Introducción a la Probabilidad y Estadística

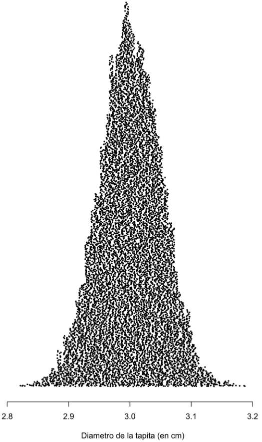
Clase XVI

Georgina Flesia

Contenidos

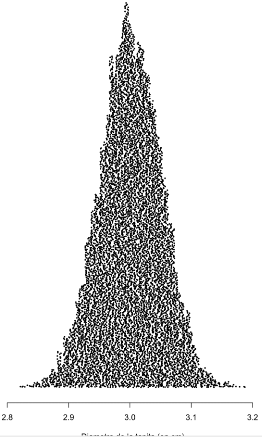
Ejemplo

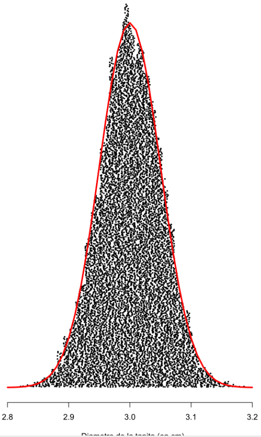
Supongamos tener una población de tapitas de gaseosa. Y observamos su diámetro. Cada punto representa una tapita muestreada de la población



Ejemplo

Un modelo paramétrico consiste en suponer la fórmula de $p(x)$ conocida, excepto por algunos parámetros. Por ejemplo, podemos suponer que es $N(\mu, \sigma^2)$





Parámetros

Definición

En un modelo paramétrico, los parámetros son los valores que determinan la densidad de la población $p(x)$ de la variable aleatoria medida en el modelo.

- Denotaremos un parámetro general por la letra θ, y la densidad correspondiente por $p(x; \theta)$.
- Si el modelo es discreto $p(x; \theta)$ denota la función de probabilidad puntual.
- En este estadio de nuestro estudio, los parámetros son los únicos valores desconocidos de la densidad (función de frecuencia) marginal o conjunta.

Estimación Puntual

Como no conocemos los parámetros de la población, el problema central es como estimarlos.

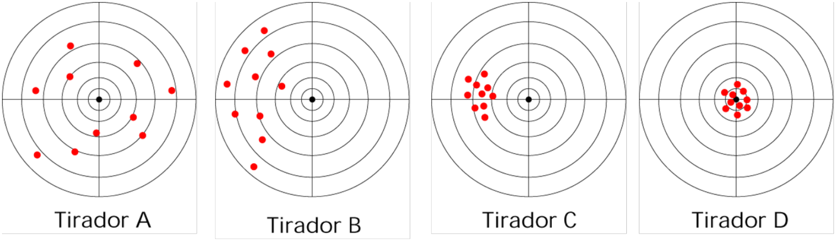
Definición

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Entonces se dice que la función de la muestra

$$\hat{\theta} = h(X_1, \cdots, X_n)$$

es un estimador puntual de θ si $\hat{\theta}$ es una variable aleatoria.

Precisión y exactitud



Estimador insesgado y no eficiente	**Estimador sesgado y no eficiente**	**Estimador sesgado y eficiente**	**Estimador insesgado y eficiente**

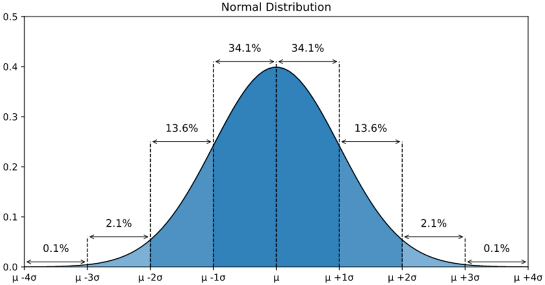
Bondad de un estimador

- Un estimador es el método para calcular una estimación.
- Características deseables en estimadores son
  1. Sesgo nulo
  2. consistencia
  3. varianza mínima en la clase de los insesgados
- Dentro de un conjunto de estimadores, podemos elegir como el mejor el que tenga el menor Error Cuadrático Medio.
- Si consideramos la clase de los insesgados, el ECM es la varianza del estimador.
- En muchas ciencias, no se reporta el valor estimado en un experimento, sino que también se reporta el error estándar

$$\hat{\theta}_{\text{obs}} \pm \sqrt{Var(\hat{\theta})}$$

Observaciones

- ¿Porqué reportar la estimación y el error?
- Porque se sabe que el valor estimado no es el valor real del parámetro, sino un valor cercano.
- Y alguna estimación de cual lejos está del verdadero valor es de interés.



Observaciones

- Cuando uno elige un estimador insesgado de mínima varianza en la clase de los insesgados, está diciendo que el método para calcular la estimación es
  - exacto
  - y con precisión máxima en esa clase.

  Y el error estándar dá idea de esa precisión. Pero no dá confianza.
- Entonces una alternativa es calcular lo que llamaremos INTERVALO DE CONFIANZA.

Estimación por intervalo

Definición

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Entonces, dado un valor α se dice que el par de variables aleatorias (funciones de la muestra)

$$L(X_1, \cdots, X_n) \qquad R(X_1, \cdots, X_n)$$

que cumplen

$$P(L(X_1, \cdots, X_n) \leq \theta \leq R(X_1, \cdots, X_n)) = 1 - \alpha$$

es un estimador por intervalo de θ de nivel $(1 - \alpha)$.

Estimación por intervalo

Definición

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Sea $\hat{\theta}$ es un estimador puntual de θ y $(L(X_1, \cdots, X_n), R(X_1, \cdots, X_n))$ un estimador por intervalo de de θ de nivel $(1-\alpha)$. Dada una observación $x_1, \cdots, x_n$,

- el valor de θ estimado es $\hat{\theta}(x_1, \cdots, x_n)$
- el intervalo estimado es $[L(x_1, \cdots, x_n), R(x_1, \cdots, x_n)]$

Observaciones

- Un Intervalo de Confianza (IC) para el parametro θ permite tener una medida de la CONFIABILIDAD y PRECISION de la estimación del parámetro.
- La PRECISIÓN de un IC tiene que ver con su longitud: cuanto menor sea su longitud, mayor es la precisión.
- La CONFIABILIDAD es medida con el nivel de confianza del intervalo, que denotaremos con $(1 - \alpha)$

Observaciones

- Los niveles mas usados son de 0.90 , 0.95 y 0.99. Cuanto mayor sea el nivel de confianza, mayor es la chance de que el IC contenga al verdadero valor poblacional.
- Luego es bueno pedirle a un IC que tenga una longitud pequeña y una alta confiabilidad de contener al parámetro poblacional.

Método del pivote

Sea $X_1, \cdots, , X_n$ una muestra aleatoria con distribución que depende de un parámetro θ, desconocido. Sea $h(X_1, \cdots, , X_n)$ una variable aleatoria con distribución conocida que no depende de θ. Sea $(1-\alpha)$ el nivel del intervalo a construir

1. Calcular los valores a y b tales que

$$P(a \leq h(X_1, \cdots, , X_n) \leq b) = 1 - \alpha$$

2. A partir de la expresión del evento $(a \leq h(X_1, \cdots, , X_n) \leq b$ hay que tratar de obtener $(L(X_1, \cdots, X_n), R(X_1, \cdots, X_n))$ tales que

$$P(L(X_1, \cdots, X_n) \leq \theta \leq R(X_1, \cdots, X_n)) = 1 - \alpha$$

3. Luego $[L(X_1, \cdots, X_n), R(X_1, \cdots, X_n)]$ es un IC aleatorio para θ con un nivel de confianza $(1-\alpha)$

Definición:

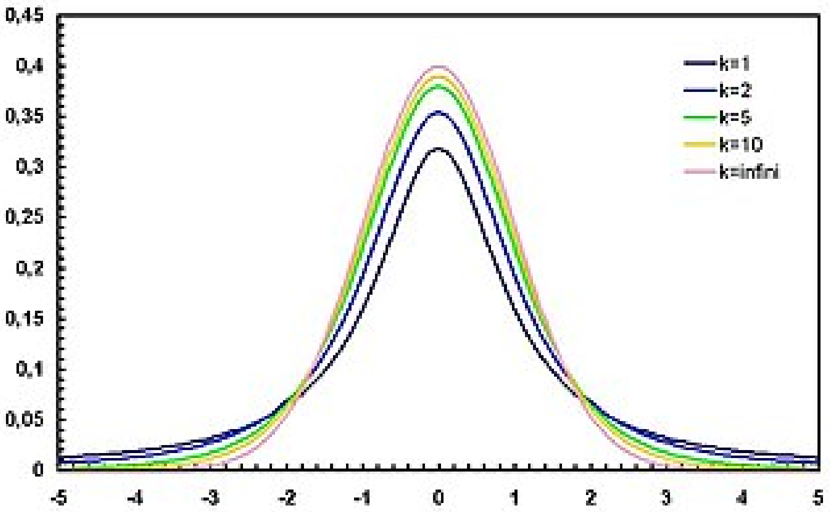
Sean X yZ dos variables aleatorias independientes tales que $Z \sim N(0;1)$ y $X \sim \chi^2_\nu$. Entonces

$$T = \frac{Z}{\sqrt{X/\nu}}$$

se dice que tiene distribución t-student con ν grados de libertad. Densidad de la distribución t de student.

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(v/2)}(1+\frac{t^2}{v})^{-(v+1)/2}, \quad \forall t \in \mathbb{R}$$

Distribución t de Student

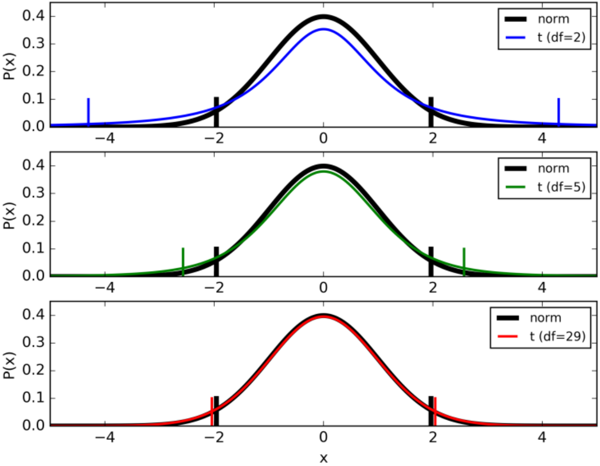


Distribuciones derivadas de la normal

Sea $X_1, X_2, \ldots, X_n$ una m.a. con distribución $N\left(\mu, \sigma^2\right)$, con μ y σ^2 desconocida.

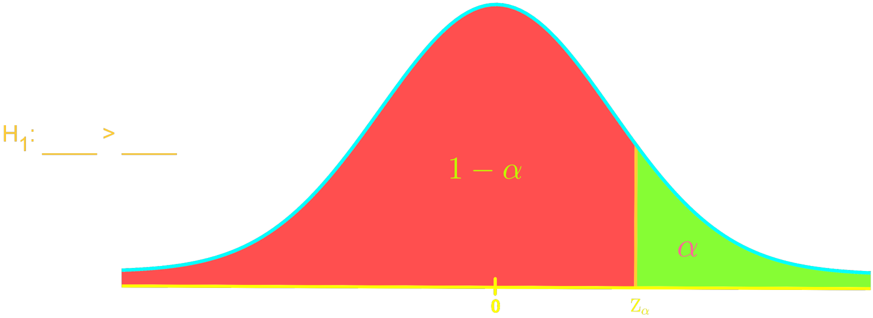
- Entonces
$$Z = \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma}$$
es una variable aleatoria normal estándar
- $$X = (n-1)\frac{S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$
es una variable aleatoria con distribución χ^2 con n-1 grados de libertad.
- X e Z son independientes
- $$\sqrt{n}\frac{(\overline{X} - \mu)}{S}$$
tiene distribución t de Student con n-1 grados de libertad.

Distribución t de Student



Notación:

Denotaremos con z_α al valor crítico, hallado en la tabla de la distribución Normal Estándar, tal que: $\Phi(z_\alpha) = 1 - \alpha$



Intervalos de confianza para la media

CASO A: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ conocida

Sea $X_1, X_2, \ldots, X_n$ una m.a. con distribución $N\left(\mu, \sigma^2\right)$, con σ^2 conocida.

- 1ro) Conocemos la distribución del promedio muestral $\bar{X}$ y esta es:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Al estandarizar $\bar{X}$ obtenemos una v.a. con distribución $\mathrm{N}(0,1)$:

$$h\left(X_1, X_2, \ldots, X_n; \theta\right) = \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

Intervalos de confianza para la media

CASO A: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ conocida

- 2do) Fijado un nivel de confianza ($1 - \alpha$), hay que buscar en la tabla normal estándar los valores de a y b tales que cumplan

$$P(a \leq \frac{(\bar{X} - \mu)}{\sigma}\sqrt{n} \leq b) = 1 - \alpha$$

  Estos valores son $a = -z_{\alpha/2}$ y $b = z_{\alpha/2}$.
- 3ro) Despejando μ resulta

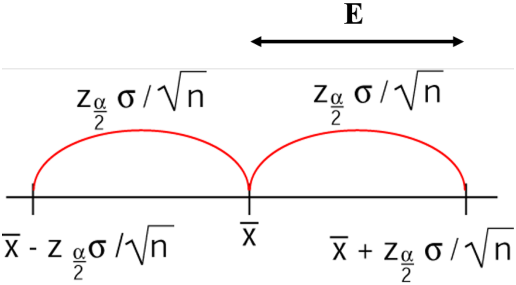
$$P\left(\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) = (1 - \alpha)$$

Intervalos de confianza para la media

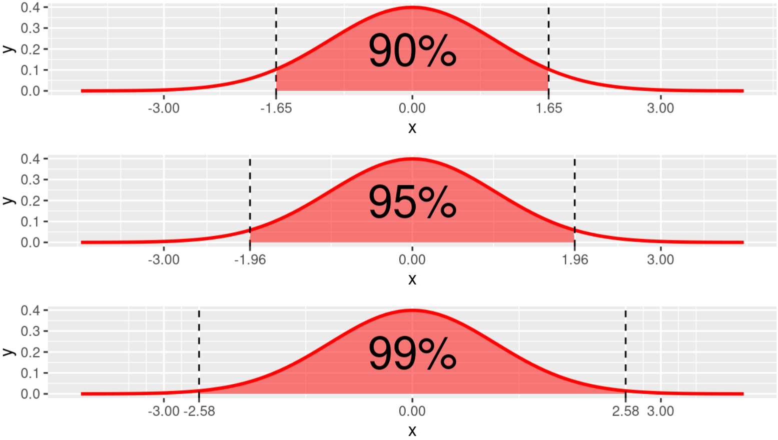
CASO A: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ conocida

Por lo tanto un IC aleatorio de nivel $(1-\alpha)$ para $\theta = \mu$ es:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

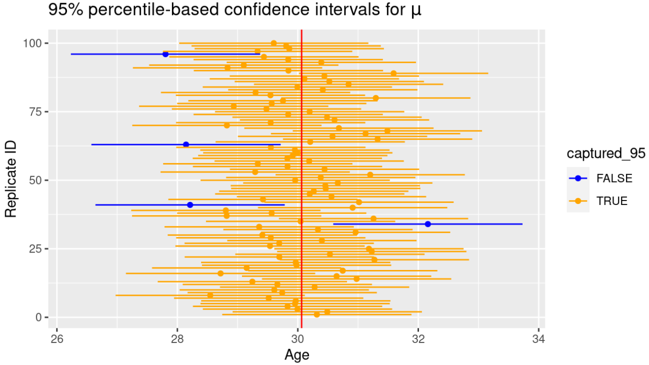


CASO A: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ conocida



Observaciones

- Un Intervalo de Confianza (IC) para el parámetro θ es un par de variables aleatorias que permite tener una medida de la CONFIABILIDAD y PRECISIÓN de la estimación del parámetro.



Observaciones

- El pivote que usamos para calcular el intervalo anterior permite construir muchos intervalos asimétricos con la misma CONFIABILIDAD $(1-\alpha)$.
- El intervalo que dimos es simétrico y puede verse que tiene la mayor PRECISIÓN. Es decir, es el intervalo de menor longitud que puede crearse con ese pivote, para un n y α fijo.
- La longitud de este intervalo es

$$L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Si se quiere obtener un intervalo de confianza de longitud a lo sumo L y una confiabilidad $(1-\alpha)$ para μ entonces hay que tomar

$$n \geq \left(2z_{\alpha/2}\frac{\sigma}{L}\right)^2$$

Observaciones

- Dada una muestra observada, $x_1, \cdots, x_n$ podemos construir el intervalo observado.
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
- El método con el que lo construimos tiene un nivel de confianza de $(1-\alpha)$. La observación, no.
- ¿Que podemos decir entonces? Que si construimos 100 intervalos observados con muestras independientes, esperamos que no mas que α% de esas muestras pueden no contener el verdadero parámetro.
- Pero no podemos decir nada de una observación específica porque no tenemos conocimiento del verdadero parámetro.

Ejemplo

Supongamos que cuando se transmite una señal con valor μ desde la ubicación A, el valor recibido en la ubicación B está distribuido normalmente con media μ y varianza 4. Es decir, si se envía μ, entonces el valor recibido es $\mu + N$ donde N, que representa el ruido, es normal con media 0 y varianza 4. Para reducir el error, supongamos que se envía el mismo valor 9 veces. Si los valores sucesivos recibidos son 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, construyamos un intervalo de confianza del 95 por ciento para μ.

Ejemplo

A partir de los valores de la muestra obtenemos

$$\overline{x} = \frac{81}{9} = 9$$

El valor critico de un nivel 95% para la distribución normal es $z_{0.25} = 1.96$ Por lo cual el intervalo de confianza estimado para μ es

$$[\overline{x} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{x} + z_{\alpha/2}\sigma/\sqrt{n}] = [9 - 1.96\ \frac{2}{3}, 9 + 1.96\ \frac{2}{3}]$$
$$\sim [7.69, 10.31]$$

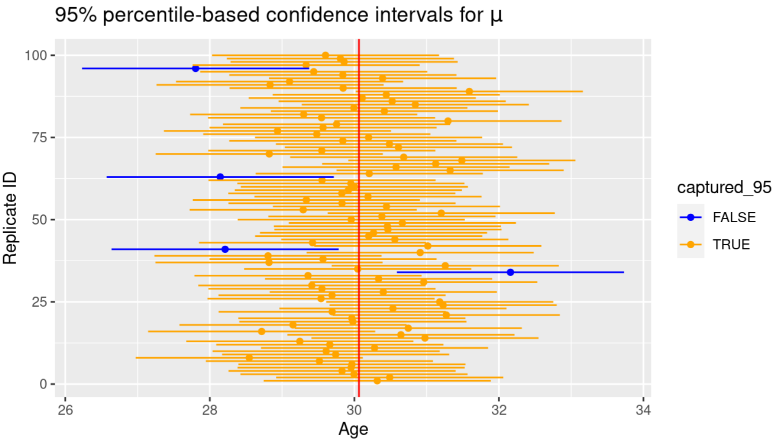
Por lo cual se suele reportar el intervalo observado $[7.69, 10.31]$ o el valor de la media muestral mas o menos el error cometido 9 ± 1.31 y se dice que tenemos una confianza del 95% de que el verdadero mensaje se encuentra en este intervalo.

Ejemplo

La interpretación de este resultado es delicada. Pueden ocurrir dos cosas

- El intervalo observado $[7.69, 10.31]$ contiene el verdadero valor de μ
- Nuestra muestra es una de las pocas muestras para las cuales $\overline{x}$ no esta a 1.31 puntos del valor de μ. Solo un 5% de todas las muestras dan estos resultados incorrectos.

Por lo tanto, no podemos asegurar que $\mu \in [7.69, 10.31]$, pero tenemos un 95% de confianza de que lo contiene, pues llegamos a estos números por un procedimiento que falla solo el 5% de las veces.


95% percentile-based confidence intervals for µ
Replicate ID
100
75
50
25
0
26
28
30
32
34
Age
captured_95
FALSE
TRUE

Ejemplo

Por experiencia pasada se sabe que los pesos de los salmones criados en una incubadora comercial son normales con una media que varía de una temporada a otra, pero con una desviación estándar que se mantiene fija en 0.3 libras. Si queremos estar un 95 por ciento seguros de que nuestra estimación del peso medio de un salmón de la temporada actual es correcta dentro de ± 0.1 libras, ¿qué tamaño de muestra se necesita?

Ejemplo

El valor critico de un nivel 95% para la distribución normal es $z_{0.25} = 1.96$. Observemos que el error en la estimación en un intervalo del 95% es

$$1.96\sigma/\sqrt{n} = .588/\sqrt{n}$$

Por lo cual si se quiere tener una certeza del 95% de que $\overline{x}$ este a lo sumo 0.1 de μ tiene que ocurrir que

$$.588/\sqrt{n} \leq 0.1$$

Esto es

$$\sqrt{n} \geq 5.88$$

por lo cual

$$n \geq 34.57$$

Esto es, un tamaño muestral de 35 o mas grande va a ser suficiente.

Intervalos de confianza para la media

CASO B: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ desconocida

Sea $X_1, X_2, \ldots, X_n$ una m.a. con distribución $N\left(\mu, \sigma^2\right)$, con σ^2 desconocida.

- 1ro) Dado que σ^2 es desconocido, ya no podemos basar nuestro intervalo en el hecho de que $\sqrt{n}(\overline{X} - \mu)/\sigma$ es una variable aleatoria normal estándar. Sin embargo, si $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ es la varianza muestral, entonces

$$h\left(X_1, X_2, \ldots, X_n; \theta\right) = \sqrt{n}\frac{(\overline{X} - \mu)}{S} \sim t_{n-1}$$

tiene distribución t con n-1 grados de libertad.

Intervalos de confianza para la media

CASO B: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ desconocida

- 2do) Fijado un nivel de confianza ($1-\alpha$), hay que buscar en la tabla de la distribución t de student los valores de a y b tales que cumplan

$$P(a \leq \frac{(\bar{X}-\mu)}{S}\sqrt{n} \leq b) = 1-\alpha$$

Estos valores son $a = -t_{\alpha/2,n-1}$ y $b = t_{\alpha/2,n-1}$.

- 3ro) Despejando μ resulta

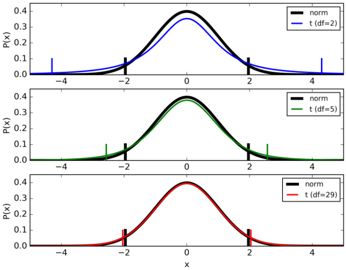
$$P\left(\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)\right) = (1-\alpha)$$

Intervalos de confianza para la media

CASO B: muestra $N(\mu, \sigma^2)$ con $\sigma > 0$ desconocida

Por lo tanto un IC aleatorio de nivel $(1-\alpha)$ para $\theta = \mu$ es:

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$



Ejemplo

Supongamos que cuando se transmite una señal con valor μ desde la ubicación A, el valor recibido en la ubicación B está distribuido normalmente con media μ y varianza DESCONOCIDA. Es decir, si se envía μ, entonces el valor recibido es $\mu + N$ donde N, que representa el ruido, es normal con media 0 y varianza σ^2. Para reducir el error, supongamos que se envía el mismo valor 9 veces. Si los valores sucesivos recibidos son 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5, construyamos un intervalo de confianza del 95 por ciento para μ usando la distribución t de student.

Ejemplo

A partir de los valores de la muestra obtenemos

$$\overline{x} = \frac{81}{9} = 9 \qquad s^2 = \frac{\sum_i^2 x_i - 9(\overline{x})^2}{8} = 9.5$$

Por lo cual $s = 3.082$. El valor critico de un nivel 95% para la distribución t con 8 grados de libertad es $t_{0.25,8} = 2.306$ Por lo cual el intervalo de confianza estimado para μ es

$$[\overline{x} - t_{\alpha/2,n-1} s/\sqrt{n}, \overline{x} + t_{\alpha/2,n-1} s/\sqrt{n}] = [9 - 2.306\ \frac{3.082}{3}, 9 + 2.306\ \frac{3.082}{3}]$$
$$\sim [6.63, 11.37]$$

el cual es un intervalo mas largo que el calculado conociendo la varianza, y usando la distribución normal $[7.69, 10.31]$. Aun si el s estimado fuese el valor 2, el uso de la distribución t da un intervalo mas grande que usando la normal.

$$[9 - 2.306\frac{2}{3}, 9 + 2.306\frac{2}{3}] = [7.46, 10.54]$$

Ejemplo (Ejercicio de Parcial)

La calibración de una bascula debe ser revisada al pesar 25 veces un espécimen de 10 Kg. Suponga que los resultados de los diferentes pesos son independientes entre si y que la variable peso esta normalmente distribuida con un desvío estándar $\sigma = 0.20$ Kg. Sea μ el verdadero valor medio de lectura de peso de la bascula.

a) ¿Cual es el nivel de confianza para el intervalo $\overline{x} \pm 2.81 \frac{\sigma}{\sqrt{n}}$ para μ?

b) ¿Cual es el valor de $z_{\alpha/2}$ para un IC del 99,7 % para μ?

c) ¿Que tan grande debería ser el tamaño de muestra tal que la longitud del IC del 95 % para μ sea a lo sumo de 0.05?

d) Si de la muestra observada se obtuvo un promedio y desvío estándar muestrales de $\overline{x} = 10.30$ Kg y $s_{n-1} = 0.19$ Kg respectivamente, obtenga un IC del 95 % para μ. Compare el intervalo obtenido con el intervalo del inciso anterior.

Intervalos de confianza para la varianza

CASO C: muestra $N(\mu, \sigma^2)$ con μ $\sigma > 0$ desconocida

Sea $X_1, X_2, \ldots, X_n$ una m.a. con distribución $N\left(\mu, \sigma^2\right)$, con μ y σ^2 desconocidos.

- 1ro) Sabemos que

$$h\left(X_1, X_2, \ldots, X_n; \theta\right) = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- 2do) Fijado un nivel de confianza ($1-\alpha$), hay que buscar en la tabla de la distribución χ^2_{n-1} los valores de a y b tales que cumplan

$$P(a \leq (n-1)\frac{S^2}{\sigma^2} \leq b) = 1-\alpha$$

  Estos valores son $a = \chi^2_{1-\alpha/2,n-1}$ y $b = \chi^2_{\alpha/2,n-1}$.

Intervalos de confianza para la varianza

CASO C: muestra $N(\mu, \sigma^2)$ con μ $\sigma > 0$ desconocida

Sea $X_1, X_2, \ldots, X_n$ una m.a. con distribución $N\left(\mu, \sigma^2\right)$, con μ y σ^2 desconocidos.

- 3ro) Despejando σ^2 resulta

$$P\left((n-1)\frac{S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq (n-1)\frac{S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = (1-\alpha)$$

Intervalos de confianza para la $N(\mu, \sigma^2)$

Intervalos de confianza $\%$ 100 $(1-\alpha)$

$$X_1, \ldots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$\bar{X} = \sum_{i=1}^{n} X_i/n, \quad S = \sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 /(n-1)}$$

Hipótesis	Parámetro	Intervalo de Confianza
σ^2 conocido	μ	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
σ^2 desconocido	μ	$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$
μ desconocido	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$

Ejemplo

Se efectuaron las siguientes observaciones de resistencia a la fractura de placas base de 18 % de acero maragizado al níquel:

$$69.571.972.673.173.373.575.575.775.876.176.2$$

$$76.277.077.978.179.679.779.980.182.283.783.7$$

Asuma que la muestra proviene de una distribución $N(\mu; \sigma^2)$.

a) Construir un intervalo de confianza del 99 % para la resistencia media a la fractura.

b) Construir un intervalo de confianza del 99 % para la desviación estándar poblacional de la resistencia a la fractura.

Intervalos de confianza asintóticos

CASO D: muestra aleatoria de tamaño grande σ conocido

Sea $X_1, X_2, \ldots, X_n$ una m.a. con $E(X_k) = \mu$ y $Var(X_k) = \sigma^2$ y n suficientemente grande para que valga el Teorema Central del Límite. Entonces se tiene que $\bar{X} = \sum_{i=1}^{n} X_i/n$ tiene distribución aproximadamente $N(\mu, \sigma^2/n)$.

- 1ro) Sabemos que si $n \geq 30$ y σ es conocido

$$h\left(X_1, X_2, \ldots, X_n; \theta\right) = \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

- 2do) Fijado un nivel de confianza ($1 - \alpha$),

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  es un intervalo de confianza aproximado para μ

Intervalos de confianza aproximados

CASO D: muestra aleatoria de tamaño grande σ desconocido

Sea $X_1, X_2, \ldots, X_n$ una m.a. con $E(X_k) = \mu$ y $Var(X_k) = \sigma^2$ ambos desconocidos

- 1ro) Sabemos que si $n \geq 40$ y n suficientemente grande para que valga el Teorema Central del Límite, por el teorema de Sluzky, como S^2 converge a σ^2, resulta

$$h\left(X_1, X_2, \ldots, X_n; \theta\right) = \sqrt{n}\frac{(\bar{X} - \mu)}{S_{n-1}} \sim N(0,1)$$

- 2do) Fijado un nivel de confianza ($1 - \alpha$),

$$\bar{X} \pm z_{\alpha/2}\frac{S_{n-1}}{\sqrt{n}}$$

es un intervalo de confianza aproximado para μ

Ejemplo

Se tiene una muestra de la duración de eco de radar de 110 relampagos en cierta región. El promedio muestral fue de 0.81 segundos y el desviación estándar muestral (s_{n1}) de 0.34 segundos. Calcular un IC de nivel aproximado 0.99 para la media de duración de eco μ

Ejemplo: Muestra Bernoulli

Sea $X_1, X_2, \ldots, X_n$ una m.a.con distribución $\mathrm{Bernoulli}(p)$ y tamaño de muestra suficientemente grande. Entonces por TCL resulta que

$$\frac{(\bar{X}-p)}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

donde $p = E(X)$ y $V(X) = p(1-p)$. Denotaremos con $\hat{p} = \bar{X}$, entonces

$$\frac{(\hat{p}-p)}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

Ejemplo: Muestra Bernoulli

Planteando la ecuación:

$$P\left(\left|\frac{(\hat{p}-p)}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq z_{\alpha/2}\right) \simeq (1-\alpha)$$

y trabajando con el evento elevado al cuadrado se puede obtener una ecuación cuadrática en p donde las raíces de esa ecuación son la cota inferior y superior de un intervalo de confianza aproximado ($1-\alpha$) para p, resultando:

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

Ejemplo: Muestra Bernoulli

Para n suficientemente grande:

- $\frac{z_{\alpha/2}^2}{2n}$ es insignificante en comparación con $\hat{p}$;
- $\frac{z_{\alpha/2}^2}{4n}$ es insignificante en comparación con $\frac{\hat{p}(1-\hat{p})}{n}$ y
- $\frac{z_{\alpha/2}^2}{n}$ es insignificante en comparación con 1.

Luego desechando esos términos insignificantes resulta este IC aleatorio de nivel aproximado $(1-\alpha)$ para $\boldsymbol{p}$ tradicional

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

siempre que $n\hat{p} \leq 10$ y $n(1-\hat{p}) \geq 10$

Ejemplo: Muestra Bernoulli

Si se quiere calcular el tamaño de muestra necesario para que un IC para p tenga nivel $(1-\alpha)$ y longitud L, entonces

$$n \geq (\frac{z_{\alpha/2}}{L})^2$$

independendientemente del valor de la proporción observada $\hat{p}$

Ejemplo: Muestra Bernoulli

En un articulo sobre estimación de fuentes de defectos visuales, se reporta que se estudiaron con un sensor de inspección 356 matrices de silicio de las cuales 201 pasaron la prueba.

a) Construir un intervalo de confianza de nivel aproximado 0. 98 para la proporción poblacional de matrices que pasan la inspección.

b) ¿Que tamaño de muestra seria necesario para que la longitud un intervalo de confianza de nivel aproximado 0.98 para la proporción poblacional de matrices que pasan la inspección sea a lo sumo 0.05, independientemente del valor $\hat{p}$?