Apuntes Matemáticas II. BQ y BIO: Tema 2 Inferencia I

R. Alberich y A. Mir

19 de febrero de 2012

Muestra aleatoria simple

- Supongamos que tenemos una población con una característica a estudiar que viene dada por la variable aleatoria X.
- Diremos muestra aleatoria simple de tamaño n de la población X a un conjunto de n variables aleatorias $X_1, \ldots, X_n$ independientes e idénticamente distribuidas, es decir todas tienen la misma distribución que la variable X.
- En la práctica, lo que tendremos serán unos valores muestrales a los que denotaremos por $x_1, \ldots, x_n$ y los llamaremos realización de la muestra.

Parámetro

- La distribución de la variable aleatoria X, objeto de nuestro estudio, puede depender de un parámetro θ o de varios.
- Por ejemplo, si X es binomial, los parámetros serán n y p; si X es Poisson, el parámetro será λ; si X es geométrica, el parámetro será p y si X es normal los parámetros serán μ y σ.
- El objetivo de la estadística inferencial es obtener información de dichos parámetros, en general desconocidos, de la variable X.
- Dicha información se puede obtener de tres formas:

  estimación puntual: Hallamos un valor aproximado del parámetro.

  estimación por intervalo: Hallamos un intervalo donde el parámetro tiene un probabilidad "alta" de estar dentro de dicho intervalo.

  contraste de hipótesis: Establecemos dos hipótesis para contrastar valores del parámetro.

Estadístico

- Sean $X_1, \ldots, X_n$ n v.a. iid que forman una m.a.s. de una población.
- Un **estadístico** es una variable aleatoria que es función de la muestra.
- Un **estimador puntual** de un parámetro θ es un estadístico que da como resultado un único valor del que se espera que se aproxime a θ.
- Una **realización del estimador** $T(x_1, \ldots, x_n) = \hat{\theta}$ en una muestra se llama **estimación puntual de parámetro**.

Estimadores básicos

Consideremos una m.a.s. $X_1, \ldots, X_n$ y una realización de la misma $x_1, \ldots, x_n$. Los principales estimadores de los parámetros poblacionales que hemos visto se muestran en la tabla 5

Parámetro Poblacional	Estimador(θ)	Estimación($\hat{\theta}$)
μ_X	$\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$	$\overline{x} = \frac{\sum_{i=1}^n x_i}{n}$
σ_X	$\tilde{S}_X = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$	$\tilde{s}_X = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}$
p	$\hat{p}_X = \frac{\sum_1^n X_i}{n}$	$\frac{\sum_1^n x_i}{n}$

Tabla 1: Pricipales estimadores.

Ejemplo

Consideremos una m.a.s. X_1, X_2, X_3, X_4, X_5 del lanzamiento de un dado $(n = 5)$.
Una realización de esta muestra es $x_1 = 2, x_2 = 3, x_3 = 3, x_4 = 5, x_5 = 6$.
Sabemos que si el dado es perfecto, $\mu = 3.5$. El estadístico de esta muestra es

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

y una estimación es

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{2 + 3 + 3 + 5 + 6}{5} = \frac{19}{5} = 3.8.$$

Ejemplo

Supongamos que queremos estimar la proporción de veces que sale 3, cuyo valor teórico vale $p_3 = \frac{1}{6}$.
El estadístico en este caso es

$$\hat{p}_3 = \frac{\text{frec. de 3 en la muestra}}{5}$$

y, usando la muestra anterior, su valor será $\frac{2}{5}$.

- ¿Qué estimador es mejor?
- Para decidirlo definiremos diversas propiedades de los estimadores.
- La más inmediata es pedirles que su valor esperado sea el valor del parámetro que estima.
- Dado $\hat{\theta}$ un estimador de un parámetro poblacional θ. Diremos que $\hat{\theta}$ es **insesgado** si $E(\hat{\theta}) = \theta$.
- Es este caso la estimación puntual se dice que es insesgada.

Ejemplo

En el ejemplo del dado y para cualquier muestra de tamaño n,

$$X_1, \ldots, X_n.$$

Se tiene que :

$$E\left(\overline{X}\right) = \mu_X$$

por lo tanto $\overline{X}$ es un estimador insesgado de μ_X.

Algunos estimadores insesgados notables

Dada una m.a.s. la media, varianza (sólo en el caso de normalidad) y proporción muestrales son estimadores insesgados de sus correspondientes parámetros poblacionales. Es decir:

- $E(\overline{X}) = \mu$.
- $E(\hat{p}) = p$.
- $E(\tilde{s}^2) = \sigma^2$.

El sesgo de un estimador

Sea $\hat{\theta}$ un estimador puntual de un parámetro poblacional θ, llamaremos **sesgo** de $\hat{\theta}$ a:

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Observación: Diremos que un estimador es insesgado si y sólo si tiene sesgo cero.

La varianza y el error estándar de un estimador

- Una propiedad buena para un estimador es la carencia de sesgo.
- Pero podría suceder que tuviera una gran variabilidad.
- Entonces, aunque su valor central sea el verdadero valor del parámetro que se estima, una realización del estadístico podría estar lejos del verdadero valor del parámetro.
- Parece pues interesante emplear aquellos estimadores que tengan varianza más pequeña.
- A la desviación típica, es decir la raíz cuadrada de la varianza, de un estimador la denominaremos **error estándar del estimador**.

Eficiencia de un estimador

- Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores de un parámetro poblacional θ obtenidos de la misma muestra.
- Diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.
- O lo que es lo mismo si el error estándar de $\hat{\theta}_1$ es más pequeño que el error estándar de $\hat{\theta}_2$;

$$\sqrt{Var(\hat{\theta}_1)} < \sqrt{Var(\hat{\theta}_2)}.$$

Ejemplo:

- Sea $x_{(1)}, \ldots, x_{(n)}$ la realización ordenada de menor a mayor de una muestra de tamaño n.
- Se define la mediana muestral como
$$Me = Q_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ es par} \end{cases}$$
- Como vimos la mediana es también un valor de tendencia central, pero ¿es un buen estimador de μ?
- Se puede demostrar que cuando la población tiene distribución normal con media μ y varianza σ_X^2 entonces $E(Me) = \mu$ y $Var(Me) = \frac{\pi}{2}\frac{\sigma_X^2}{n} \approx \frac{1.57\sigma_X^2}{n}$
- Se deduce que si la muestra es de una población normal, $\overline{X}$ es más eficiente (un 57 % más eficiente) que la mediana.

Estimador más eficiente

Diremos que un estimador insesgado $\hat{\theta}$ del parámetro θ es el **estimador más eficiente** si no existe ningún otro estimador insesgado que tenga menor varianza que él (también se le denomina estimador insesgado de varianza mínima).

Algunos estimadores más eficientes

- Si la población es normal la media muestral es el estimador insesgado más eficiente de la media poblacional.
- Si la población es normal la varianza muestral es el estimador insesgado más eficiente de la varianza poblacional.
- Si la población es binomial la proporción muestral es el estimador insesgado más eficiente de la proporción poblacional.

Métodos para calcular estimadores. (**Opcional**)

Existen muchos métodos para el encontrar estimadores:

- Método de los momentos. Momento central de orden r $m_r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^r}{n}$
- El de menor error cuadrático medio $E((\hat{\theta} - \theta)^2)$
- Convergencia en probabilidad $P(|\hat{\theta}_n - \theta| < \epsilon) \to 1$
- Estimadores máximo verosímiles.
- Otras técnicas, estimación robusta, remuestreo....

Función de verosimilitud

- Sea X una v.a. tal que su distribución (densidad o función de probabilidad) depende de un parámetro desconocido λ.
- Sea $f_X(x;\lambda)$ su densidad.
- Consideremos $X_1, \ldots X_n$ una m.a.s. de X (es decir son n v.a. iid como X).
- Sea $x_1, x_2, \ldots, x_n$ una realización de la muestra.
- Entonces la función de verosimilitud de la muestra es:

$$L(\lambda|x_1, x_2, \ldots, x_n) = f_X(x_1;\lambda) \cdot \ldots \cdot f_X(x_n;\lambda)$$

Estimador máximo verosímil

- Dada una función de verosimilitud $L(\lambda|x_1,\ldots,x_n)$ de una muestra denotaremos por $\hat{\lambda} = g(x_1,\ldots,x_n)$ el punto donde se alcanza en máximo de $L(\lambda|x_1,\ldots,x_n)$.
- Es decir $L(\hat{\lambda}) = \max_{\lambda} L(\lambda|x_1,\ldots,x_n)$.
- El valor $\hat{\lambda}$ recibe el nombre de estimador máximo verosímil.

El logaritmo de la función de verosimilitud

- En ocasiones es conveniente trabajar con el logaritmo de la función de verosimilitud.
- Al ser la función logaritmo creciente, el máximo de $\log(L(\lambda|x_1,\ldots,x_n))$ y $L(\lambda|x_1,x_2,\ldots,x_n)$ es el mismo y el primero suele ser más fácil de calcular.

Ejemplo

- Sea $X_1, \ldots X_n$ una muestra con observaciones independientes, de una población Bernouilli.
- Por ejemplo se analiza el genoma a 100 personas para saber si tienen una forma de un determinado alelo de un gen.
- Se anota un 1 si tienen ese alelo y cero en cualquier otro caso.
- Sea p la proporción poblacional de personas que tienen ese alelo.
- Entonces

$$f_X(1;p) = P(X=1) = p \text{ y } f_X(0;p) = P(X=0) = 1-p = q,$$

o lo que es lo mismo

$$f_X(x) = P(X=x) = p^x q^{1-x} \text{ si } x = 0, 1$$

Ejemplo

- La función de verosimilitud es:

$$\begin{aligned} L(p|x_1,\ldots,x_n) &= f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;p) \\ &= f_{X_1}(x_1;p)\cdot\ldots\cdot f_{X_1}(x_1;p) \\ &= p^{x_1}q^{1-x_1}\cdot\ldots\cdot p^{x_n}q^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i}q^{\sum_{i=1}^n(1-x_i)} = p^{\sum_{i=1}^n x_i}q^{n-\sum_{i=1}^n x_i} \end{aligned}$$

- El valor de p que hace máxima esta probabilidad es el más verosímil o el de máxima verosimilitud de esta muestra:

$$p^{\sum_{i=1}^n x_i}q^{n-\sum_{i=1}^n x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

Ejemplo

- Aplicando el logaritmo a la función de verosimilitud

$$\begin{aligned}\log(L(p; x : 1 \ldots, x_n)) &= \log\left(p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\right) \\ &= \left(\sum_{i=1}^n x_i\right)\log p + \left(n - \sum_{i=1}^n x_i\right)\log(1-p)\end{aligned}$$

- Derivando respecto de p obtenemos que

$$\begin{aligned}(\log(L(p; x : 1 \ldots, x_n)))' &= \left(\sum_{i=1}^n x_i\right)\frac{1}{p} - \left(n - \sum_{i=1}^n x_i\right)\frac{1}{1-p} \\ &= \frac{1}{1-p}\left(\frac{\sum_{i=1}^n x_i}{p} - n\right)\end{aligned}$$

Ejemplo

- La expresión anterior es posistiva si

$$\frac{1}{1-p}\left(\frac{\sum_{i=1}^{n} x_i}{p} - n\right) \geq 0 \Leftrightarrow \left(\frac{\sum_{i=1}^{n} x_i}{p} - n\right) \geq 0 \Leftrightarrow p \geq \frac{\sum_{i=1}^{n}}{n} = \overline{x}$$

- Por lo tanto la función de verosimilitud es creciente si $p \geq \overline{x}$ y es decreciente si $p \leq \overline{x}$, y se anula en $p = \overline{x}$ donde alcanza un máximo.
- Por lo tanto

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

- Luego el estimador máximo verosímil de p es la proporción muestral $\hat{p} = \overline{x}$, es decir, $L(\hat{p}|x_1, \ldots, x_n) \geq L(p|x_1, \ldots, x_n)$ para cualquier valor $0 \leq p \leq 1$.

Estimación por intervalos

- Una estimación por intervalos de un parámetro poblacional es una regla para determinar un rango o un intervalo donde, con cierta probabilidad, se encuentre el verdadero valor del parámetro.
- La estimación correspondiente se llama estimación por intervalo.
- Más formalmente dado un parámetro θ, el intervalo (A, B) es un intervalo de confianza del $(1-\alpha)\dot{1}00\,\%$ para el parámetro θ si

$$P(A < \theta < B) = 1 - \alpha.$$

- El valor $1-\alpha$ recibe el nombre de **nivel de confianza**
- El valor $0 < \alpha < 1$ es la "*cola*" de probabilidad sobrante que normalmente se reparte por igual $(\alpha/2)$ a cada lado del intervalo.
- Es frecuente que el nivel de confianza se dé en tanto por ciento.

Intervalo de confianza para la media de una población normal: varianza poblacional conocida

En lo que sigue expondremos distintas maneras de calcular o aproximar intervalos de confianza para distintos parámetros.

- Sea $X_1, \ldots, X_n$ una m.a.s. de una v.a. X con distribución normal y $Var(X) = \sigma^2$ conocida.
- Busquemos un intervalo de confianza al *nivel de confianza* del 97.5 % para la media poblacional μ.
- Bajo estas condiciones, la variable $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ sigue una distribución normal estándar.

Ejemplo

Comencemos calculando un intervalo centrado en 0 para que la variable aleatoria normal estándar Z tenga probabilidad 0.975.

- $0.975 = P(-\delta < Z < \delta) = F_Z(\delta) - F_Z(-\delta) = 2F_Z(\delta) - 1$
- Entonces
  $F_Z(\delta) = \frac{1.975}{2} = 0.9875$
- Consultando las tablas de la distribución normal estándar,
  $F_Z(2.24) = 0.9875$ y por lo tanto $\delta = 2.24$

Ejemplo

- Luego $P(-2.24 < Z < 2.24) = 0.975$
- En resumen, hemos obtenido que

$$0.975 = P\left(-2.24 < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 2.24\right)$$

- Por lo tanto

$$P\left(\overline{X} - 2.24\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 2.24\frac{\sigma}{\sqrt{n}}\right) = 0.975$$

Ejemplo

- Hemos encontrado un intervalo de confianza para μ.
- La probabilidad de que μ se encuentre en el intervalo

$$\left(\overline{X} - 2.24\frac{\sigma}{\sqrt{n}}, \overline{X} + 2.24\frac{\sigma}{\sqrt{n}}\right)$$

es 0.975.

- Luego es un intervalo de confianza con nivel de confianza 97.5 %
- Es decir en 97.5 de cada 100 ocasiones, en que tomemos una muestra de tamaño n y bajo estas condiciones, el verdadero valor de μ se encontrará en el intervalo de confianza de esa muestra.

Ejemplo

- Supongamos que tenemos una muestra con $n = 16$ de una v.a. normal de forma que $\overline{x} = 20$ y cuya desviación típica poblacional conocida vale $\sigma = 4$.
- Entonces un intervalo de confianza al 97.5 % para μ es:

$$\left(20 - \frac{2.24 \cdot 4}{\sqrt{16}}, 20 + \frac{2.24 \cdot 4}{\sqrt{16}}\right)$$

- La probabilidad con que el verdadero valor del parámetro μ se encuentra en el intervalo $(17.76, 22.24)$ es 0.975.
- O lo que es lo mismo $P(17.76 < \mu < 22.24) = 0.975$

Interpretación del intervalo de confianza

- En el 97.5 % de la muestras de tamaño 16 el verdadero valor del parámetro μ se encontrará dentro del intervalo de confianza para esa muestra.

Fórmula general

- En general si tenemos una m.a.s. $X_1, \ldots, X_n$ de una población normal (representado por la v.a. X) con distribución normal de media μ y varianza conocida σ^2,
- el intervalo de confianza para μ al nivel de confianza $(1-\alpha)\cdot 100\,\%$ es

$$\begin{aligned}
1-\alpha &= P(z_{\alpha/2} < Z < z_{1-\alpha/2}) \\
&= P\left(z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha/2}\right) \\
&= P\left(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

Resumen: Intervalo de confianza para μ: σ^2 conocida.

Condiciones:

a) Población Normal con media μ y varianza σ^2 conocida

b) Muestra aleatoria de tamaño n

Entonces el intervalo de confianza del $100(1-\alpha)\,\%$ para μ es:

$$\left(\overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- Donde $z_{\frac{\alpha}{2}}$ es el cuantil $\frac{\alpha}{2}$, es decir $P(Z \leq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, cuando Z tiene distribución normal estándar.
- $z_{1-\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$, es decir $P(Z \leq z_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$, cuando Z tiene distribución normal estándar.
- Notemos que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$

Ejemplo Tenemos un aparato para medir volúmenes de líquido. Para saber si está bien calibrado se toman 10 muestras consistentes en rellenar un recipiente, especialmente calibrado, de un litro. Se comprueban las mediciones obteniéndose los resultados de la siguiente tabla:

Volumen en litros	Frec. Absoluta	Volumen $\times$ Frec. Absoluta
1.000	1	1.000
1.002	2	2.004
1.004	1	1.004
1.006	2	2.012
1.008	1	1.008
1.010	2	2.020
1.012	1	1.012
Total	10	10.06

Ejemplo

Supongamos que el volumen de líquido sigue una distribución normal con varianza poblacional conocida $\sigma^2 = 4$. Calcular un intervalo de confianza al 90 % para la media del volumen.

Solución: Tenemos las siguientes condiciones:

- Población de volúmenes normal varianza $\sigma^2 = 4$ conocida.
- Muestra aleatoria de tamaño $n = 10$.

- Podemos aplicar la formula anterior, para $1-\alpha = 0.9$.
- Entonces se tiene que $\alpha = 0.1$, $\frac{\alpha}{2} = 0.05$ y $1-\frac{\alpha}{2} = 0.95$.
- Calculamos la media aritmética de las observaciones $\overline{x} = \frac{10.06}{10} = 1.006$,
- Entonces el intervalo es

$$\left(1.006 + z_{0.05}\frac{2}{\sqrt{10}}, 1.006 + z_{1-0.05}\frac{2}{\sqrt{10}}\right).$$

- Consultando las tablas de la normal $P(Z \leq 1.65) = 0.9505 \approx 0.95$ entonces $z_{0.95} = 1.65$, y $z_{0.05} = -1.65$

Ejemplo

- Sustituyendo obtenemos que

$$z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 1.65\frac{2}{\sqrt{10}} = 1.0435$$
$$z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = -1.65\frac{2}{\sqrt{10}} = -1.0435$$

- Por lo que el intervalo de confianza del 90 % para la media del volumen es: $(1.006 - 1.0435, 1.006 + 1.0435) = (-0.0375, 2.0495)$
- Lo que quiere decir que en el 90 % de la ocasiones en que tomemos una muestra de tamaño 10 el volumen medio estará comprendido entre -0.0375 y 2.0495.
- En este caso hay un abuso de la suposición de normalidad en la distribución del volumen, ya que el extremo de la izquierda es negativo.

Amplitud del intervalo de confianza

- Como de todos es conocido la amplitud (longitud) de un intervalo es la diferencia entre sus extremos superior e inferior.
- Par el caso de una muestra aleatoria de poblaciones normales de tamaño n y varianza σ^2 la amplitud A es
$$A = \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \left(\overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2 z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$
- El *error* máximo, al nivel $(1-\alpha)$, que cometemos al estimar μ por $\overline{X}$ será la mitad de la amplitud del intervalo de confianza $z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- Si queremos calcular el tamaño n de la muestra para asegurarnos que el intervalo de confianza para μ al nivel $(1-\alpha)$ tiene amplitud prefijada A (o un error $\frac{A}{2}$) se puede despejar así:
$$n = \left(2 z_{1-\frac{\alpha}{2}} \frac{\sigma}{A} \right)^2$$

Observaciones:

- El intervalo está centrado en $\overline{X}$.
- Para n y $1-\alpha$ fijos, si la varianza poblacional aumenta entonces A aumenta.
- Para una varianza poblacional conocida y $1-\alpha$ fijos, si n aumenta entonces A disminuye.
- Para una varianza poblacional conocida y n fijos, si $1-\alpha$ aumenta entonces A aumenta.

Intervalo de confianza para la media poblacional: tamaños muestrales grandes

Condiciones:

- Población con media μ y varianza σ^2 conocida o si no se estima por $\tilde{S}^2$
- Muestra aleatoria de tamaño n grande (criterio $n \geq 30$)

Entonces el intervalo de confianza del $100(1-\alpha)\,\%$ para μ es:

$$\left(\overline{X} + z_{\frac{\alpha}{2}} \frac{\tilde{S}}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\tilde{S}}{\sqrt{n}} \right)$$

En caso de que σ sea conocida pondremos σ en lugar de $\tilde{S}$

Ejemplo:

- Se tomó una muestra de 147 expertos en informes de impacto ambiental y se les pidió que calificasen en una escala de 1 (totalmente en desacuerdo) a 10 (totalmente de acuerdo) la siguiente afirmación: "A veces utilizo técnicas de investigación que garantizan la obtención de los resultados que mi cliente o jefe desea".
- La calificación media de la muestra fue 6.06 y la desviación típica muestral fue 1.43. Se pide calcular un intervalo de confianza al 90 % para la media de las puntuaciones.

Solución:

- El enunciado no nos asegura que la población sea normal pero como el tamaño de la población es grande podemos aplicar el resultado anterior.
- Tenemos $n = 147$, $\overline{x} = 6.06$, $\tilde{s} = 1.43$, $1 - \alpha = 0.9$ entonces, utilizando las tablas de la normal estánadar $\frac{\alpha}{2} = 0.05$ y por lo tanto $z_{1-0.05} = z_{0.95} \approx z_{0.9505} = 1.65$. Con R el valor extacto es `qnorm(0.95)`=1.644854.
- El intervalo para la media poblacional de las puntuaciones al nivel de confianza del 90 % es

$$\left(6.06 - 1.65\frac{1.43}{\sqrt{147}}, 6.06 + 1.65\frac{1.43}{\sqrt{147}}\right) = (5.8654, 6.2546)$$

Distribución t de Student

- Si queremos calcular un intervalo de confianza para μ en una población normal con varianza poblacional desconocida necesitamos una nueva distribución: la t de Student[1]
- Dada una muestra de n observaciones con media muestral $\overline{X}$ y desviación típica muestral $\tilde{S}_X$ procedente de una población normal con media μ la variable aleatoria:

$$t = \frac{\overline{X} - \mu}{\frac{\tilde{S}_X}{\sqrt{n}}}$$

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

[1] Student es el alias del estadístico William Sealy Gosset

Propiedad

- La distribución t de Student se aproxima a la distribución la normal estándar si el número de grados de libertad es grande. Su función de densidad es simétrica respecto al origen como la de la normal estándar.
- Es decir si t_ν es una v.a. que sigue la distribución t de Student con ν g.l. entonces:

$$P(t_\nu \leq -t) = 1 - P(t_\nu \leq t)$$

- $E(t_\nu) = 0$ si $\nu > 1$ y $Var(t_\nu) = \frac{\nu}{\nu-2}$ si $\nu > 2$

Notación

- Sea t_ν una v.a. que sigue una distribución t de Student con ν g.l. Denotaremos por $t_{\nu,\alpha}$ al valor para el que se verifica que:

$$P(t_\nu \leq t_{\nu,\alpha}) = \alpha.$$

- Luego $t_{\nu,\alpha}$ es el α cuantil de una t de Student con ν g.l. y $t_{\nu,\alpha} = -t_{\nu,1-\alpha}$.

Gráficas

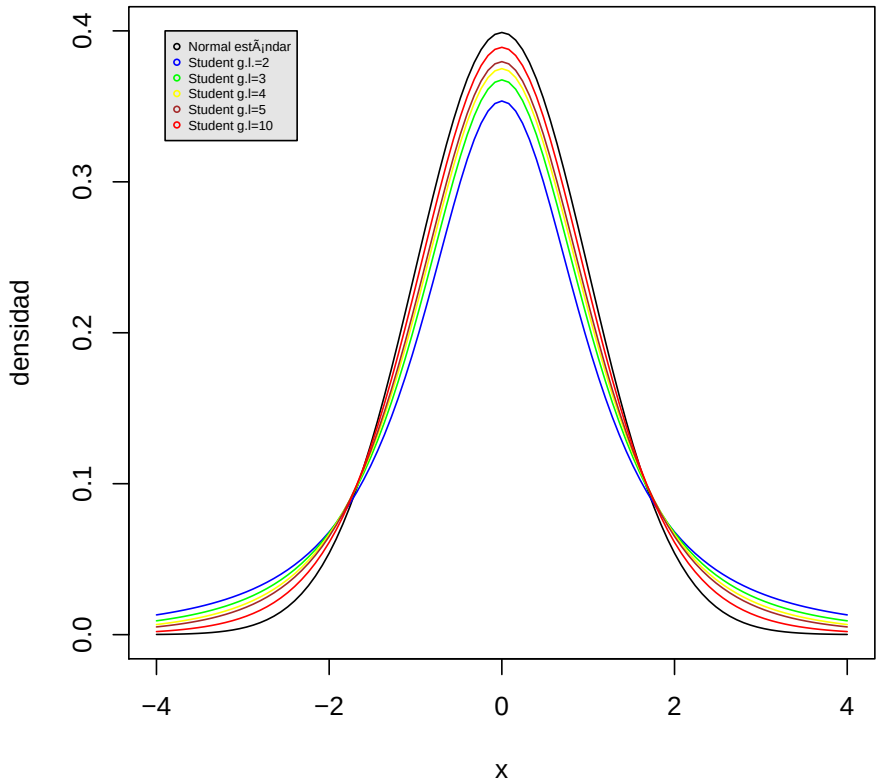

Figura 1: Gráfica de la función de densidad de la t_ν para distintos grados de libertad (g.l.)

Demo de R para la distribución t de Student

Con el código R:

```
library("TeachingDemos")
vis.t()
```

podemos ver las distintas formas que adopta la distribución t de Student para distintos grados de libertad.
En la ventana adjunta marcad la casilla de la distribución normal estándar para comparar la forma de ambas distribuciones.
Comprobad que para grados de libertad altos la t de Student se aproxima a la normal estándar.

Intervalo de confianza para la media de una población normal: varianza poblacional desconocida

Condiciones:

- Muestra aleatoria de n observaciones independientes.
- Población normal varianza desconocida

Entonces si $\overline{X}$ y $\tilde{S}_X$ son respectivamente la media y la desviación típica muestrales un intervalo de confianza al nivel $(1-\alpha)100\,\%$ para la media de la población μ es:

$$\left(\overline{X} + t_{n-1,\frac{\alpha}{2}} \frac{\tilde{S}_X}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{\tilde{S}_X}{\sqrt{n}}\right)$$

Siendo $t_{n-1,\frac{\alpha}{2}}$ y $t_{n-1,\frac{\alpha}{2}}$ los cuantiles de una v.a. t_{n-1} con distribución t de Student con n-1 g.l., respectivamente.

Ejercicio

Demostrar que la probabilidad con que μ se encuentra en el intervalo anterior es $1 - \alpha$.

Ejemplo:

La empresa *RX-print* ofrece una impresora de altísima calidad para la impresión de radiografías. En su publicidad afirma que sus *cartuchos* imprimirán un promedio de 500 radiografías*; donde el asterisco remite a una nota a pie de página donde afirma que:
"`Datos técnicos: Muestra mensual de tamaño` $n = 25$ `población supuesta normal nivel de confianza del 90 %`".
Una organización de radiólogos desea comprobar estas afirmaciones y toma también una muestra al azar de tamaño $n = 25$ obteniendo como media $\overline{x} = 518$ páginas y una desviación estándar $\tilde{s}_X = 40$. Verificar si con esta muestra la media poblacional que afirma el fabricante cae dentro del intervalo de confianza del 90 %.

Ejemplo

Solución: El problema se reduce a calcular, bajo las condiciones que afirma el fabricante el intervalo de confianza para μ con $\alpha = 0.1$.
Mirando en las tablas de la t de Student para $n-1 = 24$ g.l. tenemos que $t_{n-1,1-\frac{\alpha}{2}} = t_{24,1-0.05} = t_{24,1-0.95} = 1.71$ y $t24, 0.04 = -1.71$.
El intervalo para la media al $90\,\%$ es

$$\left(518 - 1.71\frac{40}{\sqrt{25}}, 518 + 1.71\frac{40}{\sqrt{25}}\right) = (504.32, 531.68)\,.$$

Es este caso la afirmación del fabricante queda contradicha por la muestra pues 500 cae fuera del intervalo. En cualquier caso se equivoca a favor del consumidor.

Intervalos de confianza para una proporción: Ejemplo

El procedimiento es similar al caso de la media. Comencemos con un ejemplo.

- En una muestra aleatoria de 500 familias con niños en edad escolar se encontró que 340 introducen fruta de forma diaria en la dieta de sus hijos.
- Encontrar un intervalo de confianza del 95 % para la proporción actual de familias de esta ciudad con niños en edad escolar que incorporan fruta fresca de forma diaria en la dieta de sus hijos.
- Tenemos una población binomial donde los éxitos son las familias que aportan fruta de forma diaria a la dieta de sus hijos.
- Sea X el número de familias con hijos en edad escolar que aportan diariamente fruta a su dieta en una muestra aleatoria de tamaño n.

Ejemplo

- Entonces X sigue una distribución binomial con n repeticiones y probabilidad de éxito p (proporción poblacional de familias que aportan fruta a la dieta).
- Si llamamos $\hat{p}_X = \frac{X}{n}$ a la proporción muestral, sabemos que $Z = \frac{\hat{p}_X - p}{\sqrt{\frac{p(1-p)}{n}}}$ sigue aproximadamente una distribución normal estándar.
- Pero como es evidente no conocemos p así que no tenemos más remedio que aproximar el denominador

$$\sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}$$

- Si la muestra es grande $Z = \frac{\hat{p}_X - p}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}}$ seguirá siendo aproximadamente normal estándar.

Intervalos de confianza para la proporción poblacional: (muestras grandes)

Condiciones:

- Una muestra aleatoria de tamaño n grande.
- Población Bernouilli con proporción de éxitos p (desconocida).

Bajo estas condiciones y si $\hat{p}_X$ es la proporción de éxitos en la muestra, un intervalo de confianza del parámetro p al nivel $(1-\alpha)100\,\%$ de confianza es

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

Criterio: los intervalos de confianza anteriores son fiables si $n \geq 40$.

Observaciones

- El intervalo de confianza anterior está centrado en la proporción muestral.
- Cuando n crece se reduce la amplitud del intervalo de confianza.
- La amplitud del intervalo de confianza es $A = 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}$
- De la fórmula anterior no podemos determinar el tamaño de la muestra sin conocer $\hat{p}_X$, así que nos podremos en el caso peor:
  El máximo de $\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}$ se alcanza en $\hat{p}_X = 0.5$ y en este caso vale $\sqrt{\frac{0.5(1-0.5)}{n}}$.
- Por lo tanto, en el peor de los casos, el tamaño de la muestra es $n = \frac{0.25z^2_{1-\frac{\alpha}{2}}}{(A/2)^2}$ para que la amplitud del correspondiente intervalo de confianza sea A como máximo.

Observación

Por este motivo, en las especificaciones o detalles técnicos de las encuestas se suele leer, por ejemplo:
"Universo población Balear mayor de 18 años. Encuesta telefónica, selección aleatoria de tamaño mil, error en las proporciones $\pm 3\,\%$ con una confianza del 95 % supuesto que $p = q = \frac{1}{2}$"

Intervalo de confianza para la varianza de una población normal

- Si tenemos una población normal con varianza σ^2 y una muestra aleatoria de tamaño n de esta población con varianza muestral $\tilde{S}_X^2$ entonces el estadístico

$$\chi^2_{n-1} = \frac{(n-1)S_X^2}{\sigma^2}$$

sigue una distribución χ^2 con $n-1$ g.l.

- **Notación** Si χ^2_ν es una v.a. que tiene distribución χ^2 con ν g.l. denotaremos por $\chi^2_{\nu,\alpha}$ al valor que verifica:

$$P(\chi^2_\nu \leq \chi^2_{\nu,\alpha}) = \alpha$$

- Es decir el cuantil α de una v.a. con distribución χ^2_ν. Estos valores están tabulados para distintos g.l. en la tabla de la distribución χ^2.

Ejemplo

- Sea χ^2_{10} una v.a. que tiene distribución χ^2 con 10 g.l.
- Entonces $\chi^2_{10,0.995} = 25.19$ y $\chi^2_{10,0.005} = 2.16$, es decir

$$P(\chi^2_{10} \leq 25.19) = 0.995 \text{ y } P(\chi^2_{10} \leq 2.16) = 0.005$$

- Además tendremos que

$$\begin{aligned} P(2.16 \leq \chi^2_{10} \leq 25.19) &= P(\chi^2_{10} \leq 25.19) - P(\chi^2_{10} \leq 2.16) \\ &= 0.995 - 0.005 = 0.99 \end{aligned}$$

En general

- En general dado α entre 0 y 1, tendremos que

$$1-\alpha = P(\chi^2_{\nu,\frac{\alpha}{2}} \leq \chi^2_{\nu} \leq \chi^2_{\nu,1-\frac{\alpha}{2}}).$$

- Si tenemos una muestra de tamaño n de una población normal con desviación típica muestral $\tilde{S}^2_X$, dado un nivel de confianza $1-\alpha$ tendremos que $\chi^2_{n-1} = \frac{(n-1)\tilde{S}^2_X}{\sigma^2}$.

En general

- Entonces:

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \chi^2_{n-1} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) \\ &= P\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{(n-1)S_X^2}{\sigma^2} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) \\ &= P\left(\frac{(n-1)\tilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\tilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) \end{aligned}$$

- Luego, bajo estas condiciones, un intervalo de confianza para la varianza poblacional del $(1-\alpha)100\,\%$ es

$$\left(\frac{(n-1)\tilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\tilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right).$$

Intervalo de confianza para la varianza de una población normal

Condiciones

- Población normal
- Muestra aleatoria de tamaño n con varianza muestral S_X^2

Entonces un intervalo de confianza del $(1-\alpha)100\,\%$ es

$$\left(\frac{(n-1)\tilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\tilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \right).$$

- Donde $\chi^2_{n-1,\frac{\alpha}{2}}$ es el valor que verifica

$$P(\chi^2_{n-1} < \chi^2_{n-1,\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

- Mientras que $\chi^2_{n-1,1-\frac{\alpha}{2}}$ es el valor tal que

$$P(\chi^2_{n-1} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}.$$

- Donde χ^2_{n-1} es una v.a. que sigue una distribución χ^2 con $n-1$ g.l.
- **Observación:** El intervalo de confianza para σ^2 no está centrado en $\tilde{S}^2_X$.

Ejemplo

- Un índice de calidad de un reactivo químico es el tiempo que tarda en actuar.
- El estándar es que el tiempo no debe ser superior a los 30 segundos.
- Se supone que la distribución del tiempo de actuación del reactivo es aproximadamente normal. Se realizan 30 pruebas, que forman una muestra aleatoria, en las que se mide el tiempo de actuación del reactivo.
- Los tiempos fueron:
  12, 13, 13, 14, 14, 14, 15, 15, 16, 17, 17, 18, 18, 19, 19, 25, 25, 26, 27, 30, 33, 34, 35, 40, 40, 51, 51, 58, 59, 83.
- Se pide calcular un intervalo de confianza para la varianza al nivel $95\,\%$.

Solución

- Sea X el tiempo de reacción. Haciendo los cálculos tenemos que (redondeando al segundo decimal):
  $\overline{x} = 28.37$ y $\tilde{s}_X = 17.37$.
- Como $1 - \alpha = 0.95$ tenemos que $\frac{\alpha}{2} = 0.025$, entonces mirando en las tablas de la χ^2 (y redondeando también al segundo decimal)

$$\chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{29,0.975} = 45.72 \text{ y } \chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{29,0.025} = 16.05.$$

- Por lo tanto un intervalo de confianza del 95 % para σ^2 es

$$\left(\frac{(30-1) \cdot 17.37^2}{45.72}, \frac{(30-1) \cdot 17.37^2}{16.05} \right) = (191.26, 544.96)$$

- Es decir

$$P(191.26 \leq \sigma^2 \leq 544.96) = 0.95.$$