipótesis nula, $\beta = 0$, se escribe como:

$$T = \frac{b}{\mathcal{S}/(s_X\sqrt{n})}$$

y sigue una distribución t de Student con $n-2$ grados de libertad.

El valor muestral del estadístico T es $t_{muestral} = \dfrac{\hat{\beta}}{\hat{\sigma}/(s_X\sqrt{n})}$. Además podemos reescribir dicho estadístico como $t_{muestral} = \dfrac{\dfrac{s_{XY}}{s_X^2}}{\dfrac{\sqrt{\frac{n}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right)}}{s_X\sqrt{n}}} = \dfrac{s_{XY}\sqrt{n-2}}{\sqrt{s_X^2 s_Y^2 - s_{XY}^2}} = \dfrac{\frac{s_{XY}}{s_X s_Y}\sqrt{n-2}}{\sqrt{1-\frac{s_{XY}^2}{s_X^2 s_Y^2}}} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Por lo tanto, dado que la región de aceptación para el contraste a un nivel de significación α es $RA_\alpha = \left(-t_{\alpha/2,n-2}, t_{\alpha/2,n-2}\right)$, aceptaremos la hipótesis nula de $\beta = 0$ si $t_{muestral} = \dfrac{\hat{\beta}}{\hat{\sigma}/(s_X\sqrt{n})} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \in RA_\alpha$. Esto significaría que podemos escribir $Y = \alpha$ independientemente del valor de X, por lo que no hay relación entre las variables X e Y, lo cual equivale a que $\rho = 0$.

Así pues, el contraste planteado anteriormente sobre β es equivalente a:

$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

(b) Contraste sobre un valor de la pendiente ($\beta = \beta_0$):

En determinadas ocasiones interesa saber si la pendiente de la recta de regresión de Y sobre X toma un valor determinado β_0, por lo tanto, el contraste que tenemos que plantear es:

$$\begin{cases} H_0: & \beta = \beta_0 \\ H_1: & \beta \neq \beta_0 \end{cases}$$

Al igual que en el caso anterior, utilizaremos el estadístico $T = \dfrac{b-\beta}{\mathcal{S}/(s_X\sqrt{n})}$ que sigue una distribución t de Student con $n-2$ grados de libertad. Asumiendo que es cierta la hipótesis

nula, dicho estadístico será:

$$T = \frac{b - \beta_0}{\mathcal{S}/(s_X\sqrt{n})}$$

cuyo valor muestral es $t_{muestral} = \dfrac{\hat{\beta} - \beta_0}{\hat{\sigma}/(s_X\sqrt{n})}$. Asumiremos como cierta la hipótesis nula a un nivel de significación α si $t_{muestral} \in RA_\alpha = \left(-t_{\alpha/2,n-2}, t_{\alpha/2,n-2}\right)$.

Ejemplo 7:

Analizar si la correlación lineal de la recta de regresión de Y sobre X del ejemplo 1 es significativa.

En el ejercicio 1 calculamos la recta de regresión de Y sobre X, obteniendo $y = 7.52068 + 0.45398\,x$. Además, en el ejemplo 2 obtuvimos el coeficiente de correlación de las variables, $r = 0.743$.

El contraste que tenemos que plantear para estudiar si la correlación lineal es significativa es:

$$\begin{cases} H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \end{cases}$$

donde el estadístico que utilizaremos es $T = \dfrac{b}{\mathcal{S}/(s_X\sqrt{n})}$ que sigue una distribución t de Student con $n - 2 = 15$ grados de libertad.

El valor muestral de dicho estadístico es $t_{muestral} = \dfrac{\hat{\beta}}{\hat{\sigma}/(s_X\sqrt{n})} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.743\sqrt{15}}{\sqrt{1-0.743^2}} = 4.299$. Como $t_{0.025,15} = 2.13145$ la región de aceptación para $\alpha = 0.05$ será $RA_{0.05} = (-2.13145, 2.13145)$, por lo que $t_{muestral} \notin RA_{0.05}$, así pues rechazaremos la hipótesis nula y aceptaremos que las variables X e Y están correlacionadas.

12.4.2 Contraste de significación de la regresión lineal

En el apartado anterior realizamos un contraste sobre la pendiente para analizar la significación de la correlación entre dos variables. Sin embargo, hay otras formas de contrastar la significación de la regresión lineal, tal y como veremos a continuación. Analizar la regresión lineal es realizar el contraste:

$$\begin{cases} H_0: & \mu_Y = \mu_{Y|X=x_i} = \alpha + \beta x_i \ , \ \forall i = 1, \ldots, n \\ H_1: & \mu_Y \neq \mu_{Y|X=x_i} = \alpha + \beta x_i \ , \text{ para algún } i \end{cases}$$

Este contraste es equivalente al realizado anteriormente con hipótesis nula $H_0: \beta = 0$, ya que aceptar que $\beta = 0$ es aceptar que $\mu_Y = \mu_{Y|X=x_i} = \alpha + \beta x_i = \alpha$ para todos los valores de X.

El contraste propuesto, que supone contrastar la igualdad de medias de las variables Y_i, se puede plantear como un contraste sobre las varianzas, como los ANOVA realizados en el tema anterior. Por lo tanto, tendremos que contrastar la igualdad de la varianza de las medias, $\sigma^2_{MSG} = \sigma^2_{\bar{Y}}$, con la varianza de los errores, $\sigma^2_{MSE} = \sigma^2$, de modo que reescribiremos el contraste como:

$$\begin{cases} H_0: & \sigma^2_{\bar{Y}} = \sigma^2 \\ H_1: & \sigma^2_{\bar{Y}} > \sigma^2 \end{cases}$$

donde el estimador de $\sigma^2_{MSE} = \sigma^2$ será $\mathcal{S}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\delta_i^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}(y_i - (a+bx_i))^2 = \dfrac{SSE}{n-2}$.

Para poder obtener un estimador para la varianza de las medias, que es la que nos dará una medida de la regresión lineal, debemos recordar que la suma de los cuadrados totales, $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, se puede escribir como la suma de los cuadrados de los errores, $SSE = \sum_{i=1}^{n}(y_i - (a+bx_i))^2$, más la suma de los cuadrados de las medias, $SSG = \sum_{i=1}^{n}((a+bx_i) - \bar{y})^2$, por lo que $SST = SSE + SSG$. Como $\mathcal{S}^2_Y = \dfrac{SST}{n-1}$ es un estimador de la varianza de la variable Y, σ^2_Y, que será igual a la varianza de los residuos, σ^2, sólo en el caso de que la hipótesis nula sea cierta (y no haya correlación entre las variables), la variable $\dfrac{(n-1)\mathcal{S}^2_Y}{\sigma^2} = \dfrac{SST}{\sigma^2}$ sigue una distribución χ^2 con $n-1$ grados de libertad. De igual modo, $\mathcal{S}^2 = \dfrac{SSE}{n-2} = MSE$ es un estimador de la varianza de los errores, σ^2, por lo que la variable $\dfrac{(n-2)\mathcal{S}^2}{\sigma^2} = \dfrac{SSE}{\sigma^2}$ sigue una distribución χ^2 con $n-2$ grados de libertad. Así pues, como $SSG = SST - SSE$, la variable aleatoria $\dfrac{SSG}{\sigma^2}$ sigue una distribución χ^2 con $(n-1)-(n-2) = 1$ grados de libertad.

Con todo ello, tendremos que un estimador de la varianza de las medias será $MSG = \dfrac{SSG}{1}$, y si definimos la variable aleatoria $F = \dfrac{MSG/\sigma^2}{MSE/\sigma^2} = \dfrac{MSG}{MSE}$, ésta seguirá una distribución F de Fisher con 1 grado de libertad en el numerador y $n-2$ grados de libertad en el denominador. La región de aceptación para la hipótesis nula $H_0: \sigma^2_{MSG} = \sigma^2_{MSE}$ a un nivel de significación α es $RA_\alpha = [0, F_{\alpha;1,n-2})$.

Como el valor muestral de SST es ns^2_Y y el de SSE, tal y como vimos anteriormente, es $n\left(s^2_Y - \dfrac{s^2_{XY}}{s^2_X}\right)$ podemos obtener el valor muestral de SSG, que será $ns^2_Y - n\left(s^2_Y - \dfrac{s^2_{XY}}{s^2_X}\right) = n\dfrac{s^2_{XY}}{s^2_X}$.

Por lo tanto, la tabla del ANOVA para el contraste de significación de la regresión lineal es:

Variación	*Suma de cuadrados, SS*	*g.l.*	*Cuadrados medios, MS*	*Estadístico, F*
Debida a la regresión	$SSG = \sum_{i=1}^{n} \left(\hat{\alpha} + \hat{\beta} x_i - \bar{y}\right)^2 = n \frac{s_{XY}^2}{s_X^2}$	1	$MSG = \frac{SSG}{1}$	$F = \frac{MSG}{MSE}$
Debida a los errores	$SSE = \sum_{i=1}^{n} \left(y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right)\right)^2 = n\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right)$	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n\, s_Y^2$	$n-1$		

Ejemplo 8:

Aplicar un contraste de significación para la regresión lineal de la recta de regresión de Y sobre X del ejemplo 1.

En primer lugar calculamos la tabla del ANOVA sabiendo que $n = 17$, $s_{XY} = 6.274$, $s_X^2 = 13.82$ y $s_Y^2 = 5.157$, por lo que $SSG = 17 \dfrac{6.274^2}{13.82} = 48.42$, $SSE = 17\left(5.157 - \dfrac{6.274^2}{13.82}\right) = 39.24$ y $SST = 17 \cdot 5.157 = 87.66$.

Dicha tabla será:

Variación	*Suma de cuadrados, SS*	*g.l.*	*Cuadrados medios, MS*	*Estadístico, F*
Debida a la regresión	48.42	1	48.42	18.51
Debida a los errores	39.24	15	2.62	
Total	87.66	16		

Como $f_{muestral} = 18.51 > F_{0.05;1,15} = 4.543$ rechazaremos la hipótesis nula y aceptaremos que existe regresión lineal entre las variables X e Y.

12.4.3 Contraste de hipótesis para el parámetro α

En determinadas ocasiones necesitamos realizar un contraste sobre la ordenada en el origen de una recta de regresión. En ese caso, el tipo de contraste de hipótesis que tenemos que plantear será:

$$\begin{cases} H_0: & \alpha = \alpha_0 \\ H_1: & \alpha \neq \alpha_0 \end{cases}$$

Como sabemos que el estadístico $a = \bar{Y} - b\bar{x}$ es un estimador de la ordenada en el origen y que además $a \sim N\left(\alpha, \frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}\right)$, la variable $Z = \dfrac{a-\alpha}{\frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}}$ seguirá una distribución $N(0,1)$. Sin embargo, al igual que cuando realizabamos un contraste sobre la pendiente β, como no conocemos la varianza σ^2 tendremos que estimarla utilizando el estimador $\mathcal{S}^2 = \dfrac{1}{n-2}\sum_{i=1}^{n}\delta_i^2$. Por este motivo usaremos en nuestro contraste de hipótesis el estadístico $T = \dfrac{a-\alpha}{\frac{\mathcal{S}}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}}$, que, asumiendo que es cierta la hipótesis nula, $\alpha = \alpha_0$, se escribirá como:

$$T = \frac{a-\alpha_0}{\frac{\mathcal{S}}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}}$$

y seguirá una distribución t de Student con $n-2$ grados de libertad.

El valor muestral del estadístico T es $t_{muestral} = \dfrac{\hat{\alpha}-\alpha_0}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_X^2}}}$. Como la región de aceptación para el contraste a un nivel de significación α (no confundir el nivel de significación con la ordenada en el origen) es $RA_\alpha = \left(-t_{\alpha/2,n-2}, t_{\alpha/2,n-2}\right)$, aceptaremos la hipótesis nula de $\alpha = \alpha_0$ si $t_{muestral} \in RA_\alpha$.

Ejemplo 9:

Contrastar si para la recta de regresión de Y sobre X del ejemplo 1 es posible que $\alpha = 0$.

La recta de regresión que tenemos es $y = 7.52068 + 0.45398\,x$, por lo que $\hat{\alpha} = 7.52068$. Además de ejemplos anteriores sabemos que $n = 17$, $s_X^2 = 13.82$, $\bar{x} = 11.06$ y $\hat{\sigma}^2 = 2.617 \Rightarrow \hat{\sigma} = 1.618$.

El contraste de hipótesis que tenemos que plantear es:

$$\begin{cases} H_0: & \alpha = 0 \\ H_1: & \alpha \neq 0 \end{cases}$$

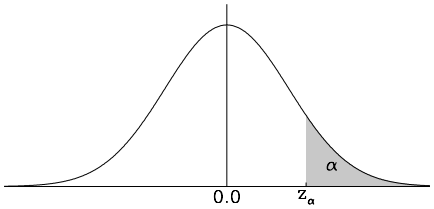
y el estadístico que utilizaremos $T = \dfrac{a}{\dfrac{\mathcal{S}}{\sqrt{n}}\sqrt{1+\dfrac{\bar{x}^2}{s_X^2}}}$, que sigue una distribución t de Student con $n-2=15$ grados de libertad, y cuyo valor muestral es $t_{muestral} = \dfrac{7.52068}{\dfrac{1.618}{17}\sqrt{1+\dfrac{11.06^2}{13.82}}} = 25.18$.

Dado que $t_{0.025,15} = 2.13145 < t_{muestral} = 25.18$ rechazaremos que la ordenada en el origen valga 0 para un nivel de significación de 0.05.

Apéndices

A Tabla de la distribución normal tipificada

Tabla de las áreas de las colas de la derecha de una distribución normal tipificada, $\alpha = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$, para valores z_α.

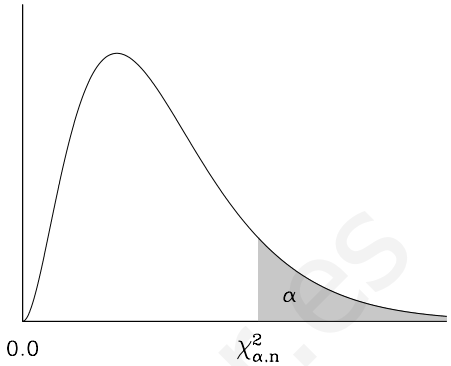


z_α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4820	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1921	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0447	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00889	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.001350	0.001306	0.001264	0.001223	0.001183	0.001144	0.001107	0.001070	0.001035	0.001001
3.1	0.000968	0.000936	0.000904	0.000874	0.000845	0.000816	0.000789	0.000762	0.000734	0.000711
3.2	0.000687	0.000664	0.000641	0.000619	0.000598	0.000577	0.000557	0.000538	0.000519	0.000501
3.3	0.000483	0.000466	0.000450	0.000434	0.000419	0.000404	0.000390	0.000376	0.000362	0.000350
3.4	0.000337	0.000325	0.000313	0.000302	0.000291	0.000280	0.000270	0.000260	0.000251	0.000242
3.5	0.000233	0.000224	0.000216	0.000208	0.000200	0.000193	0.000185	0.000178	0.000172	0.000165
3.6	0.000159	0.000153	0.000147	0.000142	0.000136	0.000131	0.000126	0.000121	0.000117	0.000112
3.7	0.000108	0.000104	0.000100	0.000096	0.000092	0.000088	0.000085	0.000082	0.000078	0.000075
3.8	0.000072	0.000070	0.000068	0.000064	0.000062	0.000059	0.000057	0.000054	0.000052	0.000050
3.9	0.000048	0.000046	0.000044	0.000042	0.000041	0.000039	0.000037	0.000036	0.000034	0.000033
4.0	0.0000317	0.0000304	0.0000291	0.0000279	0.0000267	0.0000256	0.0000245	0.0000235	0.0000225	0.0000216
4.1	0.0000207	0.0000198	0.0000190	0.0000181	0.0000174	0.0000166	0.0000159	0.0000152	0.0000146	0.0000140
4.2	0.0000134	0.0000128	0.0000122	0.0000117	0.0000112	0.0000107	0.0000102	0.0000098	0.0000094	0.0000089
4.3	0.0000085	0.0000082	0.0000078	0.0000075	0.0000071	0.0000068	0.0000065	0.0000062	0.0000059	0.0000057
4.4	0.0000054	0.0000052	0.0000049	0.0000047	0.0000045	0.0000043	0.0000041	0.0000039	0.0000037	0.0000036

B Tabla de la distribución χ^2 de Pearson

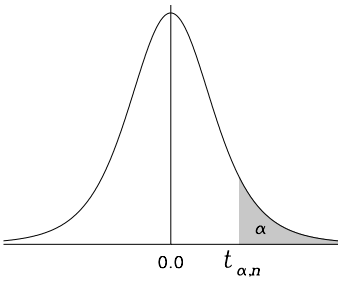
Valor de la abcisa $\chi^2_{\alpha,n}$ que encierra a la derecha un área α bajo la función de densidad de probabilidad de una χ^2 de n grados de libertad, dada por

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



	α											
n	0.999	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005	0.001
1	0.0000016	0.000039	0.000157	0.000982	0.003932	0.015791	2.70549	3.84134	5.02368	6.63454	7.87894	10.8266
2	0.002001	0.010025	0.020101	0.050636	0.102587	0.210721	4.60517	5.99146	7.37776	9.21034	10.5966	13.8155
3	0.024297	0.071721	0.114831	0.215794	0.351844	0.584370	6.25134	7.81466	9.34833	11.3448	12.8381	16.2662
4	0.090804	0.206989	0.297109	0.484419	0.710723	1.06362	7.77944	9.48773	11.1433	13.2767	14.8603	18.4668
5	0.210213	0.411742	0.554298	0.831212	1.14548	1.61031	9.23636	11.0705	12.8325	15.0863	16.7496	20.5150
6	0.391069	0.675729	0.872093	1.23735	1.63539	2.20413	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	0.598494	0.989256	1.23904	1.68987	2.16735	2.83311	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	0.857120	1.34442	1.64651	2.17974	2.73264	3.48954	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	1.15195	1.73493	2.08790	2.70039	3.32511	4.16816	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772
10	1.47879	2.15588	2.55823	3.24698	3.94031	4.86519	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883
11	1.83385	2.60322	3.05348	3.81575	4.57481	5.57778	17.2750	19.6751	21.9201	24.7250	26.7568	31.2641
12	2.21431	3.07386	3.57059	4.40380	5.22604	6.30380	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095
13	2.61722	3.56503	4.10692	5.00875	5.89186	7.04150	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282
14	3.04082	4.07472	4.66045	5.62874	6.57064	7.78954	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233
15	3.48268	4.60092	5.22935	6.26214	7.26094	8.54676	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973
16	3.94182	5.14226	5.81224	6.90768	7.96165	9.31224	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524
17	4.41609	5.69722	6.40776	7.56419	8.67176	10.0852	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902
18	4.90505	6.26485	7.01493	8.23076	9.39046	10.8649	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124
19	5.40682	6.84397	7.63273	8.90652	10.1170	11.6509	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202
20	5.92124	7.43389	8.26042	9.59078	10.8508	12.4426	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147
21	6.44668	8.03365	8.89720	10.2829	11.5913	13.2396	29.6151	32.6706	35.4789	38.9322	41.4011	46.7970
22	6.98315	8.64275	9.54251	10.9823	12.3380	14.0415	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679
23	7.52924	9.26042	10.1957	11.6886	13.0905	14.8480	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282
24	8.08504	9.88626	10.8564	12.4012	13.8484	15.6587	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786
25	8.64934	10.5197	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197
26	9.22225	11.1603	12.1982	13.8439	15.3792	17.2919	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520
27	9.80278	11.8076	12.8785	14.5734	16.1514	18.1139	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760
28	10.3910	12.4613	13.5647	15.3079	16.9279	18.9392	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923
29	10.9861	13.1212	14.2565	16.0471	17.7084	19.7677	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012
30	11.5880	13.7867	14.9535	16.7908	18.4927	20.5992	40.2560	43.7730	46.9792	50.8922	53.6720	59.7031

C Tabla de la distribución t de Student



Valor de la abcisa $t_{\alpha,n}$ que encierra a la derecha un área α bajo la función de densidad de probabilidad de una t de Student de n grados de libertad, dada por

$$f_n(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

	α									
n	0.400	0.300	0.200	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
1	0.324920	0.726543	1.37638	3.07768	6.31375	12.7062	31.8205	63.6567	318.309	636.619
2	0.288675	0.617213	1.06066	1.88562	2.91999	4.30265	6.96456	9.92484	22.3271	31.5991
3	0.276672	0.584392	0.978476	1.63775	2.35338	3.18245	4.54070	5.84091	10.2145	12.9240
4	0.270724	0.568651	0.940967	1.53321	2.13185	2.77645	3.74695	4.60410	7.17318	8.61030
5	0.267183	0.559432	0.919546	1.47589	2.01505	2.57059	3.36494	4.03216	5.89343	6.86883
6	0.264836	0.553383	0.905706	1.43976	1.94319	2.44692	3.14268	3.70744	5.20763	5.95882
7	0.263169	0.549112	0.896032	1.41493	1.89458	2.36463	2.99796	3.49949	4.78531	5.40790
8	0.261923	0.545936	0.888892	1.39682	1.85955	2.30601	2.89647	3.35540	4.50081	5.04132
9	0.260957	0.543482	0.883406	1.38303	1.83312	2.26216	2.82144	3.24984	4.29682	4.78093
10	0.260187	0.541530	0.879060	1.37219	1.81247	2.22814	2.76378	3.16928	4.14371	4.58691
11	0.259558	0.539940	0.875532	1.36343	1.79589	2.20099	2.71809	3.10581	4.02471	4.43699
12	0.259034	0.538620	0.872612	1.35622	1.78229	2.17882	2.68100	3.05455	3.92964	4.31780
13	0.258593	0.537506	0.870154	1.35017	1.77094	2.16037	2.65031	3.01228	3.85199	4.22084
14	0.258214	0.536554	0.868057	1.34503	1.76131	2.14479	2.62450	2.97685	3.78740	4.14046
15	0.257887	0.535731	0.866247	1.34061	1.75305	2.13145	2.60249	2.94672	3.73284	4.07278
16	0.257601	0.535012	0.864669	1.33676	1.74589	2.11991	2.58349	2.92079	3.68616	4.01501
17	0.257349	0.534379	0.863281	1.33338	1.73961	2.10982	2.56694	2.89824	3.64578	3.96514
18	0.257125	0.533818	0.862051	1.33039	1.73407	2.10093	2.55239	2.87845	3.61049	3.92166
19	0.256925	0.533316	0.860953	1.32773	1.72914	2.09303	2.53949	2.86094	3.57941	3.88342
20	0.256744	0.532865	0.859967	1.32534	1.72472	2.08597	2.52798	2.84535	4.55182	3.84953
21	0.256582	0.532457	0.859076	1.32319	1.72075	2.07962	2.51765	2.83137	3.52716	3.81929
22	0.256434	0.532087	0.858268	1.32124	1.71715	2.07388	2.50833	2.81876	3.50500	3.79214
23	0.256299	0.531749	0.857532	1.31946	1.71388	2.06866	2.49987	2.80734	3.48497	3.76764
24	0.256175	0.531440	0.856858	1.31784	1.71089	2.06390	2.49216	2.79695	3.46678	3.74541
25	0.256061	0.531156	0.856238	1.31635	1.70814	2.05954	2.48511	2.78744	3.45020	3.72515
26	0.255956	0.530894	0.855668	1.31497	1.70562	2.05553	2.47863	2.77872	3.43500	3.70662
27	0.255859	0.530651	0.855140	1.31371	1.70329	2.05183	2.47266	2.77069	3.42104	3.68960
28	0.255769	0.530426	0.854650	1.31253	1.70113	2.04841	2.46715	2.76327	3.40816	3.67391
29	0.255685	0.530216	0.854194	1.31144	1.69913	2.04523	2.46203	2.75639	3.39625	3.65941
30	0.255607	0.530021	0.85377	1.31042	1.69726	2.04228	2.45727	2.75000	3.38519	3.64597
40	0.255040	0.528608	0.850702	1.30308	1.68385	2.02108	2.42326	2.70446	3.30688	3.55097
50	0.254701	0.527762	0.848872	1.29872	1.67591	2.00856	2.40328	2.67780	3.26142	3.49602
60	0.254475	0.527200	0.847655	1.29582	1.67065	2.00030	2.39012	2.66029	3.23172	3.46021
70	0.254314	0.526799	0.846789	1.29377	1.66692	1.99444	2.38081	2.64791	3.21080	3.43502
80	0.254193	0.526498	0.846140	1.29223	1.66413	1.99007	2.37387	2.63870	3.19526	3.41634
90	0.254099	0.526265	0.845636	1.29103	1.66196	1.98668	2.36850	2.63157	3.18328	3.40194
100	0.254024	0.526078	0.845233	1.29008	1.66024	1.98398	2.36422	2.62590	3.17375	3.39050
200	0.253686	0.525239	0.843424	1.28580	1.65251	1.97190	2.34514	2.60064	3.13149	3.33984
300	0.253574	0.524960	0.842823	1.28438	1.64995	1.96791	2.33885	2.59232	3.11763	3.32326
400	0.253517	0.524821	0.842523	1.28367	1.64868	1.96592	2.33571	2.58818	3.11074	3.31502
500	0.253484	0.524737	0.842343	1.28325	1.64791	1.96472	2.33383	2.58570	3.10662	3.31010
∞	0.253349	0.524402	0.841623	1.28155	1.64486	1.95997	2.32635	2.57583	3.09024	3.29053

D Tabla de la distribución F de Fisher

Valor de la abcisa $F_{\alpha;n_1,n_2}$ que encierra a la derecha un área α bajo la función de densidad de probabilidad de una F de Fisher de n_1 grados de libertad en el numerador y n_2 grados de libertad en el denominador, dada por

$$f_{n_1,n_2}(x) = \begin{cases} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} & \text{, si } x>0 \\ 0 & \text{, si } x\leq 0 \end{cases}$$

Para valores de α cercanos a 1 se puede calcular teniendo en cuenta que $F_{\alpha;n_1,n_2} = \dfrac{1}{F_{1-\alpha;n_2,n_1}}$

$\alpha = 0.10$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	70	100	∞
1	39.863	49.500	53.592	55.832	57.239	58.203	58.904	59.438	59.856	60.194	60.704	61.219	61.739	62.053	62.264	62.528	62.687	62.869	63.006	63.327
2	8.5263	9.0000	9.1618	9.2434	9.2926	9.3255	9.3491	9.3668	9.3805	9.3916	9.4081	9.4247	9.4413	9.4513	9.4579	9.4662	9.4712	9.4769	9.4812	9.4912
3	5.5384	5.4623	5.3908	5.3426	5.3092	5.2847	5.2662	5.2516	5.2400	5.2304	5.2156	5.2003	5.1844	5.1747	5.1681	5.1597	5.1546	5.1487	5.1442	5.1337
4	4.5448	4.3246	4.1908	4.1073	4.0506	4.0098	3.9789	3.9549	3.9356	3.9199	3.8955	3.8703	3.8443	3.8283	3.8174	3.8036	3.7952	3.7855	3.7782	3.7607
5	4.0604	3.7798	3.6195	3.5202	3.4530	3.4045	3.3679	3.3393	3.3163	3.2974	3.2683	3.2380	3.2067	3.1873	3.1741	3.1573	3.1471	3.1353	3.1263	3.1050
6	3.7760	3.4633	3.2888	3.1808	3.1075	3.0546	3.0145	2.9830	2.9577	2.9369	2.9047	2.8712	2.8363	2.8147	2.8000	2.7812	2.7697	2.7564	2.7463	2.7222
7	3.5895	3.2575	3.0741	2.9605	2.8833	2.8274	2.7849	2.7516	2.7247	2.7025	2.6681	2.6322	2.5947	2.5714	2.5555	2.5351	2.5226	2.5082	2.4971	2.4708
8	3.4579	3.1131	2.9238	2.8064	2.7265	2.6683	2.6241	2.5894	2.5612	2.5380	2.5020	2.4642	2.4246	2.3999	2.3830	2.3614	2.3481	2.3326	2.3208	2.2926
9	3.3603	3.0065	2.8129	2.6927	2.6106	2.5509	2.5053	2.4694	2.4403	2.4163	2.3789	2.3396	2.2983	2.2725	2.2547	2.2320	2.2180	2.2017	2.1892	2.1592
10	3.2850	2.9245	2.7278	2.6053	2.5216	2.4606	2.4140	2.3772	2.3473	2.3226	2.2841	2.2435	2.2007	2.1739	2.1554	2.1317	2.1171	2.1000	2.0869	2.0554
12	3.1766	2.8068	2.6057	2.4801	2.3939	2.3310	2.2828	2.2446	2.2135	2.1878	2.1474	2.1049	2.0597	2.0312	2.0115	1.9861	1.9704	1.9520	1.9379	1.9036
15	3.0732	2.6952	2.4898	2.3615	2.2729	2.2081	2.1582	2.1185	2.0862	2.0593	2.0171	1.9722	1.9243	1.8939	1.8728	1.8454	1.8284	1.8083	1.7929	1.7551
20	2.9747	2.5893	2.3802	2.2489	2.1581	2.0913	2.0397	1.9985	1.9649	1.9367	1.8924	1.8449	1.7938	1.7611	1.7382	1.7083	1.6896	1.6674	1.6501	1.6074
25	2.9178	2.5283	2.3170	2.1843	2.0921	2.0241	1.9714	1.9293	1.8947	1.8658	1.8200	1.7708	1.7175	1.6831	1.6590	1.6272	1.6072	1.5833	1.5645	1.5176
30	2.8807	2.4887	2.2762	2.1422	2.0492	1.9803	1.9269	1.8841	1.8490	1.8195	1.7727	1.7223	1.6673	1.6316	1.6065	1.5732	1.5522	1.5269	1.5069	1.4564
40	2.8354	2.4404	2.2262	2.0910	1.9968	1.9269	1.8725	1.8289	1.7929	1.7627	1.7146	1.6624	1.6052	1.5677	1.5411	1.5056	1.4830	1.4555	1.4336	1.3769
50	2.8087	2.4120	2.1968	2.0608	1.9659	1.8954	1.8405	1.7963	1.7598	1.7292	1.6802	1.6269	1.5681	1.5294	1.5018	1.4648	1.4409	1.4119	1.3885	1.3267
70	2.7786	2.3800	2.1638	2.0270	1.9313	1.8600	1.8044	1.7596	1.7226	1.6913	1.6413	1.5866	1.5259	1.4857	1.4567	1.4176	1.3922	1.3608	1.3352	1.2652
100	2.7564	2.3564	2.1395	2.0020	1.9057	1.8339	1.7778	1.7324	1.6949	1.6632	1.6124	1.5566	1.4944	1.4528	1.4227	1.3817	1.3548	1.3212	1.2934	1.2142
∞	2.7056	2.3026	2.0838	1.9449	1.8473	1.7741	1.7167	1.6702	1.6315	1.5987	1.5458	1.4871	1.4206	1.3753	1.3419	1.2951	1.2633	1.2218	1.1849	1.0003

$\alpha = 0.05$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	70	100	∞
1	161.45	199.50	215.70	224.57	230.15	233.98	236.76	238.87	240.53	241.87	243.90	245.94	248.00	249.25	250.08	251.13	251.76	252.49	253.03	254.30
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.429	19.446	19.456	19.462	19.471	19.476	19.481	19.486	19.496
3	10.128	9.5520	9.2766	9.1171	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6601	8.6341	8.6165	8.5943	8.5809	8.5655	8.5538	8.5265
4	7.7087	6.9443	6.5913	6.3882	6.2560	6.1631	6.0942	6.0410	5.9987	5.9644	5.9117	5.8578	5.8025	5.7687	5.7459	5.7170	5.6995	5.6793	5.6641	5.6281
5	6.6079	5.7862	5.4095	5.1922	5.0504	4.9503	4.8759	4.8184	4.7725	4.7351	4.6777	4.6188	4.5582	4.5209	4.4957	4.4638	4.4444	4.4221	4.4051	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8349	3.8082	3.7743	3.7537	3.7298	3.7118	3.6689
7	5.5915	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4036	3.3758	3.3404	3.3189	3.2939	3.2749	3.2298
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1081	3.0794	3.0428	3.0204	2.9944	2.9747	2.9276
9	5.1174	4.2565	3.8626	3.6331	3.4817	3.3738	3.2928	3.2296	3.1789	3.1373	3.0730	3.0061	2.9365	2.8932	2.8637	2.8259	2.8028	2.7760	2.7556	2.7067
10	4.9646	4.1028	3.7083	3.4781	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7298	2.6996	2.6609	2.6371	2.6095	2.5884	2.5379
12	4.7473	3.8853	3.4906	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.4977	2.4663	2.4259	2.4010	2.3720	2.3498	2.2962
15	4.5431	3.6823	3.2874	3.0558	2.9011	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4035	2.3275	2.2797	2.2468	2.2043	2.1780	2.1472	2.1234	2.0659
20	4.3513	3.4928	3.0986	2.8661	2.7107	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0739	2.0391	1.9938	1.9656	1.9323	1.9066	1.8432
25	4.2417	3.3852	2.9912	2.7589	2.6028	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0900	2.0075	1.9555	1.9192	1.8718	1.8421	1.8069	1,7794	1.7110
30	4.1709	3.3158	2.9225	2.6896	2.5334	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8783	1.8409	1.7918	1.7609	1,7240	1.6950	1.6223
40	4.0848	3.2317	2.8390	2.6060	2.4493	2.3358	2.2490	2.1802	2.1240	2.0773	2.0035	1.9245	1.8389	1.7835	1.7444	1.6928	1.6600	1.6205	1.5892	1.5089
50	4.0343	3.1826	2.7902	2.5572	2.4003	2.2864	2.1992	2.1299	2.0734	2.0261	1.9515	1.8714	1.7841	1.7273	1.6872	1.6337	1.5995	1.5580	1.5249	1.4383
70	3.9778	3.1277	2.7357	2.5027	2.3455	2.2312	2.1435	2.0737	2.0166	1.9689	1.8933	1.8117	1.7223	1.6638	1.6220	1.5661	1.5300	1.4857	1.4498	1.3529
100	3.9362	3.0873	2.6957	2.4626	2.3052	2.1906	2.1025	2.0323	1.9748	1.9267	1.8503	1.7675	1.6764	1.6164	1.5733	1.5151	1.4772	1.4303	1.3917	1.2832
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.5061	1.4591	1.3940	1.3501	1.2933	1.2434	1.0003

Tabla de la distribución F de Fisher (continuación)

$\alpha = 0.025$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	70	100	∞
1	647.79	799.50	864.08	899.50	921.76	937.02	948.13	956.57	963.20	968.54	976.62	984.78	993.01	997.99	1001.3	1005.5	1008.0	1010.9	1013.1	1018.2
2	38.506	39.000	39.166	39.248	39.298	39.332	39.355	39.373	39.387	39.398	39.415	39.431	39.448	39.458	39.465	39.473	39.478	39.484	39.488	39.498
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473	14.419	14.336	14.253	14.167	14.116	14.080	14.036	14.010	13.979	13.956	13.902
4	12.218	10.649	9.9791	9.6045	9.3644	9.1973	9.0741	8.9796	8.9046	8.8439	8.7512	8.6565	8.5599	8.5009	8.4613	8.4111	8.3808	8.3458	8.3195	8.2573
5	10.007	8.4337	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192	6.5246	6.4278	6.3286	6.2679	6.2269	6.1751	6.1437	6.1074	6.0800	6.0153
6	8.8132	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613	5.3663	5.2687	5.1684	5.1069	5.0652	5.0125	4.9804	4.9434	4.9154	4.8491
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.6658	4.5678	4.4667	4.4046	4.3624	4.3089	4.2763	4.2387	4.2101	4.1423
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951	4.1997	4.1012	3.9995	3.9367	3.8940	3.8398	3.8067	3.7684	3.7393	3.6702
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1971	4.1020	4.0260	3.9639	3.8682	3.7694	3.6669	3.6035	3.5604	3.5055	3.4719	3.4330	3.4034	3.3329
10	6.9368	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.6210	3.5217	3.4185	3.3546	3.3110	3.2554	3.2214	3.1819	3.1517	3.0798
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4359	3.3736	3.2773	3.1772	3.0728	3.0077	2.9633	2.9064	2.8714	2.8307	2.7996	2.7249
15	6.1995	4.7651	4.1528	3.8043	3.5764	3.4147	3.2934	3.1987	3.1227	3.0602	2.9633	2.8621	2.7559	2.6894	2.6437	2.5850	2.5488	2.5065	2.4739	2.3954
20	5.8715	4.4613	3.8592	3.5147	3.2887	3.1283	3.0074	2.9128	2.8366	2.7737	2.6758	2.5731	2.4645	2.3959	2.3486	2.2873	2.2493	2.2045	2.1699	2.0853
25	5.6864	4.2909	3.6943	3.3534	3.1284	2.9685	2.8478	2.7531	2.6766	2.6135	2.5149	2.4110	2.3005	2.2303	2.1816	2.1183	2.0787	2.0319	1.9955	1.9055
30	5.5676	4.1821	3.5898	3.2500	3.0261	2.8667	2.7460	2.6513	2.5746	2.5112	2.4120	2.3072	2.1952	2.1237	2.0739	2.0089	1.9681	1.9195	1.8816	1.7867
40	5.4240	4.0510	3.4637	3.1262	2.9034	2.7444	2.6238	2.5289	2.4519	2.3882	2.2882	2.1819	2.0677	1.9943	1.9429	1.8752	1.8324	1.7810	1.7405	1.6371
50	5.3403	3.9749	3.3906	3.0545	2.8324	2.6735	2.5530	2.4579	2.3808	2.3168	2.2162	2.1090	1.9933	1.9186	1.8659	1.7963	1.7520	1.6984	1.6559	1.5452
70	5.2471	3.8903	3.3094	2.9748	2.7534	2.5949	2.4743	2.3791	2.3017	2.2374	2.1361	2.0277	1.9100	1.8334	1.7792	1.7069	1.6604	1.6038	1.5581	1.4357
100	5.1786	3.8284	3.2500	2.9166	2.6958	2.5374	2.4168	2.3215	2.2439	2.1793	2.0773	1.9679	1.8486	1.7705	1.7149	1.6401	1.5917	1.5320	1.4833	1.3474
∞	5.0239	3.6889	3.1161	2.7858	2.5665	2.4082	2.2875	2.1918	2.1136	2.0483	1.9447	1.8326	1.7085	1.6259	1.5660	1.4835	1.4284	1.3575	1.2956	1.0004

$\alpha = 0.01$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	70	100	∞
1	4052.2	4999.5	5402.1	5623.3	5762.3	5857.6	5927.0	5979.7	6021.1	6054.4	6104.9	6155.9	6207.3	6238.4	6259.2	6285.3	6301.1	6319.1	6332.6	6364.4
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.433	99.449	99.459	99.466	99.474	99.479	99.485	99.489	99.499
3	34.116	30.816	29.457	28.710	28.237	27.910	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.579	26.504	26.411	26.354	26.289	26.240	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.911	13.838	13.745	13.690	13.625	13.577	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.8883	9.7223	9.5527	9.4492	9.3794	9.2912	9.2379	9.1764	9.1300	9.0204
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761	7.8741	7.7183	7.5590	7.3958	7.2960	7.2285	7.1432	7.0915	7.0318	6.9867	6.8800
7	12.247	9.5466	8.4513	7.8467	7.4604	7.1914	6.9928	6.8401	6.7188	6.6201	6.4691	6.3143	6.1554	6.0580	5.9920	5.9085	5.8577	5.7991	5.7547	5.6495
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106	5.8143	5.6667	5.5151	5.3591	5.2631	5.1981	5.1156	5.0654	5.0073	4.9633	4.8588
9	10.562	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	5.2565	5.1114	4.9621	4.8080	4.7130	4.6486	4.5667	4.5167	4.4589	4.4150	4.3106
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	4.8492	4.7059	4.5581	4.4054	4.3111	4.2469	4.1653	4.1155	4.0577	4.0137	3.9090
12	9.3303	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875	4.2961	4.1553	4.0096	3.8584	3.7647	3.7008	3.6192	3.5692	3.5111	3.4668	3.3608
15	8.6832	6.3589	5.4170	4.8932	4.5556	4.3183	4.1416	4.0045	3.8948	3.8049	3.6662	3.5222	3.3719	3.2782	3.2141	3.1319	3.0814	3.0224	2.9772	2.8684
20	8.0960	5.8489	4.9397	4.4308	4.1027	3.8714	3.6987	3.5644	3.4567	3.3682	3.2311	3.0880	2.9377	2.8434	2.7785	2.6948	2.6430	2.5822	2.5353	2.4212
25	7.7698	5.5680	4.6755	4.1785	3.8541	3.6271	3.4568	3.3239	3.2172	3.1294	2.9931	2.8502	2.6993	2.6041	2.5383	2.4530	2.3999	2.3373	2.2888	2.1694
30	7.5625	5.3904	4.5110	4.0180	3.6982	3.4734	3.3045	3.1726	3.0665	2.9791	2.8431	2.7002	2.5487	2.4526	2.3860	2.2992	2.2450	2.1808	2.1307	2.0062
40	7.3141	5.1785	4.3137	3.8284	3.5131	3.2910	3.1238	2.9930	2.8876	2.8005	2.6648	2.5216	2.3689	2.2714	2.2034	2.1142	2.0581	1.9911	1.9383	1.8047
50	7.1706	5.0566	4.2004	3.7196	3.4069	3.1864	3.0202	2.8900	2.7850	2.6981	2.5625	2.4190	2.2652	2.1667	2.0976	2.0066	1.9490	1.8797	1.8248	1.6831
70	7.0114	4.9219	4.0754	3.5997	3.2900	3.0712	2.9060	2.7765	2.6719	2.5852	2.4496	2.3055	2.1504	2.0503	1.9798	1.8861	1.8263	1.7537	1.6954	1.5404
100	6.8953	4.8239	3.9847	3.5128	3.2052	2.9876	2.8233	2.6943	2.5898	2.5033	2.3676	2.2230	2.0667	1.9652	1.8933	1.7972	1.7353	1.6594	1.5977	1.4276
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073	2.3209	2.1848	2.0385	1.8783	1.7726	1.6964	1.5923	1.5231	1.4347	1.3582	1.0005

$\alpha = 0.005$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	50	70	100	∞
1	16211	20000	21605	22489	23045	23426	23704	23914	24080	24213	24415	24619	24825	24949	25032	25137	25199	25272	25326	25453
2	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.38	199.39	199.40	199.42	199.43	199.45	199.46	199.47	199.48	199.48	199.49	199.49	199.50
3	55.552	49.799	47.467	46.194	45.392	44.838	44.434	44.125	43.882	43.685	43.387	43.085	42.777	42.591	42.466	42.308	42.213	42.104	42.021	41.829
4	31.333	26.284	24.259	23.155	22.456	21.975	21.622	21.352	21.139	20.967	20.705	20.438	20.167	20.002	19.892	19.752	19.667	19.570	19.497	19.325
5	22.785	18.314	16.530	15.556	14.940	14.513	14.201	13.961	13.772	13.618	13.385	13.146	12.904	12.756	12.656	12.530	12.454	12.366	12.300	12.144
6	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.392	10.250	10.034	9.8140	9.5888	9.4511	9.3583	9.2409	9.1697	9.0877	9.0257	8.8793
7	16.236	12.404	10.882	10.051	9.5221	9.1553	8.8854	8.6781	8.5138	8.3803	8.1764	7.9678	7.7540	7.6230	7.5345	7.4225	7.3544	7.2760	7.2166	7.0760
8	14.688	11.042	9.5965	8.8051	8.3018	7.9520	7.6941	7.4959	7.3386	7.2106	7.0149	6.8143	6.6082	6.4817	6.3961	6.2875	6.2216	6.1453	6.0875	5.9505
9	13.614	10.107	8.7171	7.9559	7.4712	7.1339	6.8849	6.6933	6.5411	6.4172	6.2274	6.0325	5.8318	5.7084	5.6248	5.5186	5.4539	5.3791	5.3224	5.1875
10	12.827	9.4270	8.0808	7.3428	6.8724	6.5446	6.3025	6.1159	5.9676	5.8467	5.6613	5.4707	5.2740	5.1528	5.0706	4.9659	4.9022	4.8283	4.7721	4.6385
12	11.754	8.5096	7.2258	6.5211	6.0711	5.7570	5.5245	5.3451	5.2021	5.0855	4.9063	4.7213	4.5299	4.4115	4.3309	4.2282	4.1653	4.0924	4.0368	3.9039
15	10.798	7.7008	6.4760	5.8029	5.3721	5.0708	4.8473	4.6744	4.5364	4.4235	4.2498	4.0698	3.8826	3.7662	3.6868	3.5850	3.5225	3.4498	3.3941	3.2602
20	9.9440	6.9865	5.8208	5.1743	4.7616	4.4722	4.2569	4.0900	3.9564	3.8470	3.6779	3.5020	3.3178	3.2025	3.1234	3.0215	2.9586	2.8849	2.8282	2.6904
25	9.4754	6.5982	5.4616	4.8374	4.4309	4.1499	3.9394	3.7758	3.6447	3.5371	3.3704	3.1963	3.0133	2.8981	2.8187	2.7160	2.6522	2.5772	2.5191	2.3765
30	9.1797	6.3547	5.2414	4.6236	4.2258	3.9491	3.7415	3.5801	3.4505	3.3440	3.1787	3.0057	2.8230	2.7076	2.6278	2.5241	2.4594	2.3829	2.3235	2.1760
40	8.8279	6.0664	4.9782	4.3740	3.9844	3.7128	3.5088	3.3498	3.2220	3.1168	2.9531	2.7811	2.5984	2.4823	2.4015	2.2958	2.2295	2.1504	2.0884	1.9318
50	8.6258	5.9016	4.8281	4.2318	3.8471	3.5784	3.3764	3.2189	3.0921	2.9875	2.8247	2.6531	2.4702	2.3533	2.2717	2.1644	2.0967	2.0155	1.9512	1.7863
70	8.4027	5.7204	4.6633	4.0760	3.6967	3.4312	3.2315	3.0755	2.9498	2.8460	2.6840	2.5127	2.3291	2.2112	2.1283	2.0187	1.9488	1.8642	1.7966	1.6176
100	8.2407	5.5892	4.5443	3.9635	3.5882	3.3251	3.1271	2.9722	2.8472	2.7440	2.5825	2.4113	2.2270	2.1080	2.0239	1.9119	1.8400	1.7521	1.6809	1.4858
∞	7.8795	5.2983	4.2794	3.7151	3.3499	3.0913	2.8968	2.7444	2.6210	2.5188	2.3583	2.1868	1.9998	1.8771	1.7891	1.6692	1.5898	1.4888	1.4019	1.0005

E Aproximaciones más comunes entre funciones de probabilidad

Se presenta a continuación un resumen de las aproximaciones más comunes entre funciones de distribución de probabilidad, indicando también el régimen de validez de dicha aproximación.

Distribución	*Aproximación*	*Regla de validez*
Binomial $Bin(n,p)$	Poisson $Poi(\lambda = np)$	n grande y p pequeño $n > 50,\ p < 0.1$
Binomial $Bin(n,p)$	Normal $N(np, \sqrt{np(1-p)})$	n grande y p próximo a 0.5 $np \geq 5,\ n(1-p) \geq 5$
Poisson $Poi(\lambda)$	Normal $N(\lambda, \sqrt{\lambda})$	λ grande $\lambda \geq 5$

F Resumen de las distribuciones discretas más comunes

En la siguiente tabla se presentan las funciones de probabilidad discretas más comunes, junto con los parámetros poblacionales que las definen así como los valores de la media μ y la varianza σ^2 de cada una de ellas.

Distribución	*Parámetros*	*Función de densidad de probabilidad, f*	*Media, μ*	*Varianza, σ^2*
Uniforme	$n > 0$	$f(x_i) = \frac{1}{n}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2$
Bernoulli	$0 \leq p \leq 1$	$f(x_i) = \begin{cases} p^{x_i}(1-p)^{1-x_i} & \text{, si } x_i = 1, 0 \\ 0 & \text{, resto} \end{cases}$	p	$p(1-p)$
Binomial $Bin(n,p)$	$n > 0$ $0 \leq p \leq 1$	$f(x_i) = \begin{cases} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} & \text{, si } x_i = 0, 1, \ldots, n \\ 0 & \text{, resto} \end{cases}$	np	$np(1-p)$
Poisson $Poi(\lambda)$	$\lambda > 0$	$f(x_i) = \begin{cases} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} & \text{, si } x_i = 0, 1, 2, ... \\ 0 & \text{, resto} \end{cases}$	λ	λ

G Resumen de las distribuciones continuas más comunes

En la siguiente tabla se presentan las funciones de densidad de probabilidad continuas más comunes, junto con los parámetros poblacionales que las definen así como los valores de la media μ y la varianza σ^2 de cada una de ellas.

Distribución	*Parámetros*	*Función de densidad de probabilidad, f*	*Media, μ*	*Varianza, σ^2*
Normal $N(\mu,\sigma)$	μ $\sigma>0$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniforme	$a<b$	$f(x)=\begin{cases}\frac{1}{b-a} & \text{, si } a<x<b\\ 0 & \text{, si } x<a \text{ ó } x>b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gamma $\Gamma(\alpha,\beta)$	$\alpha>0$ $\beta>0$	$f(x)=\begin{cases}\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/2} & \text{, si } x>0\\ 0 & \text{, si } x\leq 0\end{cases}$	$\alpha\beta$	$\alpha\beta^2$
Exponencial $E(\lambda)$	$\lambda>0$	$f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{, si } x\geq 0\\ 0 & \text{, si } x<0\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Chi-cuadrado χ^2_n	$n>0$	$f_n(x)=\begin{cases}\frac{1}{2^{n/2}\Gamma\left(n/2\right)}x^{\frac{n}{2}-1}e^{-x/2} & \text{, si } x>0\\ 0 & \text{, si } x\leq 0\end{cases}$	n	$2n$
t de Student t_n	$n>0$	$f_n(t)=\frac{1}{\sqrt{n\pi}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$	0	$\frac{n}{n-2}$ si $n>2$
F de Fisher $F_{(n_1,n_2)}$	$n_1>0$ $n_2>0$	$f_{n_1,n_2}(x)=\begin{cases}\left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} & \text{, si } x>0\\ 0 & \text{, si } x\leq 0\end{cases}$	$\frac{n_2}{n_2-2}$ si $n_2>2$	$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}$ si $n_2>4$

H Intervalos de confianza para la estimación de parámetros poblacionales

En la siguiente tabla se presentan los intervalos de confianza para los parámetros poblacionales de las distribuciones más comunes.

Parámetro poblacional	*Estimador*	*Distribución*	*Estimación puntual*	*Intervalo de confianza* $(1-\alpha)\cdot 100\%$
μ de una $N(\mu,\sigma)$ con σ^2 conocida	$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$	$\bar{X}\sim N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$	$\hat{\mu}$	$I=\left(\hat{\mu}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$
μ de una $N(\mu,\sigma)$ con σ^2 desconocida (σ^2 se estima con $\mathcal{S}^2$)	$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ $(\mathcal{S}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2)$	$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t_{n-1}$	$\hat{\mu}$ $(\hat{\sigma}^2)$	$I=\left(\hat{\mu}\pm t_{\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right)$
p de una $Bin(n,p)$	$P=\frac{\#\text{éxitos}}{\#\text{ensayos}}$	$P\sim N\left(p,\sqrt{\frac{p(1-p)}{n}}\right)$	$\hat{p}$	$I=\left(\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
λ de una $Poi(\lambda)$	$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$	$\bar{X}\sim N\left(\lambda,\sqrt{\frac{\lambda}{n}}\right)$	$\hat{\lambda}$	$I=\left(\hat{\lambda}\pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}\right)$
σ^2 de una $N(\mu,\sigma)$	$\mathcal{S}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$	$(n-1)\frac{\mathcal{S}^2}{\sigma^2}\sim\chi^2_{n-1}$	$\hat{\sigma}^2$	$I=\left(\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-1}},\frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}}\right)$
$\mu_1-\mu_2$ de $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$ con σ_1 y σ_2 conocidas	$\bar{X}_1-\bar{X}_2$	$\bar{X}_1-\bar{X}_2\sim N\left(\mu_1-\mu_2,\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	$\hat{\mu}_1-\hat{\mu}_2$	$I=\left(\hat{\mu}_1-\hat{\mu}_2\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$
$\mu_1-\mu_2$ de $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$ con σ_1 y σ_2 desconocidas y $\sigma_1^2\simeq\sigma_2^2$ ($\sigma^2=\sigma_1^2=\sigma_2^2$ se estima con $\mathcal{S}^2$)	$\bar{X}_1-\bar{X}_2$ $(\mathcal{S}^2=\frac{(n_1-1)\mathcal{S}_1^2+(n_2-1)\mathcal{S}_2^2}{n_1+n_2-2})$	$T=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\mathcal{S}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_g$ con $g=n_1+n_2-2$	$\hat{\mu}_1-\hat{\mu}_2$ $(\hat{\sigma}^2)$	$I=\left(\hat{\mu}_1-\hat{\mu}_2\pm t_{\alpha/2,g}\,\hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$
$\mu_1-\mu_2$ de $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$ con σ_1 y σ_2 desconocidas y $\sigma_1^2\neq\sigma_2^2$ (σ_1^2 y σ_2^2 se estiman con $\mathcal{S}_1^2$ y $\mathcal{S}_2^2$)	$\bar{X}_1-\bar{X}_2$ $(\mathcal{S}_1^2$ y $\mathcal{S}_2^2)$	$T=\frac{(X_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\mathcal{S}_1^2}{n_1}+\frac{\mathcal{S}_2^2}{n_2}}}\sim t_g$ con g el natural más cercano a $h=\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\hat{\sigma}_1^2/n_1\right)^2}{n_1+1}+\frac{\left(\hat{\sigma}_2^2/n_2\right)^2}{n_2+1}}-2$	$\hat{\mu}_1-\hat{\mu}_2$ $(\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2)$	$I=\left(\hat{\mu}_1-\hat{\mu}_2\pm t_{\alpha/2,g}\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_1^2}{n_2}}\right)$
p_1-p_2 de $Bin(n_1,p_1)$ y $Bin(n_2,p_2)$	P_1-P_2	$P_1-P_2\sim N\left(p_1-p_2,\sqrt{\frac{p_1(1-p_1)}{n_1}-\frac{p_2(1-p_2)}{n_2}}\right)$	$\hat{p}_1-\hat{p}_2$	$I=\left(\hat{p}_1-\hat{p}_2\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$
$\frac{\sigma_1^2}{\sigma_2^2}$ de $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$	$\frac{\mathcal{S}_1^2}{\mathcal{S}_2^2}$	$F=\frac{\mathcal{S}_1^2/\sigma_1^2}{\mathcal{S}_2^2/\sigma_2^2}\sim F_{n_1-1,n_2-1}$	$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$	$I=\left(\frac{\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}}{F_{\alpha/2;n_1-1,n_2-1}},\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}F_{\alpha/2;n_2-1,n_1-1}\right)$
$\mu_D=\mu_1-\mu_2$ de $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$ siendo datos apareados (la varianza se estima con $\mathcal{S}_D^2$)	$\bar{D}=\frac{1}{n}\sum_{i=1}^{n}D_i$ con $D_i=X_{1_i}-X_{2_i}$ $(\mathcal{S}_D^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i-\bar{D}\right)^2)$	$T=\frac{\bar{D}-\mu_D}{s_D/\sqrt{n}}\sim t_{n-1}$	$\hat{\mu}_D$ $(\hat{\sigma}_D^2)$	$I=\left(\hat{\mu}_D\pm t_{\alpha/2,n-1}\frac{\hat{\sigma}_D}{\sqrt{n}}\right)$

I Contrastes de hipótesis para parámetros poblacionales

En la siguiente tabla se presentan los contrastes de hipótesis más comunes para parámetros poblacionales de una distribución.

Contraste	*Hipótesis nula, H_0*	*Hipótesis alternativa, H_1*	*Variable*	*Valor muestral*	RA_α	RC_α
Proporción p de una binomial	$p = p_0$	$p \neq p_0$	$Z = \dfrac{P - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \sim N(0,1)$	$z_m = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$\left(-z_{\alpha/2}, z_{\alpha/2}\right)$	$\left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
		$p > p_0$			$(-\infty, z_\alpha)$	$[z_\alpha, \infty)$
		$p < p_0$			$(-z_\alpha, \infty)$	$(-\infty, -z_\alpha]$
Media μ de una normal (σ^2conocida)	$\mu = \mu_0$	$\mu \neq \mu_0$	$Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$z_m = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}$	$\left(-z_{\alpha/2}, z_{\alpha/2}\right)$	$\left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
		$\mu > \mu_0$			$(-\infty, z_\alpha)$	$[z_\alpha, \infty)$
		$\mu < \mu_0$			$(-z_\alpha, \infty)$	$(-\infty, -z_\alpha]$
Media μ de una normal (σ^2 desconocida)	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$	$t_m = \dfrac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}}$	$\left(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}\right)$	$\left(-\infty, -t_{\alpha/2,n-1}\right] \cup \left[t_{\alpha/2,n-1}, \infty\right)$
		$\mu > \mu_0$			$(-\infty, t_{\alpha,n-1})$	$[t_{\alpha,n-1}, \infty)$
		$\mu < \mu_0$			$(-t_{\alpha,n-1}, \infty)$	$(-\infty, -t_{\alpha,n-1}]$
Varianza σ^2 de una normal	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = (n-1)\dfrac{\mathcal{S}^2}{\sigma^2} \sim \chi^2_{n-1}$	$\chi^2_m = (n-1)\dfrac{\hat{\sigma}^2}{\sigma_0^2}$	$\left(\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}\right)$	$\left[0, \chi^2_{1-\alpha/2,n-1}\right] \cup \left[\chi^2_{\alpha/2,n-1}, \infty\right)$
		$\sigma^2 > \sigma_0^2$			$\left[0, \chi^2_{\alpha,n-1}\right)$	$\left[\chi^2_{\alpha,n-1}, \infty\right)$
		$\sigma^2 < \sigma_0^2$			$\left(\chi^2_{1-\alpha,n-1}, \infty\right)$	$\left[0, \chi^2_{1-\alpha,n-1}\right]$

Contrastes de hipótesis para parámetros poblacionales (continuación)

En la siguiente tabla se presentan los contrastes de hipótesis más comunes para comparación de parámetros poblacionales de dos distribución.

Contraste	*Hipótesis nula,* H_0	*Hipótesis alternativa,* H_1	*Variable*	*Valor muestral*	RA_α	RC_α
Comparación de dos proporciones p_1 y p_2	$p_1 = p_2$	$p_1 \neq p_2$	$Z = \dfrac{P_1 - P_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$	$z_m = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$\left(-z_{\alpha/2}, z_{\alpha/2}\right)$	$\left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
		$p_1 > p_2$			$(-\infty, z_\alpha)$	$[z_\alpha, \infty)$
		$p_1 < p_2$			$(-z_\alpha, \infty)$	$(-\infty, -z_\alpha]$
Dos medias normales μ_1 y μ_2 (σ_1^2 y σ_2^2 conocidas)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$Z = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$z_m = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\left(-z_{\alpha/2}, z_{\alpha/2}\right)$	$\left(-\infty, -z_{\alpha/2}\right] \cup \left[z_{\alpha/2}, \infty\right)$
		$\mu_1 > \mu_2$			$(-\infty, z_\alpha)$	$[z_\alpha, \infty)$
		$\mu_1 < \mu_2$			$(-z_\alpha, \infty)$	$(-\infty, -z_\alpha]$
Dos medias normales μ_1 y μ_2 ($\sigma_1^2 = \sigma_2^2$ y desconocidas)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\mathcal{S}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2}$	$t_m = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$\left(-t_{\alpha/2,n_1+n_2-2}, t_{\alpha/2,n_1+n_2-2}\right)$	$\left(-\infty, -t_{\alpha/2,n_1+n_2-2}\right] \cup \left[t_{\alpha/2,n_1+n_2-2}, \infty\right)$
		$\mu_1 > \mu_2$			$(-\infty, t_{\alpha,n_1+n_2-2})$	$[t_{\alpha,n_1+n_2-2}, \infty)$
		$\mu_1 < \mu_2$			$(-t_{\alpha,n_1+n_2-2}, \infty)$	$(-\infty, -t_{\alpha,n_1+n_2-2}]$
Dos medias normales μ_1 y μ_2 ($\sigma_1^2 \neq \sigma_2^2$ y desconocidas)	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\mathcal{S}_1^2}{n_1} + \dfrac{\mathcal{S}_2^2}{n_2}}} \sim t_g$ con $g = INT\left(\dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1+1} + \frac{\left(s_2^2/n_2\right)^2}{n_2+1}} - 2\right)$	$t_m = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}}$	$\left(-t_{\alpha/2,g}, t_{\alpha/2,g}\right)$	$\left(-\infty, -t_{\alpha/2,g}\right] \cup \left[t_{\alpha/2,g}, \infty\right)$
		$\mu_1 > \mu_2$			$(-\infty, t_{\alpha,g})$	$[t_{\alpha,g}, \infty)$
		$\mu_1 < \mu_2$			$(-t_{\alpha,g}, \infty)$	$(-\infty, -t_{\alpha,g}]$
Media, μ_D, de datos apareados	$\mu_D = 0$	$\mu_D \neq 0$	$T = \dfrac{\bar{D} - \mu_D}{\mathcal{S}_D/\sqrt{n}} \sim t_{n-1}$	$t_m = \dfrac{\hat{\mu}_D}{\hat{\sigma}_D/\sqrt{n}}$	$\left(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}\right)$	$\left(-\infty, -t_{\alpha/2,n-1}\right] \cup \left[t_{\alpha/2,n-1}, \infty\right)$
		$\mu_D > 0$			$(-\infty, t_{\alpha,n-1})$	$[t_{\alpha,n-1}, \infty)$
		$\mu_D < 0$			$(-t_{\alpha,n-1}, \infty)$	$(-\infty, -t_{\alpha,n-1}]$
Dos varianzas normales σ_1^2 y σ_2^2	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \dfrac{\mathcal{S}_1^2/\sigma_1^2}{\mathcal{S}_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$	$f_m = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$	$\left(\dfrac{1}{F_{\alpha/2;n_2-1,n_1-1}}, F_{\alpha/2;n_1-1,n_2-1}\right)$	$\left[0, \dfrac{1}{F_{\alpha/2;n_2-1,n_1-1}}\right] \cup \left[F_{\alpha/2;n_1-1,n_2-1}, \infty\right)$
		$\sigma_1^2 > \sigma_2^2$			$[0, F_{\alpha;n_1-1,n_2-1})$	$[F_{\alpha;n_1-1,n_2-1}, \infty)$
		$\sigma_1^2 < \sigma_2^2$			$\left(\dfrac{1}{F_{\alpha;n_2-1,n_1-1}}, \infty\right)$	$\left[0, \dfrac{1}{F_{\alpha;n_2-1,n_1-1}}\right]$