3. Estimación por intervalos de confianza

Profesora: María José Lombardía Cortiña

Departamento de Matemáticas. Universidade da Coruña

Programa Oficial de Postgrado en Estadística e Investigación Operativa
Máster en Técnicas Estadísticas

UNIVERSIDADE DA CORUÑA

1. Introducción
2. Método Pivotal
3. Intervalos de confianza para una muestra
   - I.C. para la media
   - I.C. para la varianza de una población normal
   - Trabajo práctico
   - I.C. para la proporción
   - Trabajo práctico
4. Intervalos de confianza para dos muestras
   - IC para la diferencia de medias
   - I.C. para la diferencia de medias con datos apareados
   - IC para el cociente de varianzas $\frac{\sigma_X^2}{\sigma_Y^2}$
   - IC para poblaciones no normales
5. Desigualdad de Tchebychev
6. Método Bootstrap
7. Intervalos de confianza bayesianos

UNIVERSIDADE DA CORUÑA

Estimación por intervalos

- La estimación puntual trata el problema de estimar mediante un número el valor de una característica poblacional o parámetro θ desconocido (por ejemplo, la estimación del IPC de un determinado período).
- En muchos casos la estimación puntual no es suficiente en el sentido de que no nos indica el error que se comete en dicha estimación.
- Lo razonable en la práctica es adjuntar, junto a la estimación puntual del parámetro, un cierto intervalo numérico que mida el margen de error que, de acuerdo a las observaciones muestrales, pueda tener dicha estimación.
- La idea de Intervalo de Confianza, es proponer un rango de valores entre los que *posiblemente* se encuentre el verdadero valor del parámetro θ.

UNIVERSIDADE DA CORUÑA

Ejemplo

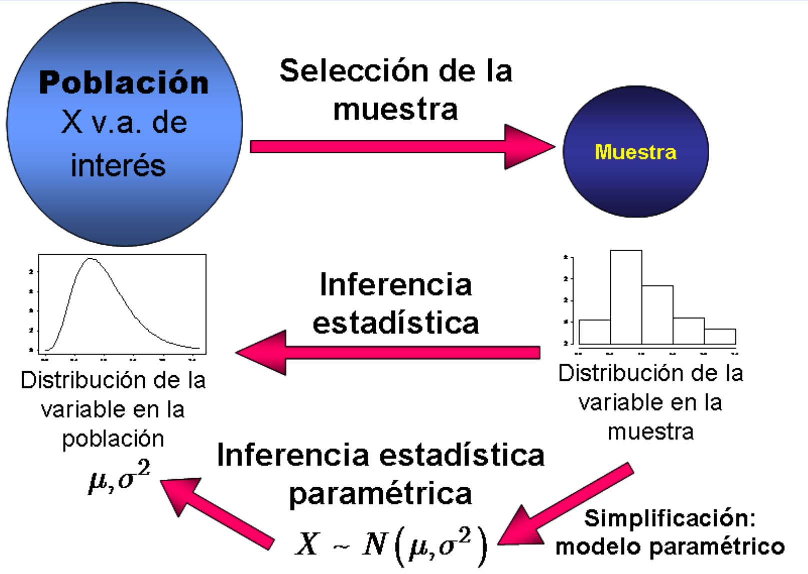
Con el fin de garantizar la salubridad de cierto alimento se ha llevado a cabo un análisis que cuenta el número de bacterias que se encuentran en 9 unidades del mismo. Los recuentos han dado los siguientes resultados:

$$157, 186, 179, 163, 171, 154, 177, 165, 168$$

Aceptando normalidad en la distribución del recuento bacteriano, elabora

(a) Estimaciones para la media y la varianza,

(b) El intervalo de confianza para la media y varianza a un nivel del 95%.


Población
X v.a. de
interés
Selección de la
muestra
Muestra
Inferencia
estadística
Distribución de la
variable en la
población
μ, σ^2
Distribución de la
variable en la
muestra
Inferencia estadística
paramétrica
$X \sim N(\mu, \sigma^2)$
Simplificación:
modelo paramétrico

Estimación por intervalos

Definición

La estimación por intervalos de confianza consiste en determinar un posible rango de valores o intervalo, en los que pueda precisarse –con una determinada probabilidad– que el valor de un parámetro se encuentra dentro de esos límites.

UNIVERSIDADE DA CORUÑA

Definición

Dada una muestra aleatoria $X_1, \ldots, X_n$, se denomina Intervalo de Confianza para el parámetro θ con nivel $1-\alpha$, a un intervalo aleatorio $\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$ –cuyos límites dependen de la muestra– de manera que:

$$\mathbb{P}\left(\widehat{\theta}_1\left(X_1, \ldots, X_n\right) \leq \theta \leq \widehat{\theta}_2\left(X_1, \ldots, X_n\right)\right) = 1-\alpha,$$

para cada $\theta \in \Theta$.

Si no se puede dar la igualdad anterior, al menos que la probabilidad en el intervalo supere $1-\alpha$.

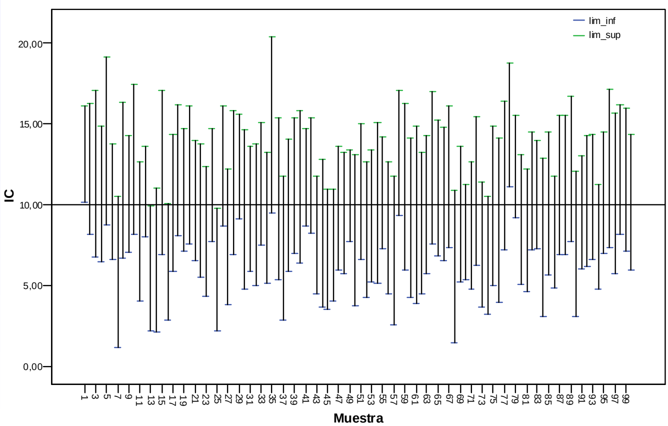
- Dada una realización muestral $x_1, \dots, x_n$.
- Una vez calculados los límites inferior y superior $\widehat{\theta}_1 (x_1, \dots, x_n)$ y $\widehat{\theta}_2 (x_1, \dots, x_n)$, respectivamente, se obtiene una intervalo numérico

$$\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$$

que constituye un intervalo de confianza para θ con nivel de confianza $1-\alpha$.

Interpretación: Si se toman infinitas muestras $\{x_1, \dots, x_n\}$ y construimos los correspondientes intervalos de confianza, el $100\,(1-\alpha)\,\%$ de ellos contendrían el verdadero valor del parámetro, mientras que $100\alpha\%$ no.

- El concepto de "confianza" en el intervalo es muy importante.
- Supongamos que queremos con una confianza del 95% que el verdadero valor del parámetro θ esté en un intervalo.
- Si en vez de considerar todas las muestras posibles consideramos 100 muestras. Entonces habrá más o menos 95 intervalos que contienen el valor del parámetro.



Observación

Como por lo general sólo vamos a disponer de una muestra, tenemos que confiar (por ejemplo con un 95% de confianza) que la muestra que tenemos pertenece al grupo de las muestras buenas (las que nos dan una estimación del intervalo que contiene el verdadero valor del parámetro).

El método pivotal

Para construir un IC (Intervalo de Confianza) para un parámetro θ usaremos el llamado Método Pivotal.

Definición

Se denomina estadístico pivotal $T(X_1, \ldots, X_n; \theta)$ *al estadístico cuya distribución en el muestreo no depende de* θ.

Ejemplo: Estadístico pivotal para la estimación de la media

Dada una población normal, $X \sim N(\mu, \sigma)$ se verifica que

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Tipificando

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

El método pivotal

Método pivotal para la construcción de IC:

0. Fijamos el nivel de confianza, $1-\alpha$ $(0<\alpha<1)$.
1. Elección del estadístico pivotal. Se elige un estadístico que dependa sólo del parámetro θ que se desea estimar y cuya distribución sea conocida y no dependa de θ (desconocido).
2. Planteamiento del enunciado probabilístico. Se plantea el enunciado probabilístico teniendo en cuenta la distribución de probabilidad del estadístico elegido en la etapa 1 y el valor $1-\alpha$ fijado para el nivel de confianza en la etapa 0. Es decir, se determinan constantes a y b tales que

$$\mathbb{P}\left(a < T\left(X_1,\ldots,X_n,\theta\right) < b\right) = 1-\alpha.$$

El método pivotal

3. Transformación del enunciado probabilístico. Si es posible despejar θ de la expresión anterior, obtenemos dos variables aleatorias $T_1^{-1}(X_1, \ldots, X_n)$ y $T_2^{-1}(X_1, \ldots, X_n)$

$$\mathbb{P}\left(T_1^{-1}(X_1, \ldots, X_n)\right) < \theta < T_2^{-1}(X_1, \ldots, X_n)\right) = 1 - \alpha$$

4. Con lo cual,

$$\left(T_1^{-1}(X_1, \ldots, X_n)\right), \; T_2^{-1}(X_1, \ldots, X_n)\right)$$

es el IC para θ con nivel de confianza $1 - \alpha$.

UNIVERSIDADE DA CORUÑA

El método pivotal

Ejemplo: Estadístico pivotal para la estimación de la media

Dada una población normal, $X \sim N(\mu, \sigma)$.

0. Fijamos el nivel de confianza, $1-\alpha$ $(0<\alpha<1)$.
1. Elección del estadístico pivotal.

$$\frac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

2. Planteamiento del enunciado probabilístico.

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

UNIVERSIDADE DA CORUÑA

El método pivotal

Ejemplo: Estadístico pivotal para la estimación de la media

Dada una población normal, $X \sim N(\mu, \sigma)$.

3. Transformación del enunciado probabilístico.

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

4. IC para μ con nivel de confianza $1 - \alpha$.

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

UNIVERSIDADE DA CORUÑA

Ejemplo

Supongamos que la variable X *representa el precio (en miles de euros) de la vivienda de alquiler en Madrid* y que se distribuye según una normal de media desconocida y varianza conocida $\sigma^2 = 20^2$.

- Para determinar el precio del alquiler medio en Madrid se toma una muestra de 70 viviendas obteniéndose que $\overline{x} = 82.5$.
- Se sabe que la variable

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

  es una normal estándar.
- Por tanto (sabiendo que $z_{0.025} = 1.96$)

$$\mathbb{P}\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Ejemplo

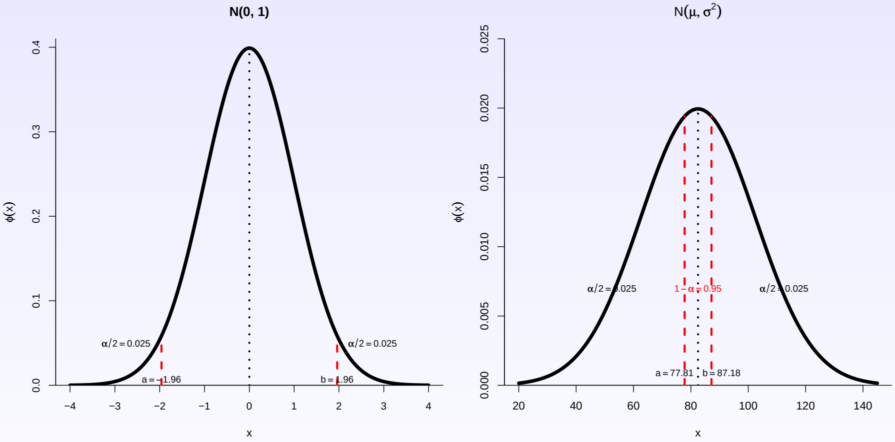
- Si de esta expresión se despeja el valor μ, se obtiene el intervalo probabilístico

$$\mathbb{P}\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- Como $\overline{x} = 82.5$ es la realización particular de $\overline{X}$, para esta muestra dada el intervalo de confianza para μ es

$$\left(82.5 - 1.96\frac{20}{\sqrt{70}}, 82.5 + 1.96\frac{20}{\sqrt{70}}\right) = (77.81, 87.18)$$

Luego el precio del alquiler medio en Madrid se encuentra en el intervalo anterior con una confianza del 95%. Recordar que tenemos que confiar (al 95%) en que este intervalo sea de los "*buenos*".


N(0, 1)
ϕ(x)
0.0
0.1
0.2
0.3
0.4
α/2 = 0.025
α/2 = 0.025
a = −1.96
b = 1.96
−4
−3
−2
−1
0
1
2
3
4
x
N(μ, σ2)
ϕ(x)
0.000
0.005
0.010
0.015
0.020
0.025
α/2 = 0.025
1 − α = 0.95
α/2 = 0.025
a = 77.81
b = 87.18
20
40
60
80
100
120
140
x

¿Cómo elegir el estadístico pivote?

- Se buscarán funciones sencillas de la muestra y del parámetro, cuya distribución no dependa del parámetro (θ).
- En los casos más generales, el estadístico pivote surge de la forma más natural.
- En algunas situaciones la búsqueda es más complicada. En estos casos, si tenemos muestras grandes, podemos recurrir como estadístico pivotal al estimador por Máxima Verosimilitud:

$$\widehat{\theta}_{MV} \text{ es asintóticamente } N\left(\theta, \sigma\left(\widehat{\theta}\right)\right).$$

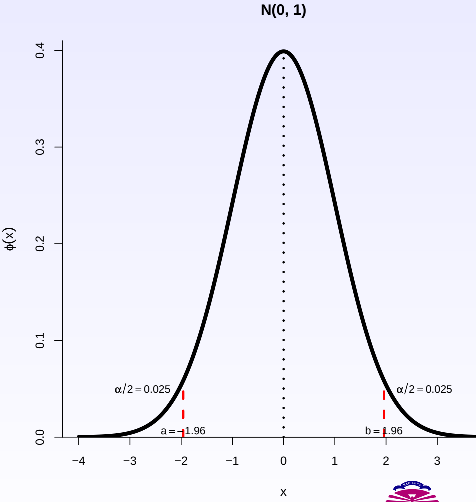
Entonces, el estadístico pivotal T es

$$T = \frac{\widehat{\theta}_{MV} - \theta}{\sigma\left(\widehat{\theta}_{MV}\right)} \text{ es asintóticamente } N(0,1)$$

UNIVERSIDADE DA CORUÑA

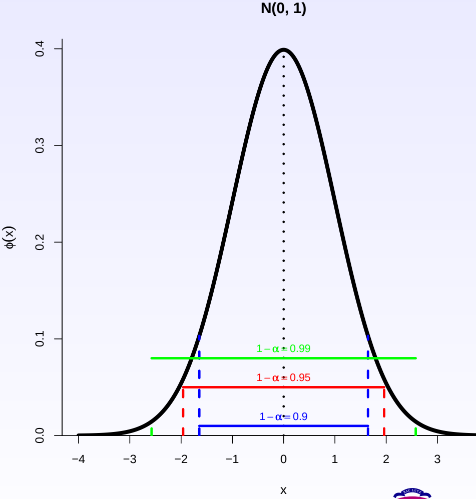
¿Cómo determinar las constantes *a* y *b*?

- Interesan constantes *a* y *b* de forma que el intervalo tenga longitud mínima. Con lo que la precisión será mayor.
- Para ello la distribución del estadístico pivote tiene que ser conocida. Si ésta es unimodal y simétrica, se tomará entonces el I.C. centrado alrededor del valor central, dejando $\alpha/2$ de probabilidad a ambos lados.
- Si la distribución es asimétrica la selección es complicada, y por sencillez la tomaremos como en el caso anterior, dejando $\alpha/2$ de probabilidad a ambos lados.



¿Cómo elegir α?

- Se elegirá α según el nivel de confianza $1-\alpha$ deseado, teniendo en cuenta que a menor α el intervalo será más largo y por tanto menos informativo.
- Normalmente se suele tomar como α uno de los siguientes valores: 0.1, 0.05 ó 0.01.



IC para una muestra

$$\begin{array}{c}\textbf{Población normal}\\ X \sim N(\mu, \sigma)\end{array} \quad \left\{ \begin{array}{l} I.C. \text{ para la media } \mu \left\{ \begin{array}{l} \sigma^2 \text{ conocida} \\ \sigma^2 \text{ desconocida} \end{array} \right. \\ I.C. \text{ para la varianza } \sigma^2 \end{array} \right.$$

$$\begin{array}{l}\textbf{Población no normal}\\ X \sim Ber(p)\end{array} \quad \Longrightarrow I.C. \text{ para la proporción } p$$

IC para la media con varianza conocida

- Supongamos una población cuya característica en estudio puede describirse mediante una variable aleatoria X con distribución $N(\mu, \sigma)$.
- Sea $X_1, \ldots, X_n$ una M.A.S. de X.
- Aplicando el método pivotal en este caso particular:

0. Fijamos el nivel de confianza, $1-\alpha$ $(0<\alpha<1)$.
1. Elección del estadístico pivotal.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Rightarrow \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

2. Planteamiento del enunciado probabilístico. Sea $z_{\alpha/2}$ el cuantil $\alpha/2$ de la distribución $N(0,1)$, es decir,

$$\mathbb{P}\left(Z \leq z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

La probabilidad de que cualquier variable aleatoria con distribución $N(0,1)$ tome valores en el intervalo $(-z_{\alpha/2}, z_{\alpha/2})$ es $1-\alpha$ y, por tanto:

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

3. Transformación del enunciado probabilístico. Si realizamos operaciones aritméticas:

- Multiplicar todos los términos por $\dfrac{\sigma}{\sqrt{n}}$.

$$\mathbb{P}\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- Restar en todos los términos $\overline{X}$.

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

- Multiplicar todos los términos por -1 para cambiar los signos

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

Se obtendría finalmente

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

4. IC para μ con nivel de confianza $1 - \alpha$. De esta forma hemos construido un intervalo probabílistico cuyos extremos son variables aleatorias

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

tales que la probabilidad de que tomen valores entre los que quede comprendido el verdadero valor de la media es $1 - \alpha$.

UNIVERSIDADE DA CORUÑA

Ejemplo

Supongamos que la variable X *representa el precio (en miles de euros) de la vivienda de alquiler en Madrid* y que se distribuye según una normal de media desconocida y varianza conocida $\sigma^2 = 20^2$.

- Para determinar el precio del alquiler medio en Madrid se toma una muestra de 70 viviendas obteniéndose que $\overline{x} = 82.5$.
- Se sabe que la variable

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

  es una normal estándar.
- Por tanto (sabiendo que $z_{0.025} = 1.96$)

$$\mathbb{P}\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Ejemplo

- Si de esta expresión se despeja el valor μ, se obtiene el intervalo probabilístico

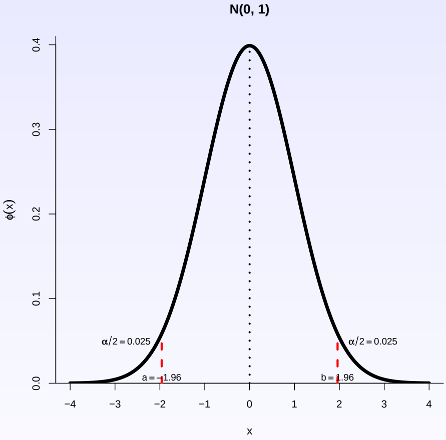
$$\mathbb{P}\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

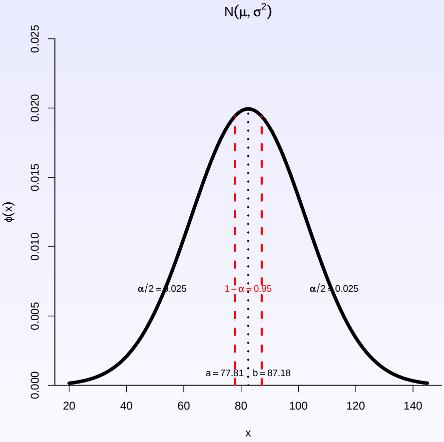
- Como $\overline{x} = 82.5$ es la realización particular de $\overline{X}$, para esta muestra dada el intervalo de confianza para μ es

$$\left(82.5 - 1.96\frac{20}{\sqrt{70}}, 82.5 + 1.96\frac{20}{\sqrt{70}}\right) = (77.81, 87.18)$$

Luego el precio del alquiler medio en Madrid se encuentra en el intervalo anterior con una confianza del 95%. Recordar que tenemos que confiar (al 95%) en que este intervalo sea de los "*buenos*".

UNIVERSIDADE DA CORUÑA

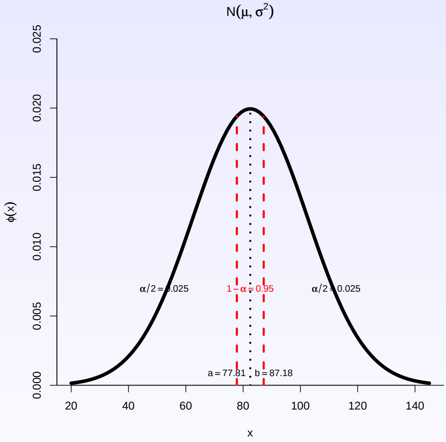

N(0, 1)
φ(x)
0.0
0.1
0.2
0.3
0.4
α/2 = 0.025
α/2 = 0.025
a = −1.96
b = 1.96
−4
−3
−2
−1
0
1
2
3
4
x


N(μ, σ²)
φ(x)
0.000
0.005
0.010
0.015
0.020
0.025
α/2 = 0.025
1 − α = 0.95
α/2 = 0.025
a = 77.81
b = 87.18
20
40
60
80
100
120
140
x

Observación

La media $\overline{X}$ está en el centro del intervalo.

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



Observación: Longitud del intervalo

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

- Cuanto mayor es la desviación típica σ, mayor es la longitud del intervalo.
- Cuanto mayor es el tamaño de la muestra n, menor es la longitud del intervalo.
- Cuanto mayor es el nivel de confianza $1 - \alpha$, mayor es la longitud del intervalo.

UNIVERSIDADE DA CORUÑA

IC para la media con varianza desconocida

En el caso de que desconozcamos el valor σ de la desviación típica, deberemos estimar dicho parámetro mediante el estimador (insesgado)

$$\hat{S}_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Eligiéndose como estadístico pivotal

$$\frac{\overline{X} - \mu}{\hat{S}_x/\sqrt{n}} \sim T_{n-1}$$

UNIVERSIDADE DA CORUÑA

En general, se procede de forma análoga a la anterior para calcular el intervalo de probabilidad al $100(1-\alpha)\%$ para la media μ de una población normal con varianza σ^2 desconocida.

0. Fijamos el nivel de confianza, $1-\alpha$ $(0<\alpha<1)$.
1. Elección del estadístico pivotal.

$$\frac{\overline{X}-\mu}{\hat{S}_x/\sqrt{n}} \sim T_{n-1}$$

2. Planteamiento del enunciado probabilístico.

$$\mathbb{P}\left(-t_{n-1,\alpha/2} < \frac{\overline{X}-\mu}{\hat{S}_x/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1-\alpha.$$

3. Transformación del enunciado probabilístico.

$$\mathbb{P}\left(\overline{X} - t_{n-1,\alpha/2}\frac{\hat{S}_x}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\alpha/2}\frac{\hat{S}_X}{\sqrt{n}}\right) = 1-\alpha.$$

4. IC para μ con nivel de confianza $1-\alpha$.

$$\boxed{\left(\overline{X}-t_{n-1,\alpha/2}\frac{\hat{S}_x}{\sqrt{n}},\ \overline{X}+t_{n-1,\alpha/2}\frac{\hat{S}_x}{\sqrt{n}}\right)}$$

UNIVERSIDADE DA CORUÑA

Ejemplo

Supongamos de nuevo que X representa el precio de la vivienda de alquiler en Madrid pero en este caso se desconoce tanto la media como la varianza (caso más habitual que el anterior).

- Con las mismas 70 viviendas se obtiene $\bar{x} = 82.5$ y una cuasivarianza $\hat{s}_x^2 = 21.4^2$.
- En este caso, se sabe que la variable

$$\frac{\overline{X} - \mu}{\hat{S}_X/\sqrt{n}} \sim T_{n-1}$$

es una T-Student con 69 grados de libertad.

UNIVERSIDADE DA CORUÑA

Ejemplo

Supongamos de nuevo que X representa el precio de la vivienda de alquiler en Madrid pero en este caso se desconoce tanto la media como la varianza (caso más habitual que el anterior).

- Por tanto, (sabiendo que $t_{69,0.025} = 2$) como $\bar{x} = 82.5$ es la realización particular de $\overline{X}$ y $\hat{s}_x = 21.4$ es la realización particular de $\widehat{S}_x$, el intervalo de confianza para μ es

$$\left(82.5 - 2\frac{21.4}{\sqrt{70}}, 82.5 + 2\frac{21.4}{\sqrt{70}}\right) = (77.38, 87.61)$$

- Luego, el precio del alquiler medio en Madrid se encuentra en el intervalo anterior con una confianza del 95%.

UNIVERSIDADE DA CORUÑA

I.C. para la media con muestras grandes

- Aunque la población base no sea Normal, por el Teorema Central del Límite, se verifica que cuando n se hace grande ($n > 30$), la distribución de $\overline{X}$ se puede aproximar por $N(\mu, \sigma/\sqrt{n})$.

- Por lo tanto, en el supuesto de σ desconocida, un intervalo de confianza para μ ($n > 30$) con nivel aproximado de confianza $1 - \alpha$ es:

$$\boxed{\left(\overline{X} - z_{\alpha/2}\frac{\hat{S}_x}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\hat{S}_x}{\sqrt{n}}\right)}$$

UNIVERSIDADE DA CORUÑA

Ejemplo

En el ejemplo anterior, se desconoce tanto la media como la varianza, sin embargo $n = 70 > 30$.

- Por tanto, (sabiendo que $z_{0.025} = 1.96$) como $\bar{x} = 82.5$ es la realización particular de $\overline{X}$ y $\hat{s}_x = 21.4$ es la realización particular de $\widehat{S}_x$, el intervalo de confianza para μ es

$$\left(82.5 - 1.96\frac{21.4}{\sqrt{70}}, 82.5 + 1.96\frac{21.4}{\sqrt{70}}\right) = (77.49, 87.51)$$

- Luego, el precio del alquiler medio en Madrid se encuentra en el intervalo anterior con un nivel de confianza aproximado del 95%.

UNIVERSIDADE DA CORUÑA

Determinación del tamaño de muestra

¿Cual debe ser el tamaño de muestra necesario para que fijado un nivel de confianza se alcance una precisión (longitud) deseada en el intervalo?

Determinación del tamaño de muestra

- Caso de σ conocida, se puede conocer el tamaño exacto:

$$L = b - a = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{despejando} \quad n = \frac{4z_{\alpha/2}\sigma^2}{L^2}.$$

- Caso σ desconocida:

$$L = b - a = 2t_{n-1,\alpha/2}\frac{\hat{S}}{\sqrt{n}},$$

siendo $\hat{S}$ y $t_{n-1,\alpha/2}$ desconocidos antes de tomar los datos. Una forma de proceder es, supuesto el tamaño muestral a determinar es suficientemente grande para aproximar $t_{n-1,\alpha/2}$ por $z_{\alpha/2}$, se toma una muestra preliminar para calcular $\hat{S}$ y se despeja n en la expresión anterior:

$$L = b - a = 2t_{n-1,\alpha/2}\frac{\hat{S}}{\sqrt{n}} \quad \text{despejando} \quad n = \frac{4z_{\alpha/2}\hat{S}^2}{L^2}.$$

Ejemplo

Los siguientes datos son los pesos (en gramos) de 16 pollos que se seleccionaron en una granja con el propósito de verificar el peso promedio: 506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496.

- Si el peso de cada pollo es una variable normal con desviación típica de 5gr.
  - (a) Obtener los intervalos de confianza al 90, 95, 99% para la media del peso de los pollos.
  - (b) Determinar el tamaño muestral necesario para que la precisión del intervalo (longitud) sea menor que la unidad, con $\alpha = 0.05$.
- Si ahora σ fuese desconocido, recalcular el apartado (a) y (b) anteriores.

Ejemplo

Los siguientes datos son los pesos (en gramos) de 16 pollos que se seleccionaron en una granja con el propósito de verificar el peso promedio: 506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496.

- Consideramos la variable X=peso en gramos $\in N(\mu, 5)$. Tenemos la siguiente información $n = 16$, $\bar{x} = 503.76$ y $\sigma = 5$

(a) $IC(\mu)$ con σ conocida:

$1-\alpha$	$z_{\alpha/2}$	Intervalo
0.90	1.64	(501.7, 505.8)
0.95	1.96	(501.3, 506.2)
0.99	2.58	(500.5, 506.9)

(b) El tamaño muestral necesario es

$$n \geq \frac{4z_{\alpha/2}\sigma^2}{L^2} = \frac{4 \cdot 1.96^2 \cdot 5^2}{1} = 384.16$$

En este caso, el tamaño de pollos debe ser mayor o igual que $n \geq 385$.

Ejemplo

Los siguientes datos son los pesos (en gramos) de 16 pollos que se seleccionaron en una granja con el propósito de verificar el peso promedio: 506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496.

- Consideramos la variable X=peso en gramos $\in N(\mu, \sigma)$. En este caso consideramos σ desconocida. A partir de la muestra obtenemos la siguiente información: $n = 16$, $\bar{x} = 503.76$ y $\hat{s} = 6.2022$.

(a) $IC(\mu)$ con σ desconocida:

$1-\alpha$	$t_{15\alpha/2}$	Intervalo
0.90	1.753	(501.03, 506.47)
0.95	2.131	(500.45, 507.05)
0.99	2.947	(499.18, 508.32)

(b) El tamaño muestral necesario es

$$n \geq \frac{4z_{\alpha/2}\hat{S}^2}{L^2} = \frac{4 \cdot 1.96^2 \cdot 6.2022^2}{1} = 591.103$$

En este caso, el tamaño de pollos debe ser mayor o igual que 592.

I.C. para la varianza, con media desconocida

- Supongamos una población cuya característica en estudio puede describirse mediante una variable aleatoria X con distribución $N(\mu, \sigma)$.
- Sea $X_1, \ldots, X_n$ una M.A.S. de X.
- Aplicando el método pivotal en este caso particular:

UNIVERSIDADE DA CORUÑA

0. Fijamos el nivel de confianza, $1 - \alpha$ $(0 < \alpha < 1)$.
1. Elección del estadístico pivotal. Tomamos el estimador insesgado de σ^2:

$$\widehat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Se utiliza el siguiente estadístico pivotal

$$\frac{(n-1)\,\widehat{S}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

2. Planteamiento del enunciado probabilístico. Se procede como en los anteriores casos, teniendo en cuenta que la distribución de χ^2_n no es simétrica $\left(\chi^2_{n-1,\alpha/2} \neq -\chi^2_{n-1,1-\alpha/2}\right)$:

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

3. Transformación del enunciado probabilístico. Despejando de esta expresión el valor σ^2 se obtiene

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

4. IC para σ^2 con nivel de confianza $1-\alpha$. De esta forma hemos construido un intervalo probabílistico cuyos extremos son variables aleatorias

$$\boxed{\left(\frac{(n-1)\,\widehat{S}^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\,\widehat{S}^2}{\chi^2_{n-1,1-\alpha/2}}\right)}$$

tales que la probabilidad de que tomen valores entre los que quede comprendido el verdadero valor de la varianza es $1-\alpha$.

UNIVERSIDADE DA CORUÑA

Ejemplo

Siguiendo con el ejemplo del precio del alquiler en Madrid se tenía que, para 70 viviendas, $\overline{x} = 82.5$ y $\widehat{s}_X^2 = 21.4^2$.

Nos preguntamos ahora cuál será el intervalo de confianza al 95% para la varianza σ^2.

- En este caso, se sabe que la variable

$$\frac{(n-1)\,\hat{S}^2}{\sigma^2} \sim \chi^2_{n-1}$$

  es una χ^2_{69} con 69 grados de libertad
- Sabiendo que $\chi^2_{69,0.975} = 47.92$ y $\chi^2_{69,0.025} = 93.86$)
-

$$\begin{aligned} 0.95 \quad &= \mathbb{P}\left(47.92 < \frac{(n-1)\hat{S}_X^2}{\sigma^2} < 93.86\right) \\ &= \mathbb{P}\left(\frac{(n-1)\hat{S}_X^2}{93.86} < \sigma^2 < \frac{(n-1)\hat{S}_X^2}{47.92}\right) \end{aligned}$$

Ejemplo

- Como $\widehat{s}_X^2 = 21.4^2$ es la realización particular de $\hat{S}_X^2$.
- el intervalo de confianza para σ^2 es

$$\left(\frac{(70-1)\,21.4^2}{93.86}, \frac{(70-1)\,21.4^2}{47.92}\right) = (336.86, 659.41)$$

Luego la varianza del precio del alquiler en Madrid se encuentra en el intervalo anterior con una confianza del 95%.

Ejemplo

Con el fin de garantizar la salubridad de cierto alimento se ha llevado a cabo un análisis que cuenta el número de bacterias que se encuentran en 9 unidades del mismo. Los recuentos han dado los siguientes resultados:

$$157, 186, 179, 163, 171, 154, 177, 165, 168.$$

Aceptando normalidad en la distribución del recuento bacteriano, elabora

(a) Estimaciones para la media y la varianza,

(b) El intervalo de confianza para la media y varianza a un nivel del 95%.

Solución

Aceptando normalidad en la distribución del recuento bacteriano,

(a) Estimaciones para la media y la varianza: 168.9 es la estimación del número medio de bacterias con una varianza de $\hat{s}^2 = 109.86$ o bien $s^2 = 97.65$.

(b) El intervalo de confianza para la media y varianza a un nivel del 95%.
   - El número medio de bacterias está entre 160.82 y 176.96, con un nivel de confianza del 95%.
   - Su varianza se encuentra entre 50.22 y 403.16, con un nivel de confianza del 95%.

UNIVERSIDADE DA CORUÑA

Práctica en R:

Estamos interesados en conocer el contenido de *cafeína* del café y su variabilidad. Para ello contamos con una muestra de 43 granos de café de distintos países. Aplicando técnicas de inferencia estadística estima el contenido medio de cafeína y su varianza, con un nivel de confianza del 95%.

- Repite el estudio para el *contenido en agua*, el *peso de la semilla*, las *grasas* y el *contenido en minerales* del café.
- Para cada variable estudiada compara el intervalo obtenido con el calculado para un nivel de confianza del 90% y del 99%. ¿Qué observas? ¿Cómo influye el nivel de confianza en la precisión del intervalo?
- ¿Qué tamaño de muestra debería tomarse para que el error cometido en la estimación de la media no sea superior a 0.10?

Información disponible en la base de datos: cafe.xls

Ejemplo

En un sondeo realizado sobre 50 alumnos de la facultad de Informática se ha encontrado que 12 de ellos no están satisfechos con el funcionamiento del Centro de Cálculo.

(a) Determinar un intervalo de confianza para la proporción de descontentos, a un nivel de confianza del 99%.

(b) ¿Cuál es el tamaño muestral necesario para que, con el nivel de confianza anterior, la longitud del intervalo de confianza sea menor o igual que una centésima?

UNIVERSIDADE DA CORUÑA

I.C. para la proporción

1. El objetivo es contruir un I.C. para la proporción de elementos p de una población que poseen determinada característica de interés, a partir de una M.A.S. de elementos de la población.

2. Para cada elemento de la población anotaremos

$$X = \begin{cases} 1 & \text{(tiene la caract.) con probab. } p \\ 0 & \text{(no la tiene) con probab. } 1-p \end{cases},$$

3. M.A.S. $X_1, X_2, ..., X_n$, copias de una variable $X \in Ber(p)$.

0. Fijamos el nivel de confianza, $1-\alpha$ $(0<\alpha<1)$.
1. Elección del estadístico pivotal. Dado que para n grande el estimador de p:
$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n}$$
tiene distribución aproximadamente $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, utilizaremos el estadístico pivote
$$T = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \quad (n>30).$$

2.-3. Planteamiento del enunciado probabilístico y transformación del enunciado probabilístico. Teniendo en cuenta esta aproximación y realizando cálculos similares a los llevados a cabo para la obtención de los I.C. para la media.

4. IC para p con nivel de confianza $1-\alpha$. El intervalo de confianza a nivel $1-\alpha$ viene dado por

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

UNIVERSIDADE DA CORUÑA

Observaciones

El intervalo anterior no es calculable (depende del valor desconocido p). Para resolver dicho problema en la práctica:

(a) Sustituir p por su estimador $\hat{p}$. El intervalo resultante es

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

(b) Considerar la situación más desfavorable, es decir tomar $p = 1/2$ y reemplazar $p(1-p)$ por su valor máximo, $1/4$. El intervalo que se obtiene (lógicamente más largo) es

$$\boxed{\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}}\right)}.$$

UNIVERSIDADE DA CORUÑA

Determinación del tamaño de muestra

¿Cual debe ser el tamaño de muestra necesario para que fijado un nivel de confianza se alcance una precisión (longitud) deseada en el intervalo?

Determinación del tamaño de muestra

- Al igual que en el caso de la media, se obtiene despejando n en la expresión de la Longitud del intervalo (L)

$$L = 2z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \implies n = \frac{4z^2_{\alpha/2}p(1-p)}{L^2}.$$

- El problema del desconocimiento de p se sigue manteniendo, por lo que las dos soluciones apuntadas anteriormente siguen siendo válidas. Tomando la segunda solución propuesta: tomar $p = 1/2$ y reemplazar $p(1-p)$ por su valor máximo, $1/4$.

$$n = \frac{z^2_{\alpha/2}}{L^2}$$

UNIVERSIDADE DA CORUÑA

Ejemplo

En un sondeo realizado sobre 50 alumnos de la facultad de Informática se ha encontrado que 12 de ellos no están satisfechos con el funcionamiento del Centro de Cálculo.

(a) Determinar un intervalo de confianza para la proporción de descontentos, a un nivel de confianza del 99%.

(b) ¿Cuál es el tamaño muestral necesario para que, con el nivel de confianza anterior, la longitud del intervalo de confianza sea menor o igual que una centésima?

UNIVERSIDADE DA CORUÑA

Ejemplo

(a) *Determinar un intervalo de confianza para la proporción de descontentos, a un nivel de confianza del 99%.*

Teniendo en cuenta la información de la muestra: $\hat{p} = 12/50 = 0.24$, el tamaño muestral $n = 50$. El nivel de confianza es $1 - \alpha = 0.99 \;\Rightarrow\; \alpha = 0.01$ y $z_{\alpha/2} = 2.58$.
Los IC para la proporción de descontentos son:

- **Si reemplazamos p por $\hat{p}$**

$$\begin{aligned} IC(p) &= \left(0.24 - 2.58\sqrt{\tfrac{0.24(1-0.24)}{50}},\ 0.24 + 2.58\sqrt{\tfrac{0.24(1-0.24)}{50}}\right) \\ &= (0.24 \pm 0.1558) = (0.084, 0.396). \end{aligned}$$

- **Si reemplazamos p por $1/2$**

$$\begin{aligned} IC(p) &= \left(0.24 - 2.58\sqrt{\tfrac{1}{4\cdot 50}},\ 0.24 + 2.58\sqrt{\tfrac{1}{4\cdot 50}}\right) \\ &= (0.24 \pm 0.1824) = (0.058, 0.422). \end{aligned}$$

Ejemplo

(a) *Determinar un intervalo de confianza para la proporción de descontentos, a un nivel de confianza del 99%*.

El porcentaje de descontentos se sitúa entre el el 8.42% y el 39.58% con un nivel de confianza del 99% o bien el 5.76% y el 42.24%, según la aproximación utilizada.

UNIVERSIDADE DA CORUÑA

Ejemplo

(b) *¿Cuál es el tamaño muestral necesario para que, con el nivel de confianza anterior, la longitud del intervalo de confianza sea menor o igual que una centésima?*.

$L \leq 0.01$, según sustituyamos p por $\hat{p} = 0.24$ o por $\hat{p} = 1/2$ en la expresión

$$n \geq \frac{4z^2_{\alpha/2}p(1-p)}{L^2},$$

se obtiene respectivamente:

- $n \geq \frac{4 \cdot 2.58^2 0.24(1-0.24)}{0.01^2} = 48565.094$, es decir, $n \geq 48566$.
- $n \geq \frac{4 \cdot 2.58^2 0.5 \cdot 0.5}{0.01^2} = 66564$, es decir, $n \geq 66564$.

UNIVERSIDADE DA CORUÑA

Práctica en R:

Estamos interesados en conocer la proporción de café arábigo que se produce y de café robusto. Para ello se toma una muestra de 43 granos de café anotando su procedencia y el tipo al que pertenecen. Aplicando técnicas de inferencia estadística haz un estudio de esta proporción:

- Mediante técnicas de Estimación puntual.
- Utilizando estimación por intervalos de confianza (para un nivel de confianza del 90%, 95% y 99%).
- ¿Cuál es el error cometido en la estimación?

Información disponible en la base de datos: cafe.xls

Se disponen de dos muestras aleatorias $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ de dos poblaciones de interés X e Y, siendo el objetivo construir intervalos que permitan comparar parámetros poblacionales de X e Y.

UNIVERSIDADE DA CORUÑA

Muestras independientes y muestras apareadas

Un equipo de sociólogos pretende comparar el tiempo medio diario que los adolescentes dedican a ver la televisión, con el que invierten en la lectura.

- $X =$ Tiempo diario dedicado a ver la televisión

$$\Downarrow$$

$$\text{M.A.S. } X_1, X_2, \ldots X_n$$

- $Y =$ Tiempo dedicado diariamente a la lectura

$$\Downarrow$$

$$\text{M.A.S. } Y_1, Y_2, \ldots Y_n$$

Interesa construir un IC para la diferencia de medias $\mu_x - \mu_y$.

UNIVERSIDADE DA CORUÑA

Muestras independientes y muestras apareadas

La construcción de dicho intervalo precisa de dos muestras aleatorias de ambas variables. Dos posibilidades:

- MUESTRAS INDEPENDIENTES: Seleccionar jóvenes al azar, de forma independiente hasta formar dos grupos de tamaño n, y evaluar X en un grupo e Y en el otro.
- MUESTRAS APAREADAS: Seleccionar al azar un único grupo con n jóvenes y evaluar conjuntamente X e Y en el mismo grupo.

Muestras independientes y muestras apareadas

La construcción de dicho intervalo precisa de dos muestras aleatorias de ambas variables. Dos posibilidades:

- MUESTRAS INDEPENDIENTES: Los resultados obtenidos en una no condicionan los resultados de la otra. El carácter aleatorio de la formación de ambos grupos consigue aumentar la probabilidad de que las características determinantes de la respuesta se distribuyan equitativamente entre ambos grupos.
- MUESTRAS APAREADAS: Las observaciones se han recogido a pares por cada unidad muestral y se habla entonces de *muestras apareadas*. Las muestras son claramente dependientes (se espera que observaciones bajas de X se apareen con observaciones altas de Y y viceversa). Un muestreo apareado reduce notablemente la variabilidad experimental al conseguir grupos muestrales más homogéneos.

Se estudiará la construcción de intervalos para ambos tipos de muestreo.

UNIVERSIDADE DA CORUÑA

Poblaciones normales independientes
$X \sim N(\mu_x, \sigma_x), \quad Y \sim N(\mu_y, \sigma_y)$

$$\begin{cases} IC(\mu_x - \mu_y) \begin{cases} \sigma_x^2,\ \sigma_y^2 \text{ conocidas} \\ \sigma_x^2 = \sigma_y^2 \text{ desconocidas} \\ \sigma_x^2,\ \sigma_y^2 \text{ desconocidas} \end{cases} \\ IC\left(\frac{\sigma_x^2}{\sigma_y^2}\right) \end{cases}$$

Poblaciones normales con datos apareados
$D = X - Y \sim N(\mu_x - \mu_y, \sigma_D)$ $\Longrightarrow IC(\mu_D)$

Poblaciones no normales
$X \sim Ber(p_x),\ Y \sim Ber(p_y)$ $\Longrightarrow IC(p_x - p_y)$

UNIVERSIDADE DA CORUÑA

IC para la diferencia de medias de poblaciones normales independientes

Ejemplo

Interesa estudiar la diferencia de precio de la vivienda en alquiler entre Madrid y Barcelona.
Supongamos que

- X representa el precio de la vivienda de alquiler en Madrid.
- Y representa el precio de la vivienda de alquiler en Barcelona.
- Además, las varianzas ....

IC para la diferencia de medias de poblaciones normales independientes

Sean $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ dos muestras aleatorias recogidas independientemente de dos poblaciones normales $X \sim N\left(\mu_X, \sigma_X^2\right)$ e $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$.
Se desea construir intervalos de confianza para la diferencia de medias $\mu_X - \mu_Y$.

Caso 1: varianzas conocidas

- Puesto que $\overline{X}$ e $\overline{Y}$ son independientes:

$$\left.\begin{array}{l} \overline{X} \sim N\left(\mu_X, \dfrac{\sigma_X}{\sqrt{n}}\right) \\ \overline{Y} \sim N\left(\mu_Y, \dfrac{\sigma_Y}{\sqrt{m}}\right) \end{array}\right\} \Rightarrow \overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$

UNIVERSIDADE DA CORUÑA

IC para la diferencia de medias de poblaciones normales independientes

- Por tanto
$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$
puede ser utilizado como estadístico pivote puesto que tiene una distribución conocida independiente de μ_X y μ_Y.
- Con argumentos idénticos a los realizados en intervalos anteriores se llega a que un intervalo de confianza para $\mu_X - \mu_Y$ con un nivel de confianza $1-\alpha$ es

$$\boxed{\left((\overline{x} - \overline{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\overline{x} - \overline{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)}$$

Ejemplo

Supongamos que X representa el precio de la vivienda de alquiler en Madrid e Y representa el precio de la vivienda de alquiler en Barcelona. Supongamos que las varianzas σ_X^2 y σ_Y^2 son conocidas. En este caso,

$$T = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

donde n es el tamaño de la primera muestra (los pisos de Madrid) y m es el tamaño de la segunda muestra (los pisos de Barcelona).
En este ejemplo de las viviendas $\overline{x} = 82.5$, $\overline{y} = 85.4$, $\sigma_X^2 = 20^2$, $\sigma_Y^2 = 25^2$, $n = 70$ y $m = 50$.

UNIVERSIDADE DA CORUÑA

Ejemplo

El intervalo de confianza al 95% para la diferencia de precios medios de alquiler entre Madrid y Barcelona es

$$\left((\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right)$$
$$= \quad (-11.2649, 5.4649)$$

Por tanto, al 95% de confianza no se puede concluir que los precios del alquiler en Barcelona sean más altos que los precios en Madrid.

UNIVERSIDADE DA CORUÑA

IC para la diferencia de medias de poblaciones normales independientes

Caso 2: varianzas desconocidas pero iguales

- Ahora supongamos que las varianzas σ_X^2 y σ_Y^2 son desconocidas pero iguales (a σ^2).

  En este caso, se sabe que la variable

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X}{\sqrt{n}} + \frac{\sigma_Y}{\sqrt{m}}\right)$$

$$\Leftrightarrow \overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}\right)$$

  y por tanto

$$T = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N\left(0, 1\right)$$

IC para la diferencia de medias de poblaciones normales independientes

- Como en la expresión anterior no se conoce el valor de σ^2, éste se puede sustituir por el valor

$$\widehat{S}^2 = \frac{(n-1)\,\widehat{S}_X^2 + (m-1)\,\widehat{S}_Y^2}{n+m-2}$$

que es una ponderación de las dos cuasivarianzas muestrales $\widehat{S}_X^2$ y $\widehat{S}_Y^2$. Se puede demostrar que

$$T = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_X - \mu_Y)}{\widehat{S}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

UNIVERSIDADE DA CORUÑA

IC para la diferencia de medias de poblaciones normales independientes

- Luego, un intervalo de confianza para $\mu_X - \mu_Y$ a un nivel de confianza $1-\alpha$ es el siguiente

$$\boxed{\left((\overline{x} - \overline{y}) - t_{n+m-2,\alpha/2}\widehat{s}\sqrt{\frac{1}{n} + \frac{1}{m}}, (\overline{x} - \overline{y}) + t_{n+m-2,\alpha/2}\widehat{s}\sqrt{\frac{1}{n} + \frac{1}{m}} \right)}$$

Nota

Puesto que la construcción de este intervalo requiere suponer que las varianzas, aunque desconocidas, son iguales, es necesario realizar previamente un contraste de hipótesis para garantizar dicha suposició n.

UNIVERSIDADE DA CORUÑA

Ejemplo

Supongamos de nuevo que X representa el precio de la vivienda de alquiler en Madrid e Y representa el precio de la vivienda de alquiler en Barcelona. Supongamos que las varianzas σ_X^2 y σ_Y^2 son desconocidas pero iguales. Se obtenían los siguientes datos: $\overline{x} = 82.5$, $\overline{y} = 85.4$, $\widehat{s}_X^2 = 21.4^2$, $\widehat{s}_Y^2 = 24.3^2$, $n = 70$ y $m = 50$.

$$\widehat{s}^2 = \frac{(n-1)\,\widehat{s}_X^2 + (m-1)\,\widehat{s}_Y^2}{n+m-2} = \frac{69 \cdot \widehat{s}_X^2 + 49 \cdot \widehat{s}_Y^2}{118} = 512.9936$$

Ejemplo

Por tanto, el intervalo de confianza para la diferencia de precios en el alquiler entre Madrid y Barcelona al 95% de confianza es

$$\left((\overline{x}-\overline{y}) \pm t_{n+m-2,\alpha/2}\widehat{s}\sqrt{\frac{1}{n}+\frac{1}{m}}\right) = (-11.2038, 5.4038)$$

Tampoco se puede concluir con un 95% de confianza que los precios del alquiler en Barcelona sean más altos que los de Madrid.

IC para la diferencia de medias de poblaciones normales independientes

Caso 3: varianzas desconocidas y distintas

- Si las varianzas no se pueden suponer iguales, el estadístico anterior no se puede utilizar. En el caso de que las varianzas fuesen conocidas, el estadístico a utilizar era

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

UNIVERSIDADE DA CORUÑA

IC para la diferencia de medias de poblaciones normales independientes

a) Tamaños muestrales grandes ($n, m \geq 30$)

En este caso se utiliza el estadístico pivote anterior estimando σ_X por $\widehat{S}_X$ y σ_Y por $\widehat{S}_Y$, es decir,

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\widehat{S}_X^2}{n} + \frac{\widehat{S}_Y^2}{m}}} \underset{\text{Asint}}{\sim} N(0,1)$$

que tiene una distribución aproximadamente normal $N(0,1)$.

IC para la diferencia de medias de poblaciones normales independientes

- Luego, un intervalo de confianza para $\mu_X - \mu_Y$ a un nivel de confianza $1-\alpha$ es el siguiente

$$\boxed{\left((\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{\widehat{S}_x^2}{n} + \frac{\widehat{S}_y^2}{m}}, (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{\widehat{S}_x^2}{n} + \frac{\widehat{S}_y^2}{m}} \right)}$$

UNIVERSIDADE DA CORUÑA

IC para la diferencia de medias de poblaciones normales independientes

b) Tamaños muestrales pequeños ($n, m < 30$)

En este caso la aproximación normal no es nada precisa. La solución más usada habitualmente es la aproximación debida a Welch, según la cual el estadístico anterior sigue una distribución t de Student con $g = n + m - 2 - \delta$ grados de libertad, con δ el entero más próximo a

$$\Delta = \frac{\left[(m-1)\dfrac{\widehat{S}_X^2}{n} - (n-1)\dfrac{\widehat{S}_Y^2}{m}\right]^2}{(m-1)\left(\dfrac{\widehat{S}_X^2}{n}\right)^2 + (n-1)\left(\dfrac{\widehat{S}_Y^2}{m}\right)^2}$$

Se comprueba que $0 \leq \delta \leq \max(n-1, m-1)$.

IC para la diferencia de medias de poblaciones normales independientes

- Luego, un intervalo de confianza para $\mu_X - \mu_Y$ a un nivel de confianza $1 - \alpha$ es el siguiente

$$\boxed{\left((\overline{X} - \overline{Y}) - t_{g,\alpha/2}\sqrt{\frac{\widehat{S}_x^2}{n} + \frac{\widehat{S}_y^2}{m}}, (\overline{X} - \overline{Y}) + t_{g,\alpha/2}\sqrt{\frac{\widehat{S}_x^2}{n} + \frac{\widehat{S}_y^2}{m}}\right)}$$

UNIVERSIDADE DA CORUÑA

Ejemplo

En un estudio realizado sobre el tipo de sedimentos hallados en dos lugares de perforación distintos, se han anotado los siguientes datos acerca del porcentaje en volumen de arcilla presente en las muestras del sondeo:

$$X : 31, 18, 17, 16, 37, 16, 32, 13, 14, 49, 25, 19, 13, 32, 27.$$

$$Y : 15, 17, 13, 25, 22, 20, 24, 12, 23, 15, 20, 18.$$

- $X =$% de arcilla del lugar A
- $Y =$% de arcilla del lugar B
- Calcular un IC al 95% para la diferencia de los valores medios de X e Y.

UNIVERSIDADE DA CORUÑA

Ejemplo

- A partir de los datos muestrales se obtienen los siguientes valores:

$$n = 15, \ \overline{X} = 23.933, \ \hat{S}_x = 10.559$$

$$m = 12, \ \overline{Y} = 18.667, \ \hat{S}_y = 4.355$$

- Supuesto que X e Y son variables normales con varianza desconocida y distintas, para el cálculo del IC se utiliza la expresión del **Caso 3**:

$$\left((\overline{X} - \overline{Y}) - t_{g,\alpha/2}\sqrt{\frac{\widehat{S}_x^2}{n} + \frac{\widehat{S}_y^2}{m}}, (\overline{X} - \overline{Y}) + t_{g,\alpha/2}\sqrt{\frac{\widehat{S}_x^2}{n} + \frac{\widehat{S}_y^2}{m}} \right)$$

- Con los datos anteriores: $\Delta = 5.5334$, $g = n + m - 2 - \delta = 15 + 12 - 2 - 6 = 19$ y $t_{g,0.025} = 2.093024$.

- Sustituyendo los valores en el $IC(\mu_x - \mu_y)$ al 95% se obtiene $(-1.0177, 11.5497)$.

UNIVERSIDADE DA CORUÑA

Determinación del tamaño muestral

Supuesto que eligiésemos muestras del mismo tamaño en ambas poblaciones ($m = n$), podemos preguntarnos por el tamaño muestral necesario para que la longitud del intervalo para la diferencia de medias sea igual a una cantidad predeterminada, para un nivel de confianza prefijado $1 - \alpha$.

UNIVERSIDADE DA CORUÑA

Determinación del tamaño muestral

- Caso 1: Si las varianzas poblacionales son conocidas. Despejando n en la longitud del intervalo

$$L = 2z_{\alpha/2}\sqrt{\frac{\sigma_x^2 + \sigma_y^2}{n}}$$

se obtiene

$$n = \frac{4z_{\alpha/2}^2(\sigma_x^2 + \sigma_y^2)}{L^2}.$$

Determinación del tamaño muestral

- Caso 2: Si las varianzas poblacionales son desconocidas pero pueden suponerse iguales. Dada una estimación preliminar de la varianza común $\widehat{S}^2$ y supuesto que n es suficientemente grande para aproximar la distribución de t por la normal estándar, despejando n en la longitud del intervalo

$$L = 2z_{\alpha/2}\widehat{S}^2\sqrt{\frac{2}{n}}$$

se obtiene

$$n = \frac{8z_{\alpha/2}^2\widehat{S}^2}{L^2}.$$

UNIVERSIDADE DA CORUÑA

Determinación del tamaño muestral

- Caso 3: Si las varianzas poblacionales son desconocidas y no se asumen iguales. supuesto que n es suficientemente grande para aproximar la distribución de t por la normal estándar, el valor de n que se obtiene al despejar en la longitud del intervalo,

$$L = 2z_{\alpha/2}\sqrt{\frac{\widehat{S}_x^2 + \widehat{S}_y^2}{n}}$$

es

$$n = \frac{4z_{\alpha/2}^2(\widehat{S}_x^2 + \widehat{S}_y^2)}{L^2},$$

donde $\widehat{S}_x^2$ y $\widehat{S}_y^2$ son estimaciones preliminares de las varianzas poblacionales.

UNIVERSIDADE DA CORUÑA

Determinación del tamaño muestral

Ejemplo

Queremos que la longitud del IC del 95% para la diferencia del precio en alguiler de la vivienda entre Madrid y Barcelona sea inferior a 8 unidades. Supuesto que $m = n$ y que las varianzas son iguales, utilizando como estimación preliminar el valor calculado en el ejemplo anterior $\hat{s}^2 = 512.9936$, el tamaño muestral requerido es

$$n > \frac{8z_{0.025}^2 \widehat{S}^2}{L^2} = \frac{8 \cdot 1.96^2 \cdot 512.9936}{8^2} = 246.34.$$

Es decir, $n \geq 247$ y $n = m$.

I.C. para la diferencia de medias con datos apareados

Muestreo Apareado

La principal característica del muestreo apareado es que ambas muestras son claramente dependientes, de modo que los estadísticos pivote usados en caso de independencia no se pueden emplear en este caso, puesto que los intervalos de confianza pueden salir excesivamente grandes o pequeños.

El motivo es que si las variables X e Y son dependientes

$$Var\left(\overline{X}-\overline{Y}\right)=Var\left(\overline{X}\right)+Var\left(\overline{Y}\right)-2Cov\left(\overline{X},\overline{Y}\right)$$

con lo que el denominador de los estadísticos pivote puede ser equivocadamente grande o pequeño, según sea la dependencia entre X e Y, es decir, según sea $Cov\left(\overline{X},\overline{Y}\right)$.

UNIVERSIDADE DA CORUÑA

I.C. para la diferencia de medias con datos apareados

La Solución

- Considerar una nueva variable aleatoria $D = X - Y$,

$$\left.\begin{array}{l} X \sim N(\mu_X, \sigma_X) \\ Y \sim N(\mu_Y, \sigma_Y) \end{array}\right\} \Rightarrow D = X - Y \sim N(\mu_D, \sigma_D).$$

  Donde $\mu_D = \mu_X - \mu_Y, \sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$.

- Tipificando

$$\frac{D - \mu_D}{\sigma_D} \sim N(0,1)$$

- Puesto que σ_D es desconocido, si estimamos su valor por la cuasivarianza muestral $\widehat{S}_D$, el estadístico pivote en este caso es

$$\frac{D - \mu_D}{\widehat{S}_D} \sim t_{n-1}$$

I.C. para la diferencia de medias con datos apareados

La Solución

- El I.C. para $\mu_X - \mu_Y$ equivale al intervalo para μ_D. Para ello disponemos de una muestra aleatoria de la variable D sin más que considerar los valores $D_i = X_i - Y_i$, obtenidos a partir de las muestras apareadas $\{(X_i, Y_i)\}_{i=1}^{n}$.
- Recordando la construcción de I.C. para la media de una población normal, se obtiene que el intervalo de probabilidad con nivel $1 - \alpha$ para $\mu_X - \mu_Y = \mu_D$ es

$$\boxed{\left((\overline{X} - \overline{Y}) - t_{n-1,\alpha/2} \frac{\widehat{S}_D}{\sqrt{n}}, (\overline{X} - \overline{Y}) + t_{n-1,\alpha/2} \frac{\widehat{S}_D}{\sqrt{n}} \right)}$$

UNIVERSIDADE DA CORUÑA

Ejemplo

Supongamos que X representa el salario de un ejecutivo antes de hacer un MBA e Y representa el salario de un ejecutivo después de hacer un MBA. La siguiente tabla muestra los salarios correspondientes a 7 ejecutivos (en miles de euros mensuales):

x_i	2.7838	2.1672	3.0627	3.1428	2.4268	3.5955	3.5946
y_i	3.8962	3.8169	4.1051	3.9720	4.1033	3.9686	5.2383
D_i	-1.1125	-1.6497	-1.0424	-0.8281	-1.6765	-0.3731	-1.6437

La última fila de la tabla representa la variable $D = X - Y$.

Ejemplo

En el caso de los 7 ejecutivos, se obtiene que $\overline{D} = -1.1894$ y $\widehat{s}_D^2 = 0.2466$. Sabiendo que $t_{6,0.025} = -t_{6,0.975} = 2.447$, el intervalo de confianza para $\mu_D = \mu_X - \mu_Y$ es

$$\begin{aligned}
&\left((\overline{x}-\overline{y}) - t_{n-1,\alpha/2}\frac{\widehat{s}_D}{\sqrt{n}}, (\overline{x}-\overline{y}) + t_{n-1,\alpha/2}\frac{\widehat{s}_D}{\sqrt{n}}\right) \\
=&\left(-1.1894 - 2.447\frac{\sqrt{0.2466}}{\sqrt{7}}, -1.1894 + 2.447\frac{\sqrt{0.2466}}{\sqrt{7}}\right) \\
=&(-1.648, -0.730)
\end{aligned}$$

- Luego la diferencia de los salarios medios de los ejecutivos antes y después de hacer un MBA se encuentra en el intervalo anterior con una confianza del 95%.
- Se observa que esta diferencia es menor que 0 luego se puede afirmar, con un 95% de confianza, que el salario medio de un ejecutivo antes de hacer el máster es menor que después de hacerlo.

Intervalo de confianza para el cociente de varianzas

Teniendo en cuenta la normalidad e independencia de ambas poblaciones, se verifica que

$$\left.\begin{array}{l} \dfrac{(n-1)\,\widehat{S}_X^2}{\sigma_X^2} \sim \chi^2_{n-1} \\ \dfrac{(m-1)\,\widehat{S}_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1} \end{array}\right\} \Rightarrow \frac{\dfrac{\widehat{S}_X^2}{\sigma_X^2}}{\dfrac{\widehat{S}_Y^2}{\sigma_Y^2}} = \frac{\widehat{S}_X^2}{\sigma_X^2}\frac{\sigma_Y^2}{\widehat{S}_Y^2} \sim F_{n-1,m-1}$$

IC para el cociente de varianzas

Por tanto, un I.C. a nivel $1-\alpha$ se obtiene de la siguiente manera

$$1-\alpha = \mathbb{P}\left(F_{n-1,m-1,1-\alpha/2} < \frac{\widehat{S}_X^2}{\sigma_X^2}\frac{\sigma_Y^2}{\widehat{S}_Y^2} < F_{n-1,m-1,\alpha/2} \right)$$

$$1-\alpha = \mathbb{P}\left(F_{n-1,m-1,1-\alpha/2}\frac{\widehat{S}_Y^2}{\widehat{S}_X^2} < \frac{\sigma_Y^2}{\sigma_X^2} < F_{n-1,m-1,\alpha/2}\frac{\widehat{S}_Y^2}{\widehat{S}_X^2} \right)$$

$$1-\alpha = \mathbb{P}\left(\frac{\frac{\widehat{S}_X^2}{\widehat{S}_Y^2}}{F_{n-1,m-1,1-\alpha/2}} > \frac{\sigma_X^2}{\sigma_Y^2} > \frac{\frac{\widehat{S}_X^2}{\widehat{S}_Y^2}}{F_{n-1,m-1,\alpha/2}} \right)$$

IC para el cociente de varianzas

IC para el cociente de varianzas al nivel $1-\alpha$

$$\boxed{\left(\frac{\dfrac{\widehat{s}_X^2}{\widehat{s}_Y^2}}{F_{n-1,m-1,\alpha/2}}, \frac{\dfrac{\widehat{s}_X^2}{\widehat{s}_Y^2}}{F_{n-1,m-1,1-\alpha/2}} \right)}$$

Nota

Con frecuencia se toma como extremo inferior del intervalo el 0. Así, el intervalo de confianza a nivel $1-\alpha$ es

$$\left(0, \frac{\dfrac{\widehat{s}_X^2}{\widehat{s}_Y^2}}{F_{n-1,m-1,1-\alpha}} \right)$$

Ejemplo

Siguiendo con el ejemplo de los alquileres en Madrid y en Barcelona, vamos a obtener un intervalo de confianza para el cociente de varianzas. Se obtenían los siguientes datos: $\overline{x} = 82.5$, $\overline{y} = 85.4$, $\widehat{s}_X^2 = 21.4^2$, $\widehat{s}_Y^2 = 24.3^2$, $n = 70$ y $m = 50$. Si $F_{69,49,0.975} = 0.5997$ y $F_{69,49,0.025} = 1.7072$, el intervalo de confianza para $\dfrac{\sigma_Y^2}{\sigma_Y^2}$ es el siguiente:

$$
\begin{aligned}
\left(\frac{\frac{\widehat{s}_X^2}{\widehat{s}_Y^2}}{F_{n-1,m-1,\alpha/2}}, \frac{\frac{\widehat{s}_X^2}{\widehat{s}_Y^2}}{F_{n-1,m-1,1-\alpha/2}} \right) &= \left(\frac{\frac{21.4^2}{24.3^2}}{1.7072}, \frac{\frac{21.4^2}{24.3^2}}{0.5997} \right) \\
&= \left(\frac{0.7756}{1.7072}, \frac{0.7756}{0.5997} \right) \\
&= (0.4543, 1.2933)
\end{aligned}
$$

UNIVERSIDADE DA CORUÑA

Intervalo de confianza para la diferencia de dos proporciones

La manera de proceder es idéntica a la del caso de una sola proporción.

- Se recogen sendas muestras aleatorias simples $X_1, \dots, X_n$ e $Y_1, \dots, Y_n$ de las variables X e Y donde

$$X_i = \begin{cases} 1 & \text{con probabilidad } p_X \\ 0 & \text{con probabilidad } 1 - p_X \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{con probabilidad } p_Y \\ 0 & \text{con probabilidad } 1 - p_Y \end{cases}$$

UNIVERSIDADE DA CORUÑA

Intervalo de confianza para la diferencia de dos proporciones

- Supuesta aceptable la aproximación normal, se tiene que

$$\left.\begin{array}{l} \dfrac{\widehat{p}_X - p_X}{\sqrt{\dfrac{p_X(1-p_X)}{n}}} \sim N(0,1) \\ \dfrac{\widehat{p}_Y - p_Y}{\sqrt{\dfrac{p_Y(1-p_Y)}{m}}} \sim N(0,1) \end{array}\right\}$$

$$\Rightarrow \quad \widehat{p}_X - \widehat{p}_Y \sim N\left(p_X - p_Y, \sqrt{\frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}}\right)$$

Intervalo de confianza para la diferencia de dos proporciones

- Mediante razonamientos análogos a los seguidos a lo largo del tema, se llega a que el intervalo de confianza para la diferencia de proporciones $p_X - p_Y$ con un nivel de confianza aproximado de $1-\alpha$ es

$$\boxed{\left((\widehat{p}_X - \widehat{p}_Y) \pm z_{1-\alpha/2} \sqrt{\frac{p_X\left(1-p_X\right)}{n} + \frac{p_Y\left(1-p_Y\right)}{m}} \right)}$$

- De nuevo, se ha de sustituir p_X y p_Y por sus estimaciones $\widehat{p}_X$ y $\widehat{p}_Y$ respectivamente.
- O bien, considerar la situación más desfavorable $p_x = p_y = 1/2$.

Ejemplo

Estamos ahora interesados en conocer si existen diferencias entre la proporción p_1 de hombres que fuman y la proporción p_2 de mujeres que fuman.

- Sea X la variable que vale 1 si un hombre fuma y vale 0 si no, e Y la variable que vale 1 si una mujer fuma y vale 0 si no.
- Si dentro de los 344 estudiantes había 160 hombres de los cuales 29 fumaban, y 184 mujeres de las cuales 54 fumaban,
- el intervalo de confianza para la diferencia de proporciones al 90% de confianza es el siguiente

$$\left((\widehat{p}_X - \widehat{p}_Y) \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{p}_X (1-\widehat{p}_X)}{n} + \frac{\widehat{p}_Y (1-\widehat{p}_Y)}{m}} \right)$$

$$= \left(\left(\frac{29}{160} - \frac{54}{184} \right) \pm 1.645 \sqrt{\frac{\frac{29}{160}\left(1-\frac{29}{160}\right)}{160} + \frac{\frac{54}{184}\left(1-\frac{54}{184}\right)}{184}} \right)$$

$$= (-0.1868, -0.0377)$$

Determinación del tamaño de muestra

Supuesto que eligiéramos muestras del mismo tamaño en ambas poblaciones ($m = n$), ¿cuál es el tamaño muestral necesario para que la longitud del intervalo de confianza para la diferencia de proporciones sea igual a una cantidad predeterminada, con un nivel de confianza $1 - \alpha$?.

UNIVERSIDADE DA CORUÑA

Determinación del tamaño de muestra

- Supuesto que el tamaño muestral a elegir es grande ($n > 30$), despejando n en la expresión para la longitud del intervalo de confianza para la diferencia de proporciones

$$L = 2z_{\alpha/2}\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{m}},$$

se obtiene

$$n = \frac{4z^2_{\alpha/2}[p_x(1-p_x) + p_y(1-p_y)]}{L^2}.$$

UNIVERSIDADE DA CORUÑA

Determinación del tamaño de muestra

- El problema del desconocimiento de p se sigue manteniendo, por lo que las dos soluciones apuntadas anteriormente siguen siendo válidas.
  - Si reemplazamos las proporciones poblacionales por las proporciones muestrales estimadas con muestras preliminares, el tamaño muestral requerido es
$$n = \frac{4z^2_{\alpha/2}[\hat{p}_x(1-\hat{p}_x)+\hat{p}_y(1-\hat{p}_y)]}{L^2}.$$
  - Si suponemos la situación más desfavorable $p_x = p_y = 1/2$, el tamaño muestral resultante es
$$n = \frac{2z^2_{\alpha/2}}{L^2}.$$

UNIVERSIDADE DA CORUÑA

Desigualdad de Tchebychev

Desigualdad de Tchebychev

Dada una variable aleatoria X, con media μ y desviación típica σ, el teorema o Desigualdad de Tchebychev afirma que para cualquier constante positiva k se verifica:

$$P\left(|X-\mu| \leq k\sigma\right) \geq 1 - \frac{1}{k^2}$$

UNIVERSIDADE DA CORUÑA

Desigualdad de Tchebychev

En paticular,

- Sea $\widehat{\theta} = T(X_1, ..., X_n)$ un estimador insesgado del parámetro θ, con desviación típica σ.
- Entonces, dado que $E\left(\widehat{\theta}\right) = \theta$, tomado $k = \frac{1}{\sqrt{\alpha}}$ (para que $1 - \alpha = 1 - \frac{1}{k^2}$),

$$P\left(|\widehat{\theta} - \mu| \leq k\sigma\right) \geq 1 - \frac{1}{k^2} \Rightarrow P\left(|\widehat{\theta} - \theta| \leq \frac{\sigma}{\sqrt{\alpha}}\right) \geq 1 - \alpha.$$

- El intervalo de confianza para θ sería

$$I.C._{1-\alpha}(\theta) = \left[\widehat{\theta} \pm \frac{\sigma}{\sqrt{\alpha}}\right].$$

Método Bootstrap

- Se puede determinar, de forma aproximada, la distribución del estadístico de interés, y por tanto los cuantiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$. La principal ventaja de este tipo de técnicas es su no "complejidad técnica" y como principal desventaja el habitual desconocimiento de la distribución de los datos.
- Sea $(X_1, ..., X_n)$ una m.a.s de la variable aleatoria X que tiene función de distribución F_θ. El método bootstrap aproxima la distribución del estadístico $T_{F_\theta}(X_1, ..., X_n)$ por la de $T_{F_n}(X_1^*, ..., X_n^*)$, siendo F_n la función de distribución empírica obtenida de la muestra $(X_1, ..., X_n)$, y $(X_1^*, ..., X_n^*)$ una muestra aleatoria simple de una v.a. con función de distribución F_n.

UNIVERSIDADE DA CORUÑA

Las etapas del método bootstrap son las siguientes:

1. Obtención de la función de distribución empírica F_n y simulación de la muestra $(X_1^*, ..., X_n^*)$.
2. Cálculo del estadístico con la muestra anterior: $T_{F_n}(X_1^*, ..., X_n^*)$
3. Repetición de las etapas anteriores B veces, para obtener los valores $T_{F_n}(X_1^*, ..., X_n^*)^{(i)} \quad i = 1, ..., B$
4. La distribución de $T_{F_\theta}(X_1, ..., X_n)$ se aproxima por la distribución empírica de los valores $T_{F_n}(X_1^*, ..., X_n^*)^{(i)} \quad i = 1,..., B$.

UNIVERSIDADE DA CORUÑA

Existen diferentes métodos de construcción de intervalos de confianza bootstrap:

- Método percentil
- Método percentil-t
- Método percentil$-t$ simetrizado

UNIVERSIDADE DA CORUÑA

Intervalos de confianza bayesianos

- Se considera que el parámetro de interés θ es una variable aleatoria con una distribución a priori $\pi(\theta)$.
- Supongamos que la distribución a priori de θ es continua y sea entonces $\pi(\theta)$ su función de densidad.
- Entonces, aplicando el teorema de Bayes, se tiene que la distribución a posteriori de θ, condicionada a la muestra $(X_1, \ldots, X_n)$, tiene función de densidad

$$\pi(\theta_0 | x_1, \ldots, x_n) = \frac{\pi(\theta_0) f_{\theta_0}(x_1, \ldots, x_n)}{\int_\Theta \pi(\theta) f_\theta(x_1, \ldots, x_n)\, d\theta}.$$

Intervalos de confianza bayesianos

- El denominador de $\pi\left(\theta_0|x_{1,}...,x_n\right)$ es independiente de θ_0, de modo que la función de densidad a posteriori de la variable aleatoria θ es proporcional a $\pi\left(\theta_0\right) f_{\theta_0}\left(x_1,...,x_n\right)$:

$$\pi\left(\theta|x_{1,}...,x_n\right) \propto \pi\left(\theta\right) f_\theta\left(x_1,...,x_n\right).$$

- El objetivo es, fijado α, determinar los valores tales que

$$\begin{aligned} & P\left(T_1\left(X_1,...,X_n\right) \leq \theta \leq T_2\left(X_1,...,X_n\right)|X_1,...,X_n\right) \\ = & \int_{T_1(X_1,...,X_n)}^{T_2(X_1,...,X_n)} \pi\left(\theta|X_1,...,X_n\right) d\theta = 1-\alpha \end{aligned}$$

Intervalos de confianza bayesianos

- Es decir, el intervalo bayesiano para θ con probabilidad $1-\alpha$ es

$$I.C._{1-\alpha}(\theta) = \left[x_{(1-\alpha)/2}, x_{\alpha/2}\right]$$

siendo $x_{\alpha/2}$ el cuantil $(1-\alpha)$ que deja a la derecha una probabilidad α de $\pi(\theta|X_1, ..., X_n)$.

UNIVERSIDADE DA CORUÑA