INTERVALOS DE CONFIANZA

La estadística en cómic
(L. Gonick y W. Smith)

EJEMPLO: ¿Será elegido el senador Astuto?

tamaño muestral

Estimador de p

variable aleatoria poblacional

$$X = \begin{cases} 1 & \texttt{si el encuestado votará a Astuto} \\ 0 & \texttt{si el encuestado no votará a Astuto} \end{cases}$$

?

$$X \rightsquigarrow B(1, p)$$ ← proporción de personas que votarán a Astuto


¿SÓLO HA PREGUNTADO A 1.000 PERSONAS? PERO ¡¡SI HAY UN MILLÓN DE VOTANTES EN ESTE ESTADO!!
ME DA IGUAL QUE SEAN UN MILLÓN O UN BILLÓN...
¡HE HECHO UN MUESTREO ALEATORIO SIMPLE! ¡ASÍ QUE LE PUEDO OFRECER UNA FIRME GARANTÍA!
Y ESA GARANTÍA... ES DEL TIPO TOTAL Y CON REEMBOLSO, ¿VERDAD?
¿EH?
EH... BUENO... LO CIERTO ES QUE ES MÁS BIEN ESTADÍSTICA...
SÍ, PUEDO AFIRMAR CON UN 95% DE SEGURIDAD QUE LA VERDADERA PROPORCIÓN DE LA POBLACIÓN, p, ESTÁ ENTRE 0,519 Y 0,581.

Intervalo de confianza para *p* con un coeficiente de confianza del 95%

$$I = [0.519,0.581]$$

CUANDO ASTUTO SE CALMA, HOLMES LE EXPLICA LO QUE QUIERE DECIR CON **UN 95% DE SEGURIDAD**: SABE QUE SU PROCEDIMIENTO DE ESTIMACIÓN TIENE UNA **PROBABILIDAD DE UN 95%** DE PRODUCIR UN INTERVALO EN EL QUE SE ENCUENTRE *p*.


VAMOS A CONSIDERAR A UNA ARQUERA QUE DISPARA A UNA DIANA. SUPONGAMOS QUE DA EN EL BLANCO DE 10 CENTÍMETROS UN 95% DE LAS VECES QUE DISPARA. ES DECIR, SÓLO UNA FLECHA DE CADA 20 NO DA EN EL BLANCO.


Y DETRÁS DE LA DIANA ENCONTRAMOS A UN VALEROSO DETECTIVE, QUE NO VE EL BLANCO. LA ARQUERA DISPARA UNA SOLA FLECHA.


EL DETECTIVE CONOCE EL NIVEL DE HABILIDAD DE LA ARQUERA Y DIBUJA UN CÍRCULO CON RADIO DE 10 CENTÍMETROS ALREDEDOR DE LA FLECHA. ¡TIENE UNA **SEGURIDAD DE UN 95%** DE QUE EN ESE CÍRCULO SE ENCUENTRA EL CENTRO DE LA DIANA!
10 CM
HA RAZONADO QUE SI DIBUJABA CÍRCULOS DE 10 CENTÍMETROS DE RADIO ALREDEDOR DE **MUCHAS FLECHAS**, EL BLANCO SE ENCONTRARÍA DENTRO DE ESOS CÍRCULOS UN 95% DE LAS VECES.
CENTRO AUTÉNTICO
X = FLECHA

DEFINICIÓN DE INTERVALO DE CONFIANZA

Sea X una variable aleatoria cuya distribución depende de un parámetro θ, y sea $(X_1,\ldots, X_n)$ una muestra aleatoria simple de X.

Si $T_1(X_1,\ldots, X_n)$ y $T_2(X_1,\ldots, X_n)$ son dos estimadores tales que

número desconocido

$$\boxed{p(T_1(X_1,\ldots,X_n)\leq \theta \leq T_2(X_1,\ldots,X_n)) = 1-\alpha \qquad \forall\theta\in\Theta}$$

variables aleatorias

números

al intervalo $I = [T_1(x_1,\ldots, x_n) , T_2(x_1,\ldots, x_n)]$ se le llama intervalo de confianza para θ de coeficiente de confianza 1-α.

Subconjunto del espacio paramétrico

<u>Interpretación:</u> De los distintos intervalos numéricos construidos a partir de sucesivos muestreos, un porcentaje del (1-α)100% contiene al verdadero valor del parámetro desconocido θ.

<u>Ejemplo:</u> Hallar un intervalo de confianza para la media, de coeficiente de confianza 1-α, de una población normal con varianza conocida.

$$p(\underbrace{T_1(X_1,\ldots,X_n)}_{\bar{X}-c\ ?} \leq \mu \leq \underbrace{T_2(X_1,\ldots,X_n)}_{\bar{X}+c\ ?}) = 1-\alpha$$

Estimador de máxima verosimilitud de μ

queremos que la probabilidad que el intervalo no cubre se reparta en dos colas iguales

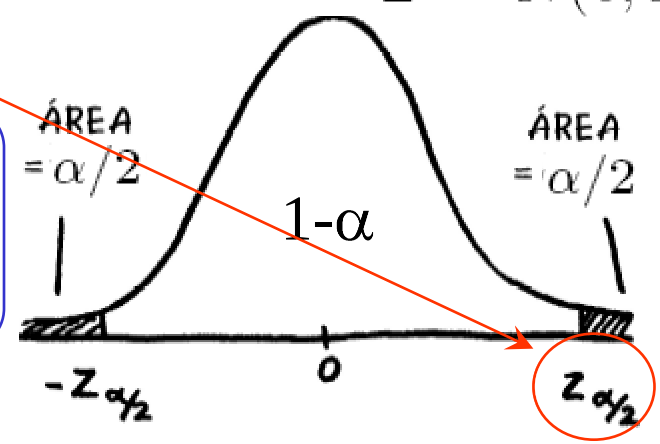
Buscamos un número *c* tal que

$$p(\bar{X} - c \leq \mu \leq \bar{X} + c) = 1 - \alpha$$

$$\Updownarrow \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

$$p\left(\frac{-c}{\sigma/\sqrt{n}} \leq \; Z \leq \frac{c}{\sigma/\sqrt{n}} \right) = 1 - \alpha$$

$$Z \rightsquigarrow N(0,1)$$



$$\Updownarrow$$

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

Notación:

$$p(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

$$\Updownarrow$$

$$c = z_{\alpha/2} \; \frac{\sigma}{\sqrt{n}} \Leftrightarrow \boxed{I = \left[\bar{x} - z_{\alpha/2} \; \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \; \frac{\sigma}{\sqrt{n}} \right]}$$

Definición: Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una caracteristicaX de una población con función de masa $P_\theta(x)$ (caso discreto), o con función de densidad $f_\theta(x)$(caso continuo) donde $\theta \in \Theta$ es desconocido. Una Cantidad pivotal $C(X_1, X_2, ..., X_n, \theta)$ es una función cuya distribución no depende del parámetro θ

•En el resto del tema se ven otras cantidades pivotales que generan otros intervalos de confianza

La lista de intervalos de confianza se encuentra en la pagina Web

Clave de la construcción de Intervalos de Confianza:
Cantidades pivotales (funciones de la muestra aleatoria
Que no dependen del parametro)

$$X \mapsto N(\mu, \sigma), \sigma \text{ conocida}$$

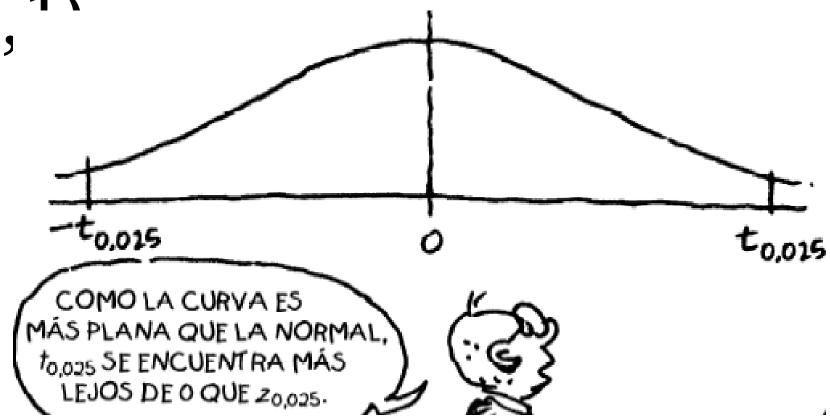
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1) \qquad I = \left[\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$X \mapsto N(\mu, \sigma), \sigma \text{ no conocida}$$

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \rightsquigarrow t_{n-1} \quad I = \left[\bar{x} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] \qquad \texttt{si} \;\; n \leq 30$$

La distribución t de student

$$Y \mapsto X_1 \mapsto X_i \mapsto X_n \mapsto N(0,1)$$



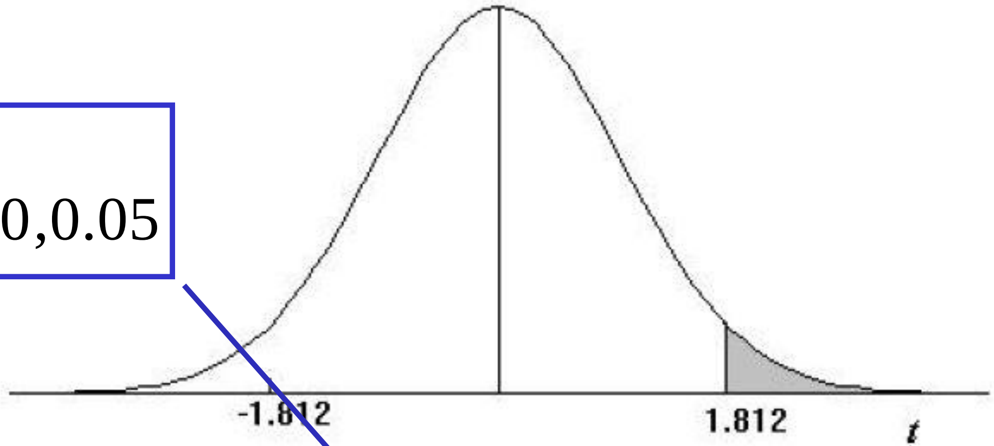
$$t_n \mapsto \frac{Y}{\sqrt{\sum_{i=1}^{n} X_i^2}}$$

La distribución t de student se trabaja con su correspondiente tabla

TABLA 2: DISTRIBUCIÓN t DE STUDENT

Puntos de porcentaje de la distribución t

$t_{10,0.05}$



Ejemplo

Para $\phi = 10$ grados de libertad:

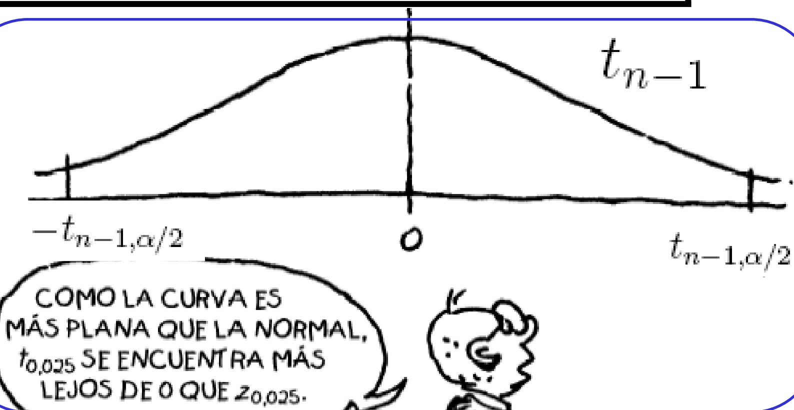
$P[\, t > 1.812] = 0.05$
$P[\, t < -1.812] = 0.05$

α / r	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318

Intervalo de confianza al nivel 1-α para la μ de una población normal

(A) Si es conocida:

$$I = \left[\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$



(B) Si σ es desconocida:

$$p(t_{n-1} \geq t_{n-1,\alpha/2}) = \frac{\alpha}{2}$$

$$I = \left[\bar{x} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] \qquad \texttt{si } n \leq 30$$

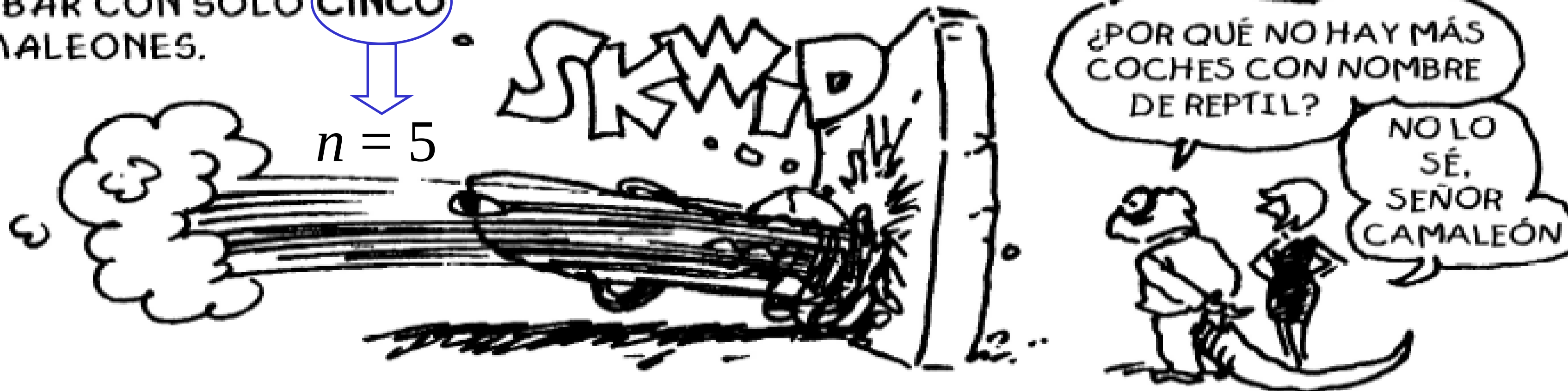
$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \rightsquigarrow t_{n-1}$$

$$I = \left[\bar{x} \mp z_{\alpha/2}\frac{s}{\sqrt{n}}\right] \qquad \texttt{si } n > 30$$

$t_{n-1} \approx N(0,1)$
si n es grande

Ejemplo: Supongamos que la Camaleón Motors tiene que hacer pruebas de choque con sus coches para determinar el coste medio de reparación tras una colisión frontal a unos 20 kilómetros por hora. ¡Resulta muy caro! Así que deciden probar con sólo cinco camaleones.

$n = 5$



Los datos de los desperfectos en dólares son 150, 400, 720, 500 y 930.

¿Dónde podemos situar la media con una confianza del 95%?

NOTA: ADMITIR NORMALIDAD EN LOS DATOS

$\bar{x}$ = 540 DÓLARES s = 299 DÓLARES

$1-\alpha = 0.95$

$\alpha = 0.05$

$$\sqrt{\frac{1}{4}\left((150-540)^2+(400-540)^2+(720-540)^2+(500-540)^2+(930-540)^2\right)}$$

$$I = \left[\bar{x} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] = \left[540 \mp 2.78\,\frac{299}{\sqrt{5}}\right]$$

$$t_{4,0.025} = 2.78$$

$$\boxed{I = [168, 912]}$$

ASÍ QUE LO MEJOR QUE PODEMOS DECIR CON UN 95% DE CONFIANZA ES QUE EL COSTE MEDIO POR REPARACIÓN DE DAÑOS ESTARÁ ENTRE 168 Y 912 DÓLARES.

PARA CALCULAR ESTE INTERVALO DE CONFIANZA UTILIZANDO LA *t* DE STUDENT, HEMOS REALIZADO UNA **PRESUNCIÓN NO COMPROBADA**: HEMOS DADO POR HECHO QUE LOS COSTES DE REPARACIÓN DE LOS COCHES TIENEN UNA **DISTRIBUCIÓN APROXIMADAMENTE NORMAL**, ES DECIR, QUE SI HICIÉSEMOS CHOCAR 1.000 CAMALEONES, EL HISTOGRAMA DE LOS COSTES DE REPARACIÓN SERÍA SIMÉTRICO Y CON FORMA DE MONTAÑA. **NO PODEMOS SABERLO** CON TAN SÓLO CINCO DATOS... PERO QUIZÁ AÑOS DE EXPERIENCIA CON ANTIGUOS MODELOS HAN PRODUCIDO HISTOGRAMAS DE LOS COSTES DE REPARACIÓN DE LA PARTE FRONTAL DEL COCHE CON DISTRIBUCIÓN NORMAL: UNA INFORMACIÓN QUE APOYARÍA NUESTRA DECISIÓN DE UTILIZAR LA *t* DE STUDENT.

Intervalo de confianza al nivel 1-α para la μ de una población no normal con muestras grandes (σ finita)

(A) Si σ es conocida:

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

$$I = \left[\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \quad \texttt{si} \;\; n > 30$$

(B) Si σ es desconocida:

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \approx N(0,1)$$

$$I = \left[\bar{x} \mp z_{\alpha/2}\frac{s}{\sqrt{n}}\right] \quad \texttt{si} \;\; n > 100$$

Intervalo de confianza al 1-α para el parámetro *p* de una binomial

$$I = \left[\widehat{p} \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

$$\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \approx N(0,1)$$

si $n > 100$

Intervalo de confianza al 1-α para el parámetro λ de una Poisson

$$I = \left[\widehat{\lambda} \mp z_{\alpha/2}\sqrt{\frac{\widehat{\lambda}}{n}}\right]$$

$$\frac{\widehat{\lambda}-\lambda}{\sqrt{\widehat{\lambda}/n}} \approx N(0,1)$$

si $n > 100$

EJEMPLO: Después de extraer una muestra aletoria simple de tamaño 1000, Holmes observó que 550 personas pensaban votar al senador Astuto. ¿Podemos afirmar con un confianza del 99% que Astuto será reelegido?

$n = 1000$

$1-\alpha = 0.99$

$\alpha = 0.01$

$\alpha/2 = 0.005$

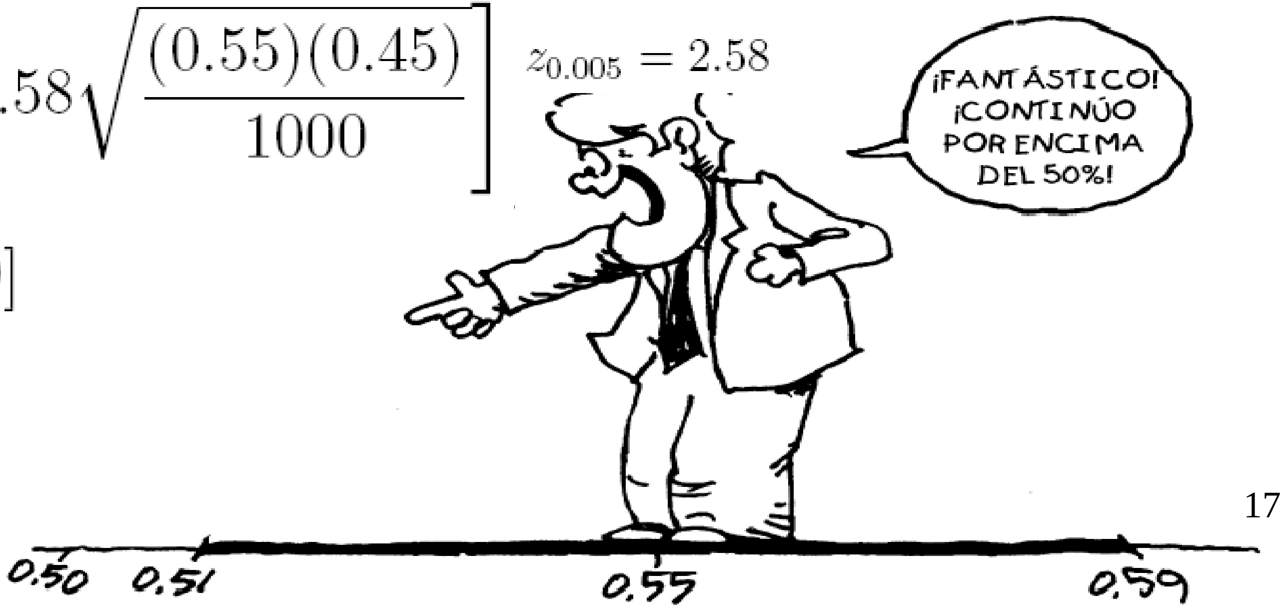
$z_{0.005} = 2.58$

$\widehat{p} = 0.55$

$$I = \left[\widehat{p} \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

$$= \left[0.55 \mp 2.58\sqrt{\frac{(0.55)(0.45)}{1000}}\right]$$

$$= [0.51, 0.59]$$



<u>EJEMPLO</u>: Admitiendo que el número de erratas por página de cierto libro sigue una distribución de Poisson, determinar un intervalo de confianza al 95% del número medio de erratas por página que contiene dicho libro, teniendo en cuenta que se eligieron al azar y con reemplazamiento 100 páginas en las que se observó una media muestral de 0.04 erratas por página.

$$\alpha = 0.05 \qquad \widehat{\lambda} = 0.04 \qquad n = 100$$

$$z_{\alpha/2} = 1.96$$

número medio de erratas por página

$$X = \texttt{número de erratas por página} \rightsquigarrow \mathcal{P}(\lambda)$$

$$I = \left[\widehat{\lambda} \mp z_{\alpha/2}\sqrt{\frac{\widehat{\lambda}}{n}}\right] = \left[0.04 \mp 1.96\sqrt{\frac{0.04}{100}}\right]$$

$$= \boxed{[0.0008, 0.0792]}$$

<u>Observación</u>: Cuanto más corto sea el intervalo de confianza más precisa es la estimación que proporciona, pero, al disminuir la longitud del intervalo, si mantenemos fijo el tamaño muestral, también disminuye el coeficiente de confianza.

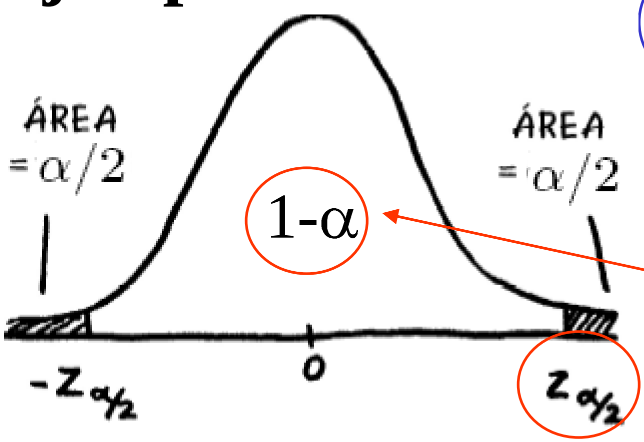
Ejemplo anterior:

longitud del intervalo

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

tamaño muestral

1-α es el coeficiente de confianza



disminuye al disminuir la longitud del intervalo

¿Cómo podemos mejorar la estimación?

Aumentando el tamaño muestral, ya que entonces la longitud del intervalo disminuye.

MÍNIMO TAMAÑO MUESTRAL

P: ¿Cuál es el tamaño muestral para fijar el error E con un intervalo De confianza α?

Caso 1: $N(\mu,\sigma)$, con σ conocida

$$z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq E \Leftrightarrow n \geq (z_{\frac{\alpha}{2}}\frac{\sigma}{E})^2$$

¿Cual es el mínimo tamaño muestral necesario para obtener una precisión dada?

EJEMPLO: Supongamos que la altura de los individuos de cierta población sigue una distribución $N(\mu, 7.5)$. Hallar el mínimo tamaño muestral necesario para estimar la altura media con un error inferior a 2 y con una confianza del 90%.

$X \rightsquigarrow N(\mu, 7.5)$

altura de un individuo

$\bar{x}$ es el estimador de μ

$$I = \left[\bar{x} \mp z_{0.05}\frac{\sigma}{\sqrt{n}}\right] = \left[\bar{x} \mp 1.645\frac{7.5}{\sqrt{n}}\right]$$

Intervalo de confianza para μ al 90%

Con una confianza del 90% $\mu \in I$

$$\Rightarrow |\bar{x} - \mu| \leq 1.645\frac{7.5}{\sqrt{n}}$$

$$|\bar{x} - \mu| < 2$$

$$\Rightarrow 1.645\frac{7.5}{\sqrt{n}} < 2 \Leftrightarrow n > 38.05 \Rightarrow \boxed{n = 39}$$

Caso 2: N(μ,σ), con σ desconocida

$$t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \le E \Leftrightarrow ?????$$

$$\boxed{z_{\frac{\alpha}{2}}}$$

Se calcula usando una pequeña muestra piloto

Ejemplo:

Ejemplo En un estudio sobre el tiempo de desarrollo de una Especie de insectos se escogió una muestra modelo de 13 Individuos. La media fue de 4 horas y la cuasidesviación típica De 3. Asumiendo normalidad, ¿Cuántos individuos habrá Que observar para estimar la media μ con un error inferiór a 0.2 Y un nivel de confianza del 0.95?

$$1,96 * 3 \leq \sqrt{n} * 0.2 \Rightarrow n \geq 864,36$$

Intervalo de confianza al nivel 1-α para la varianza de una población normal

(A) Si μ es desconocida:

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

$$I = \left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} , \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right]$$

(B) Si μ es conocida:

Notación:
$p(\chi^2_{n-1} \geq \chi^2_{n-1,\alpha/2}) = \frac{\alpha}{2}$

$$I = \left[\frac{nT}{\chi^2_{n,\alpha/2}} , \frac{nT}{\chi^2_{n,1-\alpha/2}} \right]$$

$$\frac{nT}{\sigma^2} \rightsquigarrow \chi^2_n$$

donde $T = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

Intervalo de confianza al nivel 1-α para el cociente de las varianzas de dos poblaciones normales independientes

(A) Si μ_1 y μ_2 son desconocidas:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F_{n_1-1,n_2-1}$$

$$I = \left[\frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right]$$

(B) Si μ_1 y μ_2 son conocidas:

Notación:
$p(F_{n_1,n_2} \geq F_{n_1,n_2,\alpha/2}) = \frac{\alpha}{2}$

$$I = \left[\frac{T_1/T_2}{F_{n_1,n_2,\alpha/2}}, \frac{T_1/T_2}{F_{n_1,n_2,1-\alpha/2}}\right] \text{donde}$$

$$\frac{T_1/\sigma_1^2}{T_2/\sigma_2^2} \rightsquigarrow F_{n_1,n_2}$$

$$T_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\mu_1)^2 \text{ y } T_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i-\mu_2)^2$$

Intervalo de confianza al nivel 1-α para la diferencia de medias de dos poblaciones normales independientes

(A) Si σ_1 y σ_2 son conocidas:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0,1)$$

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}$$

(B) Si $\sigma_1 = \sigma_2$ desconocidas.

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp t_{n_1+n_2-2,\alpha/2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

(C) Si $\sigma_1 \neq \sigma_2$ desconocidas:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_f$$

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp t_{f,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

donde f es el entero más próximo a

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1+1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2+1}} - 2$$

Intervalo de confianza al nivel 1-α para la diferencia de la medias de dos poblaciones independientes no necesariamente normales

(A) Si σ_1 y σ_2 son conocidas:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right] \quad \texttt{si} \;\; n_1, n_2 > 15$$

(B) Si σ_1 y σ_2 son desconocidas:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1)$$

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right] \quad \texttt{si} \;\; n_1, n_2 > 50$$