Tema 1. Inferencia estadística para una población

Contenidos

- Inferencia estadística
- Estimadores puntuales
  - Estimación de la media y la varianza de una población
- Estimación de la media de la población mediante intervalos de confianza
  - Intervalos de confianza para la media de una población normal con varianza conocida
  - Intervalos de confianza para la media en muestras grandes
    - Intervalos de confianza para la proporción en una población
  - Intervalos de confianza para la media de una población normal con varianza desconocida
- Estimación de la varianza de la población mediante intervalos de confianza
  - Intervalos de confianza para la varianza de una población normal

Tema 1. Inferencia estadística para una población

Objetivos de aprendizaje

Al final de este tema debieras ser capaz de:

- Estimar parámetros de la población desconocidos a partir de datos muestrales
- Construir intervalos de confianza para los parámetros de la población desconocidos a partir de datos muestrales:
  - En el caso de una distribución normal: intervalos de confianza para la media y la varianza de la población
  - En muestras grandes: intervalos de confianza para la media de la población y la proporción
- Interpretar el significado de un intervalo de confianza
- Entender el efecto del tamaño muestral, el nivel de confianza, etc sobre la longitud del intervalo de confianza
- Calcular un tamaño muestral necesario para controlar la longitud de un intervalo de confianza

Tema 1. Inferencia estadística para una población

Referencias

- Newbold, P. "Estadística para Administración y Economía"
  - Capítulos 7 y 8 (8.1-8.6)
- Ross, S. "Introducción a la Estadística"
  - Capítulo 8

Inferencia Estadística: palabras clave (i)

- Población: el conjunto de toda la información numérica relativa a una cantidad de interés.
  - Identificaremos el concepto de población con el de una variable aleatoria X.
  - La ley o distribución de la población es la distribución de X, F_X.
- Muestra: un subconjunto observado (por ejemplo, de tamaño n) de valores de la población.
  - Representada como una colección de n variables aleatorias $X_1, X_2, \ldots, X_n$, típicamente $\boxed{\text{iid (independientes e idénticamente distribuidas)}}$.
- Parámetro: una constante que caracteriza a X o F_X.

Inferencia Estadística: palabras clave (ii)

- Inferencia estadística: el proceso mediante el que se llega a conclusiones sobre una población a partir de las medidas o las observaciones realizadas sobre una muestra de individuos de la población.
- Estadístico: una variable aleatoria definida como una función de una muestra aleatoria, $Y = f(X_1, X_2, \ldots, X_n)$
- Estimador de un parámetro: una variable aleatoria, por ejemplo T, función de una muestra aleatoria, $T = T(X_1, X_2, \ldots, X_n)$, que se emplea para aproximar (estimar) el valor de un parámetro de la población desconocido.
- Estimación: una realización concreta del estimador, por ejemplo T, correspondiente a una muestra observada, $x_1, x_2, \ldots, x_n$, y que proporciona una aproximación al valor del parámetro de interés.

Inferencia estadística: ejemplo

Queremos conocer $\mu_X = \mathrm{E}[X]$		Tenemos n copias de X		Tenemos n valores observados de $X_1, X_2, \ldots, X_n$
$X \sim F$	$\Rightarrow$	$X_1, X_2, \ldots, X_n \sim F$ Muestra	$\Rightarrow$	$x_1, x_2, \ldots, x_n$ Muestra observada
$\Downarrow$		$\Downarrow$		$\Downarrow$
$\mu_X = \mathrm{E}[X]$ Valor esperado de X	$\Leftarrow$	Estimador de μ_X (variable aleatoria) $\bar{X}$ Media muestral	$\Leftarrow$	Estimación de μ_X (un número) $\bar{x}$ Media muestral

Estimadores puntuales: introducción

- Un estimador puntual de un parámetro de una población es una función, por ejemplo T, de la información muestral $\underline{X}_n = (X_1, \ldots, X_n)$ que toma un valor numérico.
- Ejemplos de parámetros de poblaciones, estimadores y estimaciones:

Parámetro población	$T(\underline{X}_n)$	Estimador: notación	Estimación: notación
Media pobl. μ_X	media muestral $\frac{X_1+\ldots+X_n}{n}$	$\bar{X} = \hat{\mu}_X$	$\bar{x}$
Prop. pobl. p_X	prop. muestral	$\hat{p}_X$	$\hat{p}_x$
Var. pobl. σ_X^2	var. muestral $\frac{\sum_i X_i^2 - n(\bar{X})^2}{n}$	$\hat{\sigma}_X^2$	$\hat{\sigma}_x^2$
Var. pobl. σ_X^2	quasi var. muestral $\frac{\sum_i X_i^2 - n(\bar{X})^2}{n-1} = \frac{n}{n-1}\hat{\sigma}_X^2$	s_X^2	s_x^2
...	...	...	...
En general, θ_X	...	$\hat{\theta}_X$	$\hat{\theta}_x$

Estimadores puntuales: propiedades (i)

¿Qué características querríamos que tuviese un estimador?

- Ausencia de sesgo. Esta propiedad se da cuando un estimador tiene sesgo igual a cero. ¿Qué es el sesgo? El sesgo es la diferencia entre el valor esperado del estimador y el valor del parámetro de interés.

$$\text{Sesgo}[\hat{\theta}_X] = \mathrm{E}[\hat{\theta}_X] - \theta_X$$

Población parámetro	Estimador $T(\underline{X}_n)$	Sesgo	Insesgado?	Estimador insesgado de mínima varianza?
Media pobl. μ_X	$\bar{X}$	$\mathrm{E}[\bar{X}] - \mu_X = 0$	Sí	Sí, si X normal
Prop. pobl. p_X	$\hat{p}_X$	$\mathrm{E}[\hat{p}_X] - p_X = 0$	Sí	Sí
Var. pobl. σ_X^2	$\hat{\sigma}_X^2$	$\mathrm{E}[\hat{\sigma}_X^2] - \sigma_X^2 \neq 0$	No	No
Var. pobl. σ_X^2	s_X^2	$\mathrm{E}[s_X^2] - \sigma_X^2 = 0$	Sí	Sí, si X normal
En general, θ_X	$\hat{\theta}_X$	$\mathrm{E}[\hat{\theta}_X] - \theta_X$	A menudo	Rara vez

Estimadores puntuales: propiedades (ii)

- Eficiencia. Se mide por la varianza del estimador. Un estimador con menos varianza es más eficiente.
- La eficiencia relativa de dos estimadores insesgados $\hat{\theta}_{X,1}$ y $\hat{\theta}_{X,2}$ para un parámetro θ_X se define como

$$\text{Eficiencia relativa}(\hat{\theta}_{X,1}, \hat{\theta}_{X,2}) = \frac{\text{Var}[\hat{\theta}_{X,1}]}{\text{Var}[\hat{\theta}_{X,2}]}$$

Nota:
  - En algunos casos se emplea la definición inversa.
  - En todo caso, un estimador con menor varianza es más eficiente.

Estimadores puntuales: propiedades (iii)

- Un criterio más general para seleccionar estimadores (incluyendo estimadores insesgados y sesgados) es el error cuadrático medio, definido como

$$\text{ECM}[\hat{\theta}_X] = \text{E}[(\hat{\theta}_X - \theta_X)^2] = \text{Var}[\hat{\theta}_X] + (\text{Sesgo}[\hat{\theta}_X])^2$$

Nota:
  - El error cuadrático medio de un estimador insesgado es igual a su varianza.
  - Un estimador con menor ECM es mejor.
  - El estimador insesgado de mínima varianza tiene la menor varianza/ECM entre todos los estimadores.
- Como encontrar una buena definición para un estimador T?
  - En algunos casos se conoce un estimador óptimo: estimador insesgado de mínima varianza
  - Si no es así, existen distintos métodos de construcción de estimadores que proporcionan resultados razonables, por ejemplo:
    - Estimación máximo verosímil
    - Método de momentos

Estimación puntual: ejemplo

Ejemplo: 7.1 (Newbold) Las ratios precio-beneficio para una muestra aleatoria de diez acciones negociadas en la bolsa de NY en un día concreto fueron

$$10 \quad 16 \quad 5 \quad 10 \quad 12 \quad 8 \quad 4 \quad 6 \quad 5 \quad 4$$

Emplee un procedimiento de estimación insesgado para obtener estimaciones puntuales para los siguientes parámetros de la población: media, varianza, proporción de valores que exceden 8.5.

$$\begin{aligned}
\bar{x} &= \frac{80}{10} = 8 \\
s_x^2 &= \frac{782 - 10(8)^2}{10-1} = 15{,}78 \\
\hat{p}_x &= \frac{1+1+0+1+1+0+0+0+0+0}{10} \\
&= 0{,}4
\end{aligned}$$

Estimación puntual: ejemplo

Ejemplo: Sea $\hat{\mu}_X = \frac{2}{n(n+1)}(X_1 + 2X_2 + \ldots + nX_n)$ un estimador de la media de la población basado en una MAS $\underline{X}_n$. Compare este estimador con la media muestral, $\bar{X}$.

Sabemos que $\bar{X}$ es un estimador insesgado de μ_X, con varianza $\frac{\sigma_X^2}{n}$.

$\hat{\mu}_X$ también es insesgado:

$$\begin{aligned}
E[\hat{\mu}_X] &= E\left[\frac{2}{n(n+1)}(X_1 + 2X_2 + \ldots + nX_n)\right] \\
&= \frac{2}{n(n+1)}(E[X_1] + 2E[X_2] + \ldots + nE[X_n]) \\
&=_{\text{id}} \frac{2}{n(n+1)}(\mu_X + 2\mu_X + \ldots + n\mu_X) \\
&= \frac{2\mu_X}{n(n+1)} \overbrace{(1+2+\ldots+n)}^{n(n+1)/2} = \mu_X
\end{aligned}$$

$\Rightarrow$ Sesgo$[\hat{\mu}_X] = 0$

Y su varianza/ECM es:

$$\begin{aligned}
V[\hat{\mu}_X] &= V\left[\frac{2}{n(n+1)}(X_1 + 2X_2 + \ldots + nX_n)\right] \\
&=_{\text{indep.}} \left(\frac{2}{n(n+1)}\right)^2 (V[X_1] + 2^2 V[X_2] + \ldots + n^2 V[X_n]) \\
&=_{\text{id}} \frac{4}{n^2(n+1)^2}\sigma_X^2 \overbrace{(1^2 + 2^2 + \ldots + n^2)}^{n(n+1)(2n+1)/6} \\
&= \frac{2(2n+1)}{3n(n+1)}\sigma_X^2 \\
ECM[\hat{\mu}_X] &= V[\hat{\mu}_X] + 0^2 = \frac{2(2n+1)}{3n(n+1)}\sigma_X^2
\end{aligned}$$

$$\text{Eficiencia relativa}(\bar{X}, \hat{\mu}_X) = \frac{\sigma_X^2/n}{\frac{2(2n+1)}{3n(n+1)}\sigma_X^2} = \frac{3(n+1)}{2(2n+1)}$$

Puede verse que para $n \geq 2$ este cociente es menor que 1, y por tanto $\bar{X}$ es un estimador más eficiente para μ_X.

De estimaciones puntuales a estimación por intervalos de confianza

- Hasta ahora hemos considerado la estimación puntual de un parámetro desconocido de una población que, partiendo de una MAS de n observaciones de X, proporciona una aproximación razonable para ese parámetro desconocido.
- Una estimación puntual no tiene en cuenta la variabilidad del proceso de estimación, debida entre otras causas a:
  - El tamaño muestral - una muestra mayor debiera proporcionar una información más precisa sobre el parámetro de la población.
  - Variabilidad en la población - una muestra de una población con menos varianza debiera proporcionar estimaciones más precisas
  - Que se conozcan otros parámetros de la población.
  - etc

Estas limitaciones pueden tratarse mediante el uso de estimaciones por intervalos de confianza, esto es, un método que proporciona un intervalo de valores al que es probable que pertenezca el valor del parámetro.

Estimadores por intervalos de confianza e intervalos de confianza

Sea $\underline{X}_n = (X_1, X_2, \ldots, X_n)$ una MAS de una población X con función de distribución F_X que depende de un parámetro desconocido θ.
Un estimador por intervalos de confianza de θ con un nivel de confianza $(1-\alpha) = 100(1-\alpha)\,\%$ es un intervalo $(T_1(\underline{X}_n), T_2(\underline{X}_n))$ que satisface

$$P\left(\theta \in (T_1(\underline{X}_n), T_2(\underline{X}_n)\right) = 1-\alpha$$

Interpretación: tenemos una probabilidad $(1-\alpha)$ de que el parámetro desconocido de la población pertenecerá a $(T_1(\underline{X}_n), T_2(\underline{X}_n))$.
Un intervalo de confianza para θ con un nivel de confianza $1-\alpha$ es el valor observado del estimador por intervalos de confianza,

$$(T_1(\underline{x}_n), T_2(\underline{x}_n))$$

Interpretación: podemos tener una confianza de $(1-\alpha)$ de que el valor del parámetro desconocido de la población estará en $(T_1(\underline{x}_n), T_2(\underline{x}_n))$.
Niveles de confianza habituales

α	0.01	0.05	0.10
$100(1-\alpha)\,\%$	99 %	95 %	90 %

Obteniendo un estimador por intervalos de confianza: procedimiento

1. Se busca una cantidad (aleatoria) relacionada con el parámetro desconocido θ y con la muestra $\underline{X}_n$, $C(\underline{X}_n, \theta)$, cuya distribución sea conocida y no dependa del valor del parámetro - esta cantidad se conoce como la cantidad pivotal o el pivote para θ
2. Se pueden usar los cuantiles $1-\alpha/2$ y $\alpha/2$ de esa distribución, y la definición del estimador por intervalos de confianza, para plantear la ecuación

$$P(\overbrace{1-\alpha/2 \text{ cuantil} < C(\underline{X}_n, \theta) < \alpha/2 \text{ cuantil}}^{\text{doble desigualdad}}) = 1-\alpha.$$

3. Para obtener los extremos del estimador por intervalo de confianza, $T_1(\underline{X}_n)$ y $T_2(\underline{X}_n)$, se resuelve la doble desigualdad en θ.
4. El intervalo de confianza al $100(1-\alpha)\,\%$ para θ es $(T_1(\underline{x}_n), T_2(\underline{x}_n))$.

Intervalo de confianza para la media de la población, población normal con varianza conocida

1. Sea $\underline{X}_n$ una MAS de tamaño n obtenida de X. Bajo los supuestos:
   - X sigue una distribución normal con parámetros μ_X y σ_X^2
   - σ_X^2 es conocida (muy poco realista)
2. La cantidad pivotal para μ_X es:

$$\boxed{\frac{\bar{X}-\mu_X}{\sigma/\sqrt{n}} \sim N(0,1)}$$

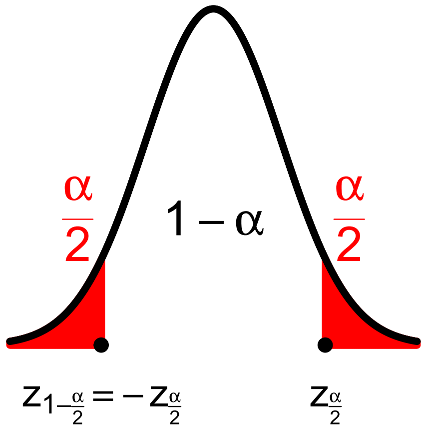
Nota: la desviación típica de $\bar{X}$, $\sigma_X/\sqrt{n}$, (o de cualquier otro estadístico) se conoce como su error estándar

Intervalo de confianza para la media de la población, población normal con varianza conocida

3. Si $z_{1-\alpha/2}$ y $z_{\alpha/2}$ son los cuantiles superiores $(1-\alpha/2)$ y $(\alpha/2)$ de la distribución $N(0,1)$, tenemos

$$\boxed{P(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha}$$

Densidad normal estándar



Si $Z \sim N(0,1)$ entonces
$\mathrm{E}[Z]=0$, $\mathrm{V}[Z]=1$

4. Se tiene que $P(\overbrace{z_{1-\alpha/2}}^{-z_{\alpha/2}} < \overbrace{\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}}}^{Z} < z_{\alpha/2}) = 1-\alpha$

Intervalo de confianza para la media de la población, población normal con varianza conocida

5. Resolvemos la doble desigualdad para μ_X:

$$\begin{array}{rcl} -z_{\alpha/2} & < \frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} < & z_{\alpha/2} \\ -z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} & < \bar{X}-\mu_X < & z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \\ -z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}-\bar{X} & < -\mu_X < & -\bar{X}+z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \\ z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}+\bar{X} & > \mu_X > & \bar{X}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \end{array}$$

para obtener el estimador por intervalos de confianza

$$(\overbrace{\bar{X}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}}^{T_1(\underline{X}_n)}, \overbrace{\bar{X}+z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}}^{T_2(\underline{X}_n)})$$

6. El intervalo de confianza es:

$$\mathrm{IC}_{1-\alpha}(\mu_X) = \left(\bar{x}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \bar{x}-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) = \left(\bar{x}\mp z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$$

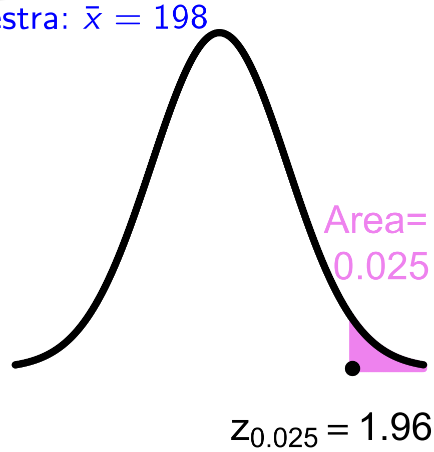
Ejemplo: cálculo de un intervalo de confianza para μ_X

Ejemplo: 8.2 (Newbold) Un proceso industrial produce bolsas de azucar refinado. Los pesos de las bolsas siguen una distribución normal con desviación típica igual a 12 g. Una muestra aleatoria de venticinco bolsas tiene un peso promedio de 198 g. Encuentre un intervalo de confianza al 95 % para el peso promedio en la población de las bolsas de azucar fabricadas con este proceso.

Población:
$X =$ "peso de una bolsa (en g)"
$X \sim N(\mu_X, \sigma_X^2 = 12^2)$

MAS: $n = 25$

Muestra: $\bar{x} = 198$



$$\boxed{\text{Objetivo: } IC_{0,95}(\mu_X) = \left(\bar{x} \mp z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)}$$

$$\begin{aligned}
\sigma_X &= 12 \\
n = 25 & \quad \bar{x} = 198 \\
1-\alpha = 0{,}95 &\Rightarrow \alpha/2 = 0{,}025 \\
z_{\alpha/2} &= z_{0,025} = 1{,}96 \\
IC_{0,95}(\mu_X) &= \left(198 \mp 1{,}96\frac{12}{\sqrt{25}}\right) \\
&= (198 \mp 4{,}7) \\
&= (193{,}3, 202{,}7)
\end{aligned}$$

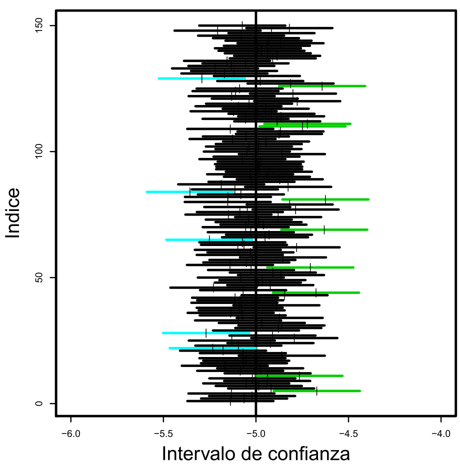
Interpretación: Podemos tener una confianza del 95 % de que μ_X está en $(193{,}3, 202{,}7)$

Interpretación frecuentista del IC: nivel de confianza

En este ejemplo simulado se han generado 150 muestras de tamaño $n = 50$, de una distribución $\boxed{X \sim N(\mu_X = -5, \sigma_X^2 = 1^2)}$ y se construyeron 150 $IC_{1-\alpha}(\mu_X)$ con $\alpha = 0{,}1$ y otros 150 intervalos con $\alpha = 0{,}01$.

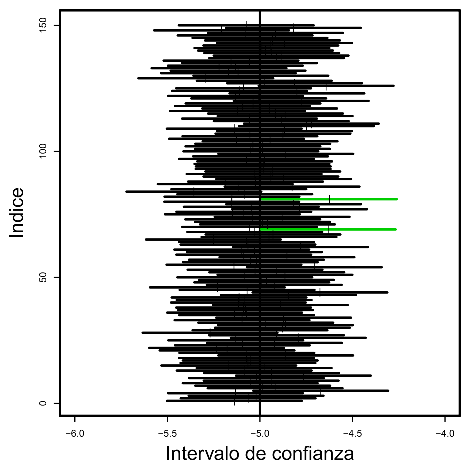
μ_X está en aprox. $150(0{,}9) = 135$ interv.
(pero no en $150(0{,}1) = 15$)

$\boxed{(1-\alpha) = 0{,}9,\ n = 50}$



μ_X está en aprox. $150(0{,}99) = 148{,}5$ interv.
(pero no en $150(0{,}01) = 1{,}5$)

$\boxed{(1-\alpha) = 0{,}99,\ n = 50}$



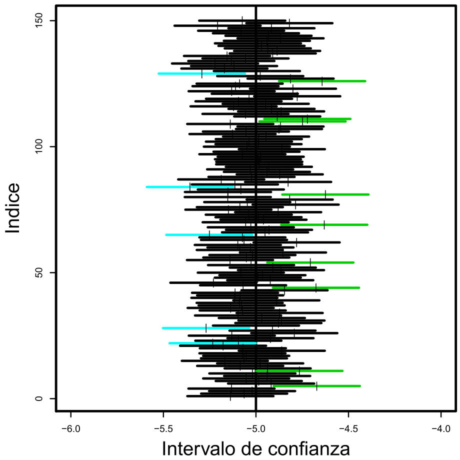
La longitud del intervalo, $\boxed{L = \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} - \left(\bar{x} - \frac{z_{\alpha/2}\sigma_X}{\sqrt{n}}\right) = 2\frac{z_{\alpha/2}\sigma_X}{\sqrt{n}}}$, aumenta al aumentar el nivel de confianza (a igualdad de otros valores). ¿Por qué?

Interpretación frecuentista del IC: tamaño muestral

Construimos ahora 150 muestras de tamaño $n = 50$ y otras 150 de tamaño $n = 200$, de una distribución $\boxed{X \sim N(\mu_X = -5, \sigma_X^2 = 1^2)}$.

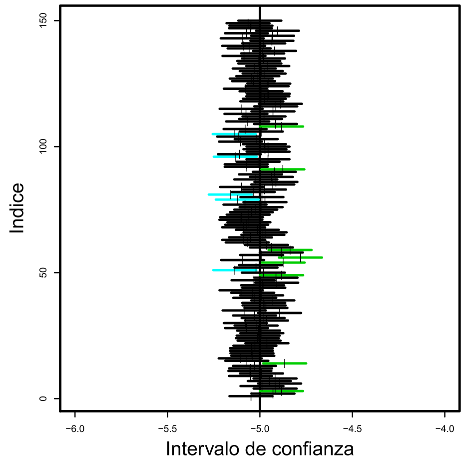
μ_X en aprox. $150(0{,}9) = 135$ interv. (pero no en $150(0{,}1) = 15$)

$\boxed{(1-\alpha) = 0{,}9,\ n = 50}$



μ_X en aprox. $150(0{,}9) = 135$ interv. (pero no en $150(0{,}1) = 15$)

$\boxed{(1-\alpha) = 0{,}9,\ n = 200}$



La longitud de los intervalos decrece cuando el tamaño muestral aumenta (suponiendo que los demás valores no cambien). ¿Por qué?

Pregunta: ¿Cuál es el efecto del valor de σ sobre la longitud?

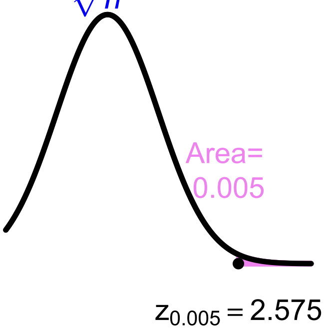
Ejemplo: estimación del tamaño muestral

Ejemplo: 8.14 (Newbold) La longitud de las barras de acero fabricadas en un proceso industrial siguen una distribución normal con desviación típica 1.8 mm. El encargado del proceso desea obtener un intervalo de confianza al 99 % para dicha longitud, con un tamaño menor o igual a 0.5 mm a cada lado de la media muestral. ¿Qué tamaño muestral sería necesario para tener esta propiedad?

Población:
$X =$ "longitud de la barra (en mm)"
$X \sim N(\mu_X, \sigma_X^2 = 1{,}8^2)$

MAS: $n = ?$

$$\mathrm{IC}_{0{,}99}(\mu_X):\ \overbrace{2\frac{z_{\alpha/2}\sigma}{\sqrt{n}}}^{\text{longitud}} \leq 2(0{,}5) = 1$$



Objetivo: n tal que longitud IC ≤ 1

$$\begin{aligned} 2\frac{z_{\alpha/2}\sigma}{\sqrt{n}} &\leq 1 \\ \sqrt{n} &\leq \frac{z_{\alpha/2}\sigma}{2(1)} \\ n &\leq \frac{z_{\alpha/2}^2\sigma^2}{2^2} \\ n &\leq \frac{2{,}575^2(1{,}8^2)}{2^2} = 85{,}93 \end{aligned}$$

Para satisfacer la petición del encargado se necesitaría una muestra de tamaño al menos igual a 86 observaciones.

Intervalo de confianza para la media de la población en muestras grandes

1. Sea $\underline{X}_n$ una MAS de tamaño n de X. Bajo las hipótesis:
   - X sigue una distribución (no necesariamente normal) con parámetros μ_X y σ_X^2
   - el tamaño muestral n es grande ($n \geq 30$)
2. La cantidad pivotal para μ_X basada en el Teorema Central del Límite es

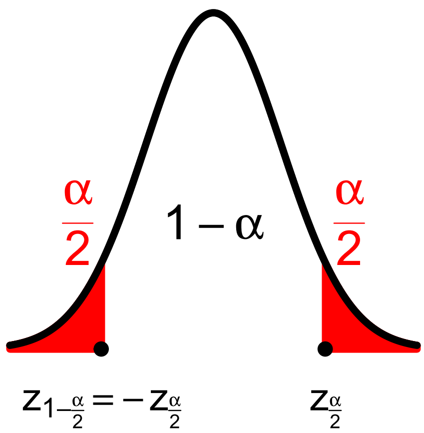
$$\boxed{\frac{\bar{X} - \mu_X}{\hat{\sigma}_X/\sqrt{n}} \sim^{\text{approx.}} N(0,1)}$$

Intervalo de confianza para la media de la población en muestras grandes

3. Por tanto, si $z_{1-\alpha/2}$ y $z_{\alpha/2}$ son los cuantiles superiores $(1-\alpha/2)$ y $(\alpha/2)$ de $N(0,1)$, tenemos

$$\boxed{P(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha}$$

Densidad normal estándar



4. Imponemos la condición $P(\overbrace{z_{1-\alpha/2}}^{-z_{\alpha/2}} < \overbrace{\frac{\bar{X} - \mu_X}{\hat{\sigma}_X/\sqrt{n}}}^{Z} < z_{\alpha/2}) = 1 - \alpha$

Intervalo de confianza para la media de la población en muestras grandes

5. Resolvemos la doble desigualdad para μ_X:

$$-z_{\alpha/2} < \frac{\bar{X} - \mu_X}{\hat{\sigma}_X/\sqrt{n}} < z_{\alpha/2}$$

para obtener el estimador por intervalos de confianza

$$(\overbrace{\bar{X} - z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}}^{T_1(\underline{X}_n)}, \overbrace{\bar{X} + z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}}^{T_2(\underline{X}_n)})$$

6. El intervalo de confianza es:

$$\mathrm{IC}_{1-\alpha}(\mu_X) = (\bar{x} - z_{\alpha/2}\frac{\hat{\sigma}_x}{\sqrt{n}}, \bar{x} - z_{\alpha/2}\frac{\hat{\sigma}_x}{\sqrt{n}})$$

Intervalo de confianza para la proporción en la población en muestras grandes

Aplicación de ICs para la media en muestras grandes

Sea $\underline{X}_n$, $n \geq 30$, una MAS de una distr. Bernoulli con parámetro p_X ($\mu_X = \mathrm{E}[X] = p_X$ y $\sigma_X = \sqrt{p_X(1-p_X)}$). La proporción muestral $\hat{p}_X$ es un caso especial de media muestral con observaciones cero-uno, $\hat{p}_X = \bar{X}$.

Por tanto, del TCL,

$$\frac{\hat{p}_X - p_X}{\underbrace{\sqrt{p(1-p)/n}}_{\sigma_X/\sqrt{n}}} \sim \text{approx. } N(0,1)$$

Este resultado sigue siendo válido si la desviación típica de la población se estima (no se conoce)

$$\frac{\hat{p}_X - p_X}{\underbrace{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}_{\hat{\sigma}_X/\sqrt{n}}} \sim \text{approx. } N(0,1)$$

En muestras grandes, el intervalo de confianza para p_X es:

$$\mathrm{IC}_{1-\alpha}(p_X) = \left(\hat{p}_x - z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}}, \hat{p}_x + z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}}\right)$$

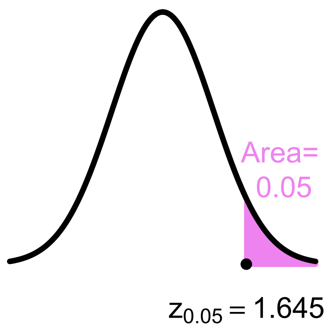
Ejemplo: cálculo de un intervalo de confianza para p_X

Ejemplo: 8.6 (Newbold) A una muestra aleatoria de 344 ejecutivos de compras se les realizó la pregunta "¿Cuál es la política de su empresa en relación con los regalos que su personal de compras pueda recibir de sus proveedores?" 83 de estos ejecutivos respondieron que cada empleado podía tomar su propia decisión. Calcule un intervalo de confianza al 90 % para la proporción en la población de los ejecutivos que dan libertad sobre estos regalos a sus empleados.

Población:
$X = 1$ si un ejecutivo permite tomar decisiones a su personal y 0 en otro caso
$X \sim \text{Bernoulli}(p_X)$

MAS: $n = 344$ $\boxed{\text{grande}}$

Muestra: $\hat{p}_x = \frac{83}{344} = 0{,}241$



Objetivo: $\boxed{IC_{0,9}(p_X) = \left(\hat{p}_x \mp z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}}\right)}$

$$\hat{p}_x = 0{,}241 \qquad n = 344$$

$$1-\alpha = 0{,}9 \quad \Rightarrow \quad \alpha/2 = 0{,}05$$

$$z_{\alpha/2} = z_{0,05} = 1{,}645$$

$$\begin{aligned} IC_{0,9}(p_X) &= \left(0{,}241 \mp 1{,}645\sqrt{\frac{0{,}241(1-0{,}241)}{344}}\right) \\ &= (0{,}241 \mp 0{,}038) \\ &= (0{,}203,\ 0{,}279) \end{aligned}$$

Interpretación: Podemos tener una confianza del 90 % de que la proporción de ejecutivos que permiten tomar sus propias decisiones a sus empleados, p_X, está en $(0{,}203, 0{,}279)$

Intervalo de confianza para la media de la población: población normal con varianza desconocida

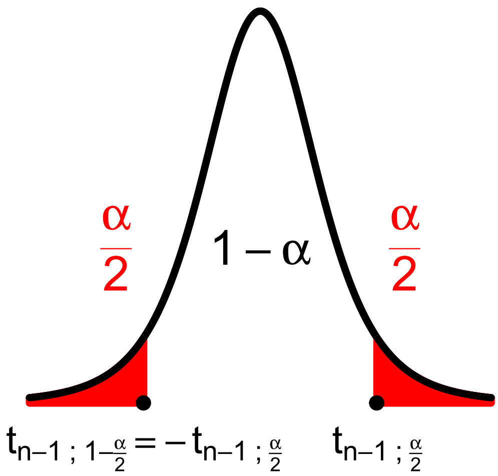
1. Sea $\underline{X}_n$ una MAS de tamaño n de X. Bajo las hipótesis:
   - X sigue una distribución normal con parámetros μ_X y σ_X^2
   - σ_X^2 es desconocida (muy realista)
2. La cantidad pivotal para μ_X es

$$\boxed{\frac{\bar{X} - \mu_X}{s_X/\sqrt{n}} \sim t_{n-1}}$$

Intervalo de confianza para la media de la población: población normal con varianza desconocida

3. Si $t_{n-1;1-\alpha/2}$ y $t_{n-1;\alpha/2}$ son los cuantiles superiores $(1-\alpha/2)$ y $(\alpha/2)$ de una distribución t de Student con $n-1$ grados de libertad (gl), tenemos

$$\boxed{P(t_{n-1;1-\alpha/2} < \overbrace{T}^{\sim\, t_{n-1}} < t_{n-1;\alpha/2}) = 1-\alpha}$$



Densidad t de Student

Recuerda: si $T \sim t_n$, $\mathrm{E}[T] = 0$, $\mathrm{V}[T] = \frac{n}{n-2}$

4. Imponemos la condición

$$P(\overbrace{t_{n-1;1-\alpha/2}}^{-t_{n-1;\alpha/2}} < \overbrace{\frac{\bar{X}-\mu_X}{s_X/\sqrt{n}}}^{T\,\sim\, t_{n-1}} < t_{n-1;\alpha/2}) = 1-\alpha$$

Intervalo de confianza para la media de la población: población normal con varianza desconocida

5. Resolvemos la doble desigualdad para μ_X:

$$-t_{n-1;\alpha/2} \quad < \frac{\bar{X}-\mu_X}{s_X/\sqrt{n}} < \quad z_{\alpha/2}$$

y obtenemos el estimador por intervalos de confianza

$$(\overbrace{\bar{X} - t_{n-1;\alpha/2}\frac{s_X}{\sqrt{n}}}^{T_1(\underline{X}_n)}, \overbrace{\bar{X} + t_{n-1;\alpha/2}\frac{s_X}{\sqrt{n}}}^{T_2(\underline{X}_n)})$$

6. El intervalo de confianza es:

$$\mathrm{IC}_{1-\alpha}(\mu) = (\bar{x} - z_{\alpha/2}\frac{s_x}{\sqrt{n}}, \bar{x} - z_{\alpha/2}\frac{s_x}{\sqrt{n}})$$

Ejemplo: calcular un intervalo de confianza para μ_X

Ejemplo: 8.4 (Newbold) Se ha medido el consumo de combustible en una muestra aleatoria de seis coches del mismo modelo, obteniendo en mpg: 18.6, 18.4, 19.2, 20.8, 19.4, 20.5. Calcule un intervalo de confianza al 90 % para el consumo medio, suponiendo que la población sigue una distribución normal.

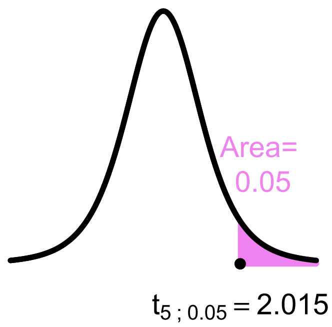
Población: $X =$ "mpg de un coche de este modelo" $X \sim N(\mu_X, \sigma_X^2)$

$\boxed{\sigma_X^2 \text{ desconocida}}$

MAS: $n = 6$ $\boxed{\text{pequeña}}$

Muestra: $\bar{x} = \frac{116{,}9}{6} = 19{,}4833$

$$s_x^2 = \frac{2282{,}41 - 6(19{,}4833)^2}{6-1} = 0{,}96$$



$$\boxed{\text{Objetivo: } IC_{0,9}(\mu_X) = \left(\bar{x} \mp t_{n-1;\alpha/2}\frac{s_x}{\sqrt{n}}\right)}$$

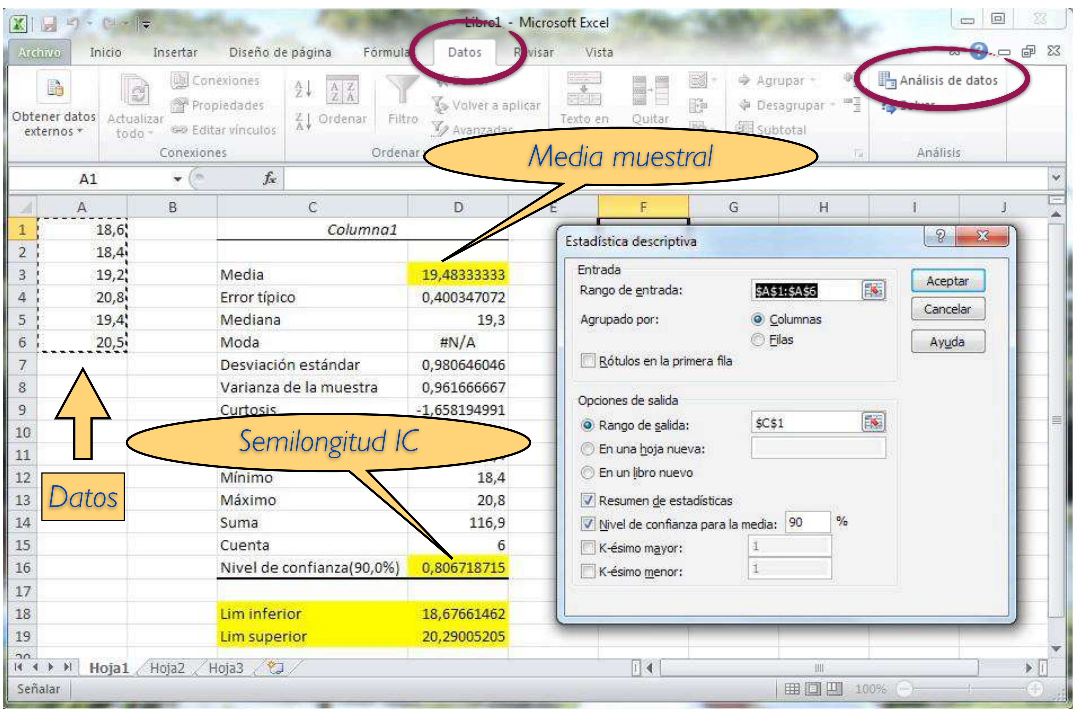
$$\begin{aligned}
s_x &= \sqrt{0{,}96} = 0{,}98 \\
n = 6 & \quad \bar{x} = 19{,}48 \\
1-\alpha = 0{,}9 &\Rightarrow \alpha/2 = 0{,}05 \\
t_{n-1;\alpha/2} &= t_{5;0{,}05} = 2{,}015 \\
IC_{0,9}(\mu_X) &= \left(19{,}48 \mp 2{,}105\frac{0{,}98}{\sqrt{6}}\right) \\
&= (19{,}48 \mp 0{,}81) \\
&= (18{,}67, 20{,}29)
\end{aligned}$$

Interpretación: Tenemos una confianza del 90 % de que el consumo promedio de este modelo, μ_X, estará entre 18.67 y 20.29 mpg

Ejemplo: calcular un intervalo de confianza para μ_X

Ejemplo: 8.4 (cont.) en Excel: Ir al menú: Datos, submenú: Análisis de Datos, escoger la función: Estadística Descriptiva.

Columna A datos en amarillo (media muestral, semilongitud $t_{n-1;\alpha/2}\frac{s_x}{\sqrt{n}}$, extremo inferior (celda: D3-D16), extremo superior (celda: D3+D16)).



Distribuciones t de Student y χ^2 (chi-cuadrado)

- Sabemos que $T \sim t_n$ si $T = \frac{Z}{\sqrt{\chi^2_n/n}}$, donde $Z \sim N(0,1)$ y χ^2_n sigue una distribución chi-cuadrado con gl $= n$, y ambas son independientes.
- También, χ^2_n es la distribución de la suma de los cuadrados de n variables aleatorias $N(0,1)$ independientes.
- Por ejemplo, la cuasi varianza muestral reescalada sigue una distribución chi cuadrado con $n-1$ grados de libertad.

$$\frac{(n-1)s_X^2}{\sigma_X^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma_X^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma_X}\right)^2 \sim \chi^2_{n-1}$$

¿Por qué $n-1$ y no n?

Si conociesemos el valor de μ_X, el número de grados de libertad sería n, porque tendríamos n variables aleatorias *iid* $\frac{X_i - \mu_X}{\sigma_X}$

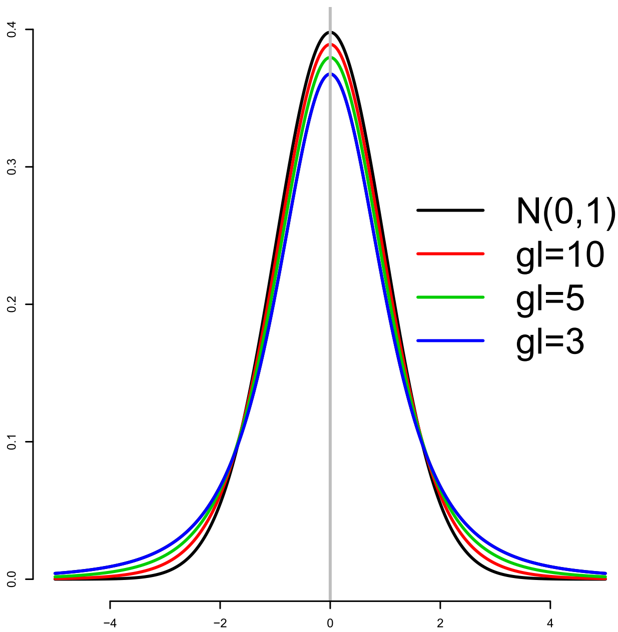
Si tenemos que estimar μ_X mediante $\bar{X}$, los gl son $n-1$, porque solo tenemos $n-1$ variables aleatorias *iid* $\frac{X_i - \bar{X}}{\sigma_X}$ (si se conocen los valores de $n-1$ de ellas, se puede deducir fácilmente el valor de la restante)

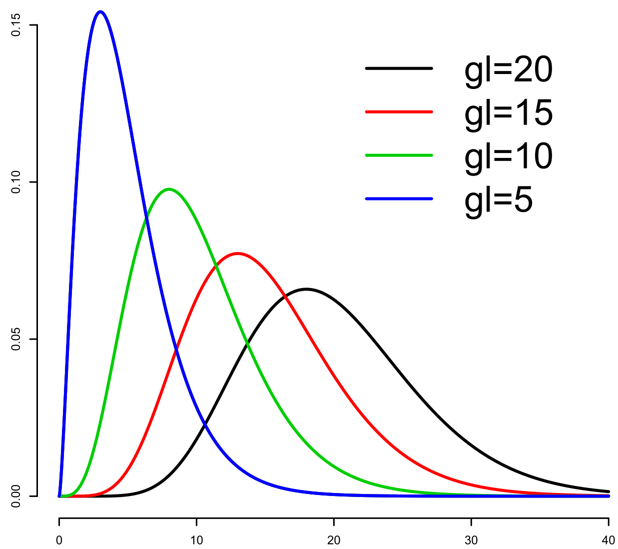
Decimos que empleamos un grado de libertad en el cálculo de μ_X

Distribuciones t de Student y χ^2 (chi-cuadrado)

Densidades de t y $N(0,1)$

Densidades de χ^2





Intervalo de confianza para la varianza de la población, población normal

1. Sea $\underline{X}_n$ una MAS de tamaño n de X. Bajo las hipótesis:
   - X sigue una distribución normal con varianza σ_X^2
2. La cantidad pivotal para σ_X^2 es

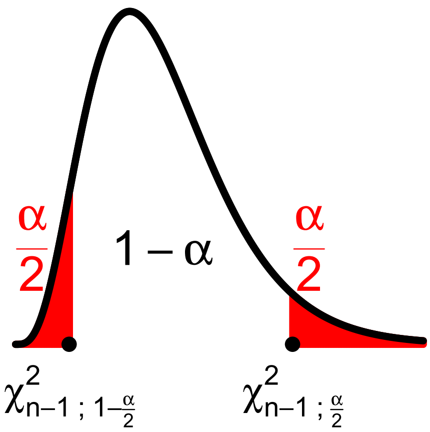
$$\boxed{\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}}$$

Intervalo de confianza para la varianza de la población, población normal

3. Por tanto, si $\chi^2_{n-1;1-\alpha/2}$ y $\chi^2_{n-1;1-\alpha/2}$ son los cuantiles superiores $(1-\alpha/2)$ y $(\alpha/2)$ de una distribución chi-cuadrado con $n-1$ grados de libertad, tenemos

$$\boxed{P(\chi^2_{n-1;1-\alpha/2} < \chi^2_{n-1} < \chi^2_{n-1;\alpha/2}) = 1-\alpha}$$

Densidad chi-cuadrado



Recuerda: $E[\chi^2_n] = n$, $V[\chi^2_n] = 2n$

4. Imponemos la condición $P(\chi^2_{n-1;1-\alpha/2} < \overbrace{\frac{(n-1)s_X^2}{\sigma^2}}^{\chi^2_{n-1}} < \chi^2_{n-1;\alpha/2}) = 1-\alpha$

Intervalo de confianza para la varianza de la población, población normal

5. Resolvemos la doble desigualdad para σ_X^2:

$$\begin{array}{ccccc}
\chi^2_{n-1;1-\alpha/2} & < & \frac{(n-1)s_X^2}{\sigma_X^2} & < & \chi^2_{n-1;\alpha/2} \\
\frac{1}{\chi^2_{n-1;1-\alpha/2}} & > & \frac{\sigma_X^2}{(n-1)s_X^2} & > & \frac{1}{\chi^2_{n-1;\alpha/2}} \\
\frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}} & > & \sigma_X^2 & > & \frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}}
\end{array}$$

y obtenemos el estimador por intervalos de confianza

$$\left(\frac{(n-1)S_X^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)S_X^2}{\chi^2_{n-1;1-\alpha/2}} \right)$$

6. El intervalo de confianza es:

$$\text{IC}_{1-\alpha}(\sigma_X^2) = \left(\frac{(n-1)s_x^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s_x^2}{\chi^2_{n-1;1-\alpha/2}} \right)$$

Ejemplo: calcular un intervalo de confianza para σ_X^2 y σ_X

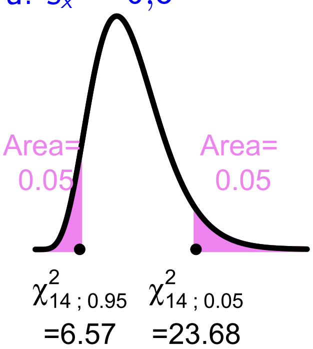
Ejemplo: 8.8 (Newbold) Una muestra aleatoria de quince pastillas para el dolor de cabeza tiene una cuasi desviación típica de 0.8 % en la concentración del ingrediente activo. Calcule un IC al 90 % para la varianza de la población para estas pastillas. Obtenga también un IC para la desviación típica de la población.

$$\boxed{\text{Objetivo: } \text{IC}_{0,9}(\sigma_X^2) = \left(\frac{(n-1)s_x^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s_x^2}{\chi^2_{n-1;1-\alpha/2}} \right)}$$

Población:
X = concentración del ingrediente activo en una pastilla (in %)" $X \sim N(\mu_X, \sigma_X^2)$

MAS: $n = 15$

Muestra: $s_x = 0{,}8$



$$s_x^2 = 0{,}8^2 = 0{,}64 \qquad n = 15$$

$$\begin{array}{rcl}
1-\alpha = 0{,}9 & \Rightarrow & \alpha/2 = 0{,}05 \\
\chi^2_{n-1;1-\alpha/2} & = & \chi^2_{14;0{,}95} = 6{,}57 \\
\chi^2_{n-1;\alpha/2} & = & \chi^2_{14;0{,}05} = 23{,}68 \\
\text{IC}_{0,9}(\sigma_X^2) & = & \left(\frac{14(0{,}64)}{23{,}68}, \frac{14(0{,}64)}{6{,}57} \right) \\
 & = & (0{,}378, 1{,}364) \Rightarrow \\
\text{IC}_{0,9}(\sigma_X) & = & (\sqrt{0{,}378}, \sqrt{1{,}364}) \\
 & = & (0{,}61, 1{,}17)
\end{array}$$

Para obtener IC(σ_X) aplicamos $\sqrt{\ }$ a los extremos de IC(σ_X^2)

Fórmulas para intervalos de confianza

Resumen para una población

- Sea $\underline{X}_n$ una muestra aleatoria simple de una población X con media μ_X y varianza σ_X^2

Parámetro	Hipótesis	Cantidad pivotal	$(1-\alpha)$ Intervalo Conf.
Media	Datos normales Varianza conocida	$\frac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1)$	$\mu_X \in \left(\bar{x} - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$
	Datos no normales Muestra grande	$\frac{\bar{X}-\mu_X}{\hat{\sigma}_X/\sqrt{n}} \sim_{approx.} N(0,1)$	$\mu_X \in \left(\bar{x} - z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}\right]$
	Datos Bernoulli Muestra grande	$\frac{\hat{p}_X-p_X}{\sqrt{\hat{p}_X(1-\hat{p}_X)/n}} \sim_{approx.} N(0,1)$	$p_X \in \left(\hat{p}_X \mp z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}\right]$
	Datos normales Varianza desconocida	$\frac{\bar{X}-\mu_X}{s_X/\sqrt{n}} \sim t_{n-1}$	$\mu_X \in \left(\bar{x} - t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}}\right)$
Varianza	Datos normales	$\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$	$\sigma_X^2 \in \left(\frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}}\right)$
Desv. típica	Datos normales	$\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$	$\sigma_X \in \left(\sqrt{\frac{(n-1)s_X^2}{\chi^2_{n-1;\alpha/2}}}, \sqrt{\frac{(n-1)s_X^2}{\chi^2_{n-1;1-\alpha/2}}}\right)$

Intervalos de confianza para la media de la población: ¿Qué usar cuándo?

$\boxed{X \sim \text{distribución con media } \mu_X \text{ y desviación típica } \sigma_X}$

- $X \sim$ normal
  - σ conocida → basada en z (exacta)
  - σ desconocida → basada en t (exacta)
- $X \nsim$ normal
  - n pequeña → métodos más allá de Est II
  - n grande → basada en z (aprox. TCL)