Intervalos de Confianza

Introducción a la Estadística

Fiona Franco Churruarín
fionafch96@gmail.com

UTDT

Febrero 2022

Motivación

Hasta ahora estudiamos cómo dar una respuesta a las preguntas: ¿cuál es su adivinanza *puntual* para el valor de la media poblacional? ¿por qué usó ese procedimiento?

Algo que no cuantificamos es que tan inseguros estamos de nuestra adivinanza. Como todo lo que observamos es una muestra, y estamos haciendo una afirmación sobre la población, no podemos estar 100% seguros que nuestra adivinanza esté cerca del valor verdadero.

Si volvemos a la pregunta de arriba, nuestra respuesta hasta acá sería algo como: "La media muestral, 1.65, porque es un estimador insesgado y eficiente".

Motivación

Lo que nos gustaría agregar es una oración como:

- "Si te tengo que decir un número, te diría 1.65, pero la verdad no tengo idea"; o (en el mejor de los casos)
- "Si me preguntás por un solo número, te digo 1.65, y si la media poblacional no es 1.65 pega en el palo"

¿Por qué podría interesarnos esto? Para saber si nuestra estimación, siguiendo el hilo 1.65, nos dió 1.65 de pura suerte, o si en repetidas ocasiones de muestreo y estimación nos daría 1.65 y números cercanos. Proveer esta información nos ayuda a establecer **que tan confiable** es nuestra estimación.

Ejemplo

Considere dos personas que intentan estimar el mismo parámetro, ambas utilizan el mismo estimador (la media muestral), pero obtienen distintas muestras:

- Persona 1, dos observaciones: 1.60, 1.80 $\rightarrow$ estimación: 1.70
- Persona 2, 30 observaciones: 8 observaciones de 1.68, 14 observaciones de 1.70, y 8 observaciones de 1.72 $\rightarrow$ estimación: 1.70

Si sólo un número fuera solicitado a ambas personas, las dos dirían que la su estimación para la media poblacional es 1.70, pero la verdad que la segunda persona estaría mucho mas segura por dos motivos:

1. obtuvo un mayor tamaño de muestra; y
2. los datos de su muestra estan menos dispersos alrededor de 1.70 que los datos de la primera.

Intervalos de Confianza

En muchos casos se recurre a la herramienta conocida como **intervalos de confianza** para cuantificar el margen de error que hay en nuestro cómputo.

Esencialmente se reporta una estimación puntual (el 1.70), un nivel de confianza, y un rango de valores posibles para el parámetro poblacional acorde a un dado nivel de confianza.

Suficiente motivación... allá vamos.

IC - Construcción

Recordemos que la media muestral es una variable aleatoria $\bar{X}$ con distribución exacta o aproximada normal, con media μ y varianza $\frac{\sigma^2}{n}$.

Entonces utilizando un valor α entre 0 y 1 (que luego elegiremos con algún criterio) podemos decir lo siguiente:

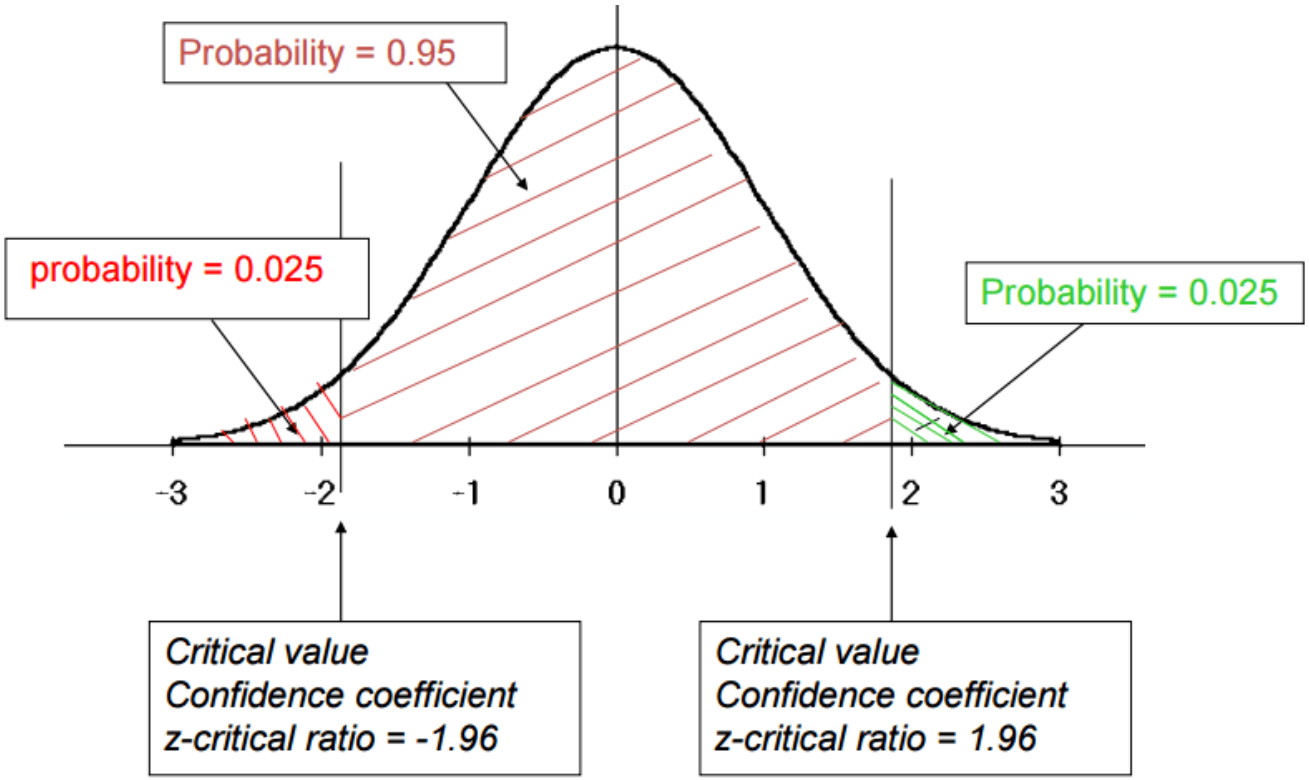
$$P(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\frac{\alpha}{2}}) = 1 - \alpha \tag{1}$$

Donde $z_{\frac{\alpha}{2}}$ es el valor que verifica, para una variable aleatoria $Z \sim N(0,1)$:

$$P(Z > z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

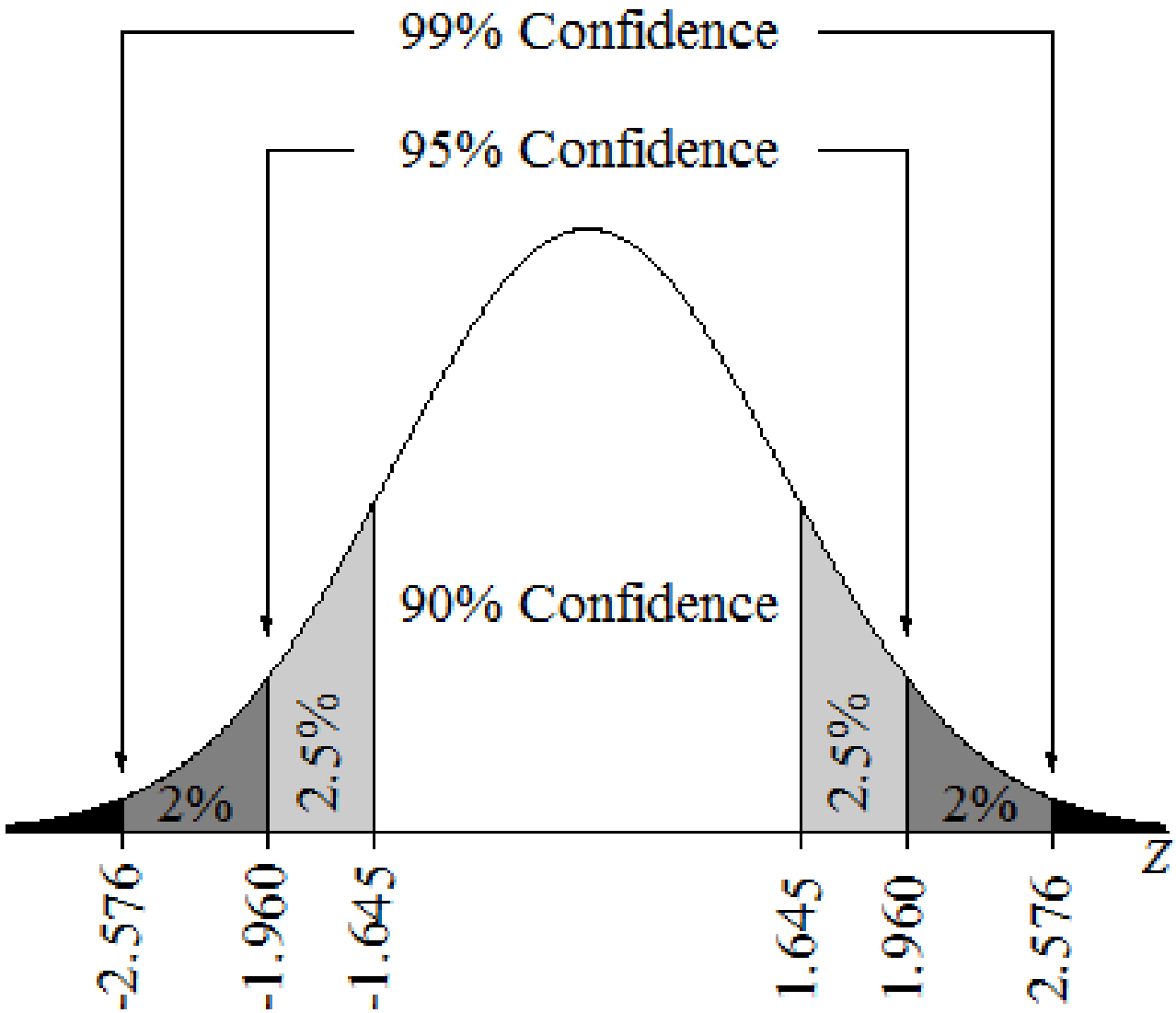
IC - Construcción

Gráficamente:



IC - Construcción

Los niveles de confianza más utilizados son 90%, 95% y 99%:



IC - Construcción

Notar que de la expresión 1 podemos escribir la siguiente:

$$P(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha \tag{2}$$

Así, suele denotarse:

$$IC_{\mu}^{(1-\alpha)\%} = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

IC - Interpretación

Notemos donde está la variable aleatoria en la expresión (2), que usamos para construir el intervalo de confianza:

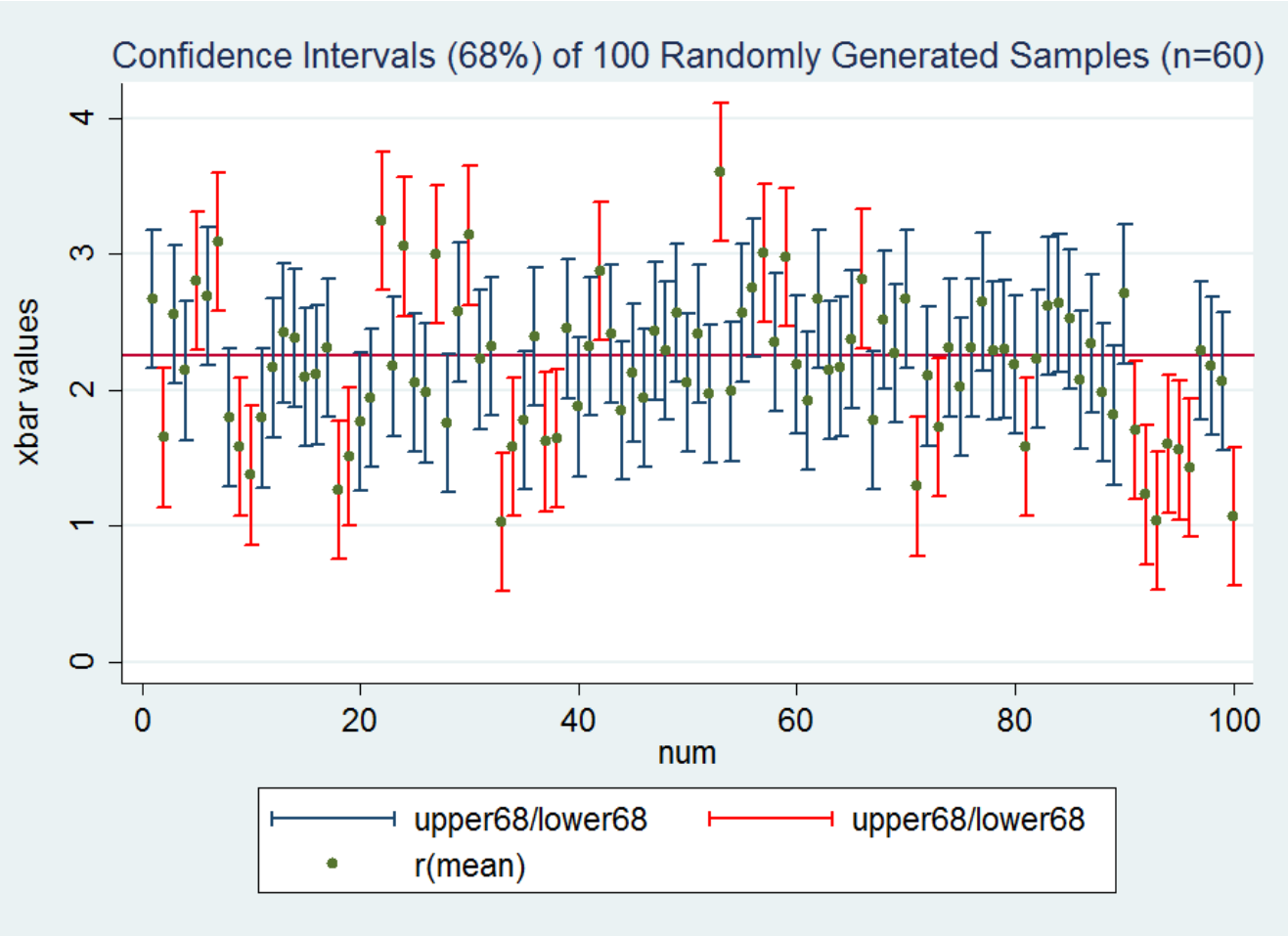
$$P(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

¿Que nos dice y que no nos dice esto?

- NO nos dice que la probabilidad de que la media poblacional μ esté dentro del intervalo es $1-\alpha$.
- SI nos dice que la **confiabilidad** del intervalo es de $(1-\alpha)\%$.

La **confiabilidad** nos habla de que ocurre si repetimos este procedimiento muchísimas veces: $(1-\alpha)\%$ de las veces el intervalo contendrá al parámetro verdadero.

IC - Interpretación



De 100 muestras, con un nivel de confianza de 68%, alrededor de 68 (70 exactamente) contienen a la media verdadera.

Ejemplo

Se conoce, de estudios anteriores, que el costo variable de construcción de determinado tipo de vivienda prefabricada, por metro cuadrado, se distribuye normalmente con un desvío estándar de $135. Se tomó una muestra aleatoria de 12 viviendas con las que se calculó un costo variable promedio de $1440.

Provea un intervalo de confianza al 95% para la media del costo de construcción por metro cuadrado para este tipo de viviendas.

¿Cuál es la probabilidad de que la media poblacional del costo variable de construcción por metro cuadrado esté contenido en el intervalo que calculó?

Varianza poblacional desconocida

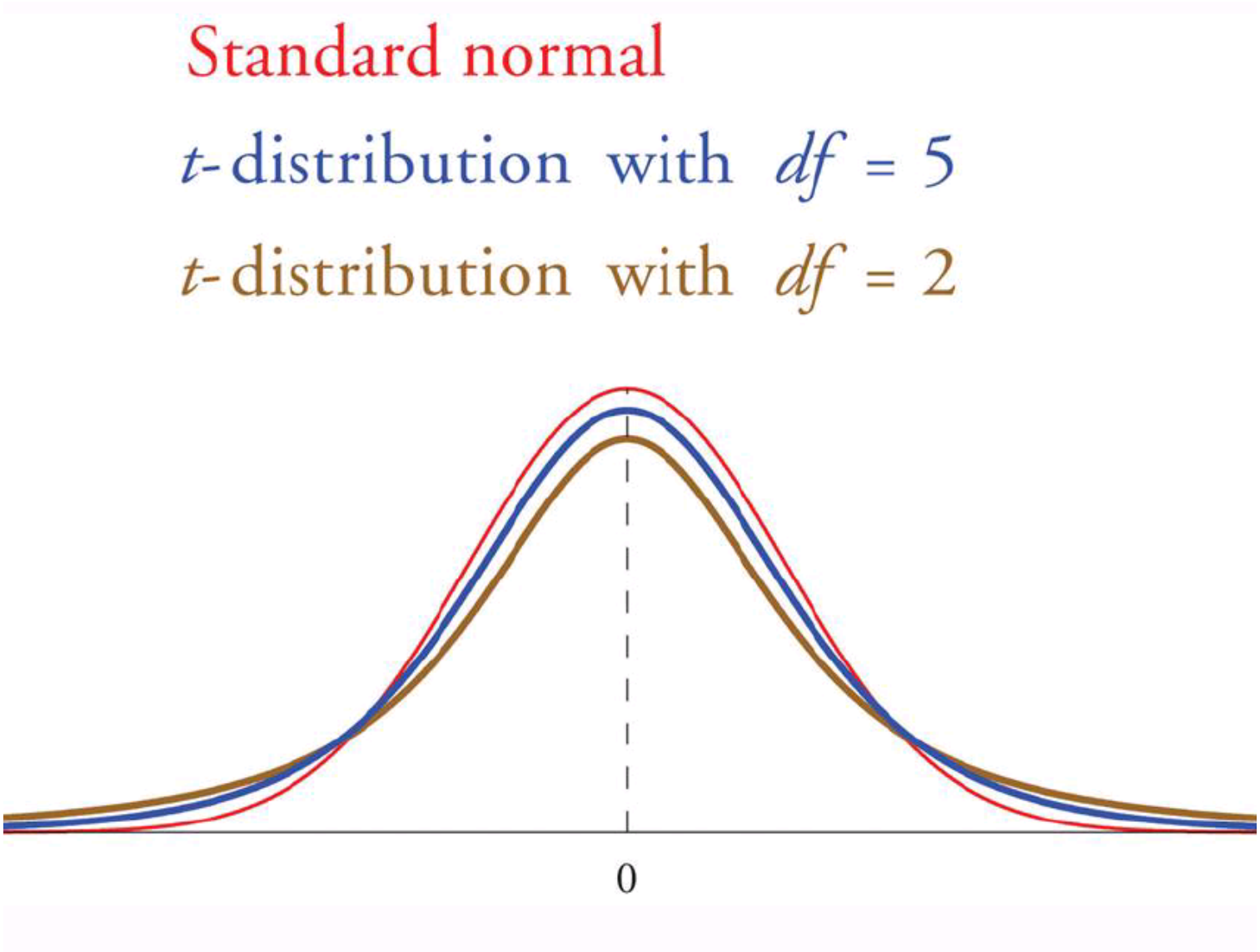
Hasta aquí hicimos un supuesto implícito que es conocer el valor σ, que típicamente es desconocido.

Por el hecho de no conocer σ^2 y estimarla usando S^2, entonces:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

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

Distribución t de Student



Varianza poblacional desconocida

De manera análoga al caso con σ conocida,

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

Así,

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

$$IC_{\mu}^{(1-\alpha)\%} = \left[\bar{X} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}} \quad , \quad \bar{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$$

Ejemplo

En un estudio de los costos anuales de alquiler de departamentos en una ciudad del este, en base a una muestra aleatoria de 36 departamentos se estiman:

- un costo promedio de alquiler de 11.535 dólares por año; y
- una desviación estándar de 875 dólares al año.

Construya un intervalo de confianza del 99% para la media del costo anual de alquiler de departamentos.

Para pensar

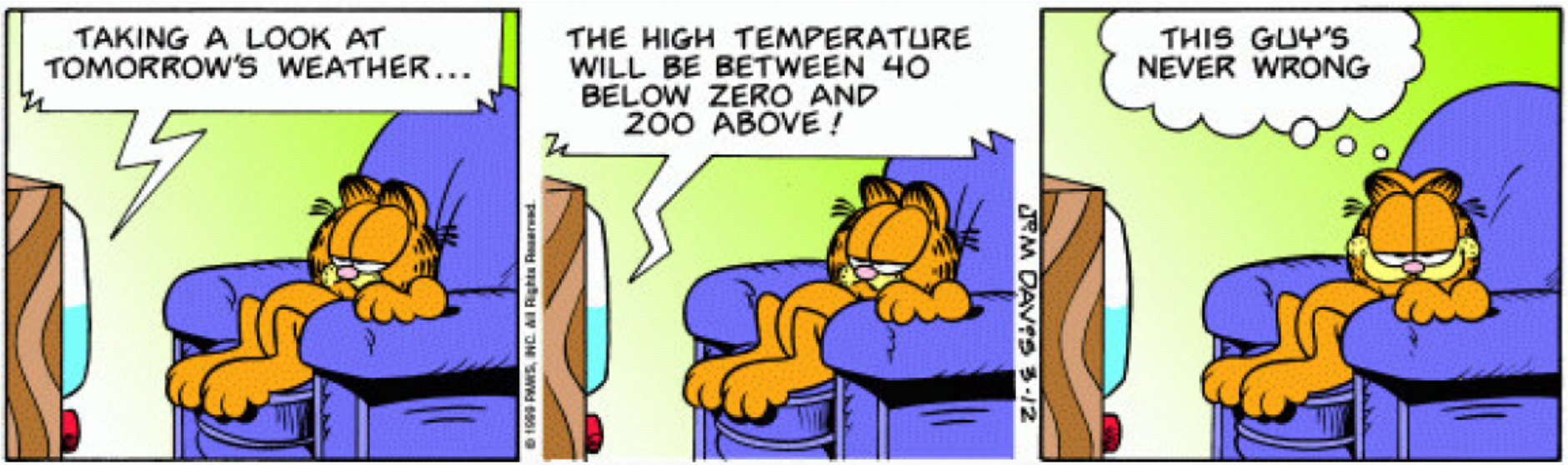
¿Qué pasará con el **margen de error** o la **amplitud del intervalo** (aumentará o disminuirá) si...

- aumento el tamaño de la muestra (n)?
- tomo otra muestra de igual tamaño, pero con mayor dispersión (S)?
- trabajo con un mayor nivel de confianza ($1-\alpha$)?

$$amplitud = \Delta(n, S, \alpha) = 2z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

Para pensar

¿Cómo construyo un intervalo de confianza con 100% de confianza?



Trade-off entre 'utilidad' del intervalo y su confiabilidad.

Intervalos de confianza para la varianza (σ^2)

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

Ejemplo: una muestra aleatoria de 15 píldoras para el dolor de cabeza mostró un desvío estándar de 0.8% en la concentración del ingrediente activo. Encontrar un intervalo de confianza del 90% para la varianza poblacional de estas píldoras.

Como $\alpha = 0.10$, a partir de la tabla de la chi-cuadrado:

$$\chi^2_{n-1,\alpha/2} = \chi^2_{14,0.05} = 23.68 \text{ y } \chi^2_{n-1,1-\alpha/2} = \chi^2_{14,0.95} = 6.57.$$

Entonces el intervalo de confianza para la varianza poblacional es:

$$IC^{90\%}_{\sigma^2} = [0.378, 1.364]$$

Intervalo de confianza para la proporción (p)

Sean $X_1, X_2, \ldots, X_n$ una muestra i.i.d. a partir de una distribución Bernoulli, $B(1,p)$. Entonces, $Y = \sum_{i=1}^{n}$ tiene una **distribución binomial**, $B(n,p)$.

Sabemos que:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}Y$$

Al seguir una distribución binomial, también sabemos que:

$$Var(Y) = np(1-p) \Rightarrow Var(\hat{p}) = \frac{p(1-p)}{n}$$

La cual se puede estimar con:

$$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n}$$

Intervalo de confianza para la proporción (p)

Si:

- n es una muestra lo suficientemente grande
- p no es extremadamente alto o pequeño

Entonces,

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{aprox.}{\sim} N(0,1)$$

- Referencia recomendada: `Quincunx - Galton Board`

Intervalo de confianza para la proporción (p)

Entonces,

$$P\left(\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\right) = 1-\alpha$$

$$IC_{p}^{(1-\alpha)\%} = \left[\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \quad , \quad \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\right]$$

Ejemplo

Una encuesta con respecto a la política de jubilaciones, reveló que una alta proporción de los ciudadanos es muy pesimista con respecto a sus perspectivas al momento de jubilarse. Al ser preguntados si consideran que su jubilación será suficiente, 60% de los 6100 trabajadores entrevistados indicaron que no sería suficiente. Calcular el intervalo de confianza para la proporción de todos los trabajadores de 18 años o más que consideran que, al jubilarse, su ingreso no será suficiente (asuma un nivel de confianza del 95%).

Determinando el tamaño muestral para μ

Queremos fijar la amplitud del IC y obtener el n necesario para alcanzar dicha amplitud. Sea Δ la amplitud del IC:

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

Supongamos que conocemos σ, entonces:

$$n \geq \left[2\frac{z_{\alpha/2}\sigma}{\Delta}\right]^2$$

Es decir, el tamaño muestral **mínimo** es:

$$n^* = \left[2\frac{z_{\alpha/2}\sigma}{\Delta}\right]^2$$

n^* asegura que el intervalo de confianza $1-\alpha$ de μ tenga a lo sumo una longitud de Δ.

Determinando el tamaño muestral para μ

¿Cómo hacemos para conocer σ^2?

En general no lo conocemos, sino que lo estimamos:

- pruebas pilotos
- experimentos pasados

En la práctica se utiliza la misma fórmula ya sea conocida o no σ^2.

Ejemplo

Un call center está interesado en determinar la duración esperada de una llamada telefónica de la forma más precisa posible. Los requerimientos son que el intervalo al 95% de la duración media de la llamada debería tener una amplitud de 1 minuto. Supongan que en el call center se llevó a cabo una prueba piloto a partir de la cual se estimó que la dispersión de la duración de las llamadas es de 5 minutos. ¿Cuál es el tamaño muestral que se requiere para estimar la duración esperada de las llamadas con la precisión deseada?

Tamaño muestral para p

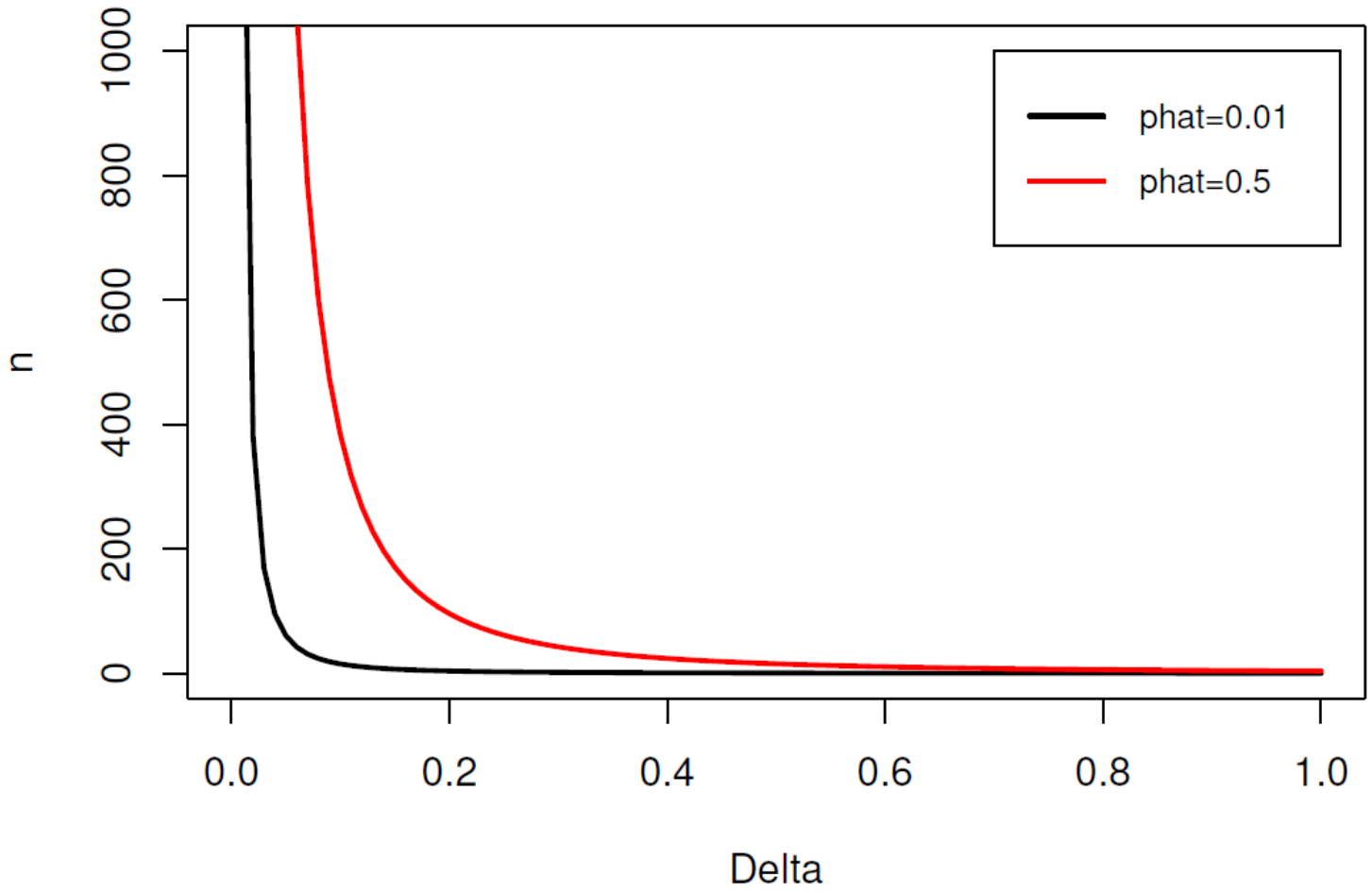
En el caso de la proporción (p), la amplitud del IC es:

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

Entonces,

$$n \geq \left[2\frac{z_{\alpha/2}}{\Delta}\right]^2 \hat{p}(1-\hat{p})$$

Tamaño muestral y amplitud del intervalo



Para pensar

Volvamos al ejemplo del costo medio del metro cuadrado. Suponga que hay dos empresas que usted podría contratar para construir. Realiza algunas estimaciones sobre la media del costo variable por metro cuadrado y calcula un intervalo de confianza al 95% para cada estimación, en dólares:

- Empresa 1 $\rightarrow IC^{95\%}_{\mu_1} = [1305, 1650]$
- Empresa 2 $\rightarrow IC^{95\%}_{\mu_2} = [1590, 1850]$

¿Cree usted que hay **evidencia suficiente** para establecer que la media del costo variable por metro cuadrado es mayor en la empresa 2 que en la empresa 1?

Un caso no tan de juguete...

NOTA `La Nación` 19/04/2015 (semana previa a las PASO en C.A.B.A.):

$ 27
Capital, GBA y La Plata
Recargo por envío al interior:
Córdoba, Santa Fe, Entre Ríos,
La Pampa y Buenos Aires: $3,80
Resto del país: $4,50
Recargo con Corriere della Sera $6
Año 146 | Número 51.584
Atención al lector: 5199-4777

LA NACION

Domingo 19 de abril de 2015 | lanacion.com

Mín. 14° • Máx. 26°
Algo nublado.
Vientos del Norte.
Espectáculos, página 16

ENCUESTA DE POLIARQUÍA PARA LA NACION

Rodríguez Larreta aventaja a Michetti por 4,9 puntos en la Capital

Cuando falta una semana para las PASO porteñas, está tercero el camporista Recalde y cuarto, Lousteau, de ECO; el macrismo suma 47,7% de intención de voto

H. RODRÍGUEZ LARRETA — PRO — 26,3%
GABRIELA MICHETTI — PRO — 21,4%
MARIANO RECALDE — FPV — 12,7%
MARTÍN LOUSTEAU — ECO — 10,6 %

En el país, Scioli supera a Macri; Massa, más lejos

El gobernador tiene hoy una intención de voto del 33,4%, pero no evitaría un ballottage

DANIEL SCIOLI — FPV — 33,4%
MAURICIO MACRI — PRO — 27,3%
SERGIO MASSA — F. RENOVADOR — 20,1%
MARGARITA STOLBIZER — SURGEN — 6,4%

Un caso no tan de juguete...

La letra chica, literalmente:

Fuente: Poliarquía. **Universo:** Personas residentes en la Ciudad de Buenos Aires, en hogares particulares con teléfono, mayores de 18 años de edad. **Tipo de encuesta:** Telefónica (Cati e IVR). **Características de la muestra:** Probabilística, polietápica y estratificada no proporcionalmente en tres zonas (Norte, Centro y Sur). Con posterioridad al trabajo de campo la muestra fue ponderada a fin de otorgar a cada localidad el peso que le corresponde en el total del conglomerado y de respetar el nivel de estudios de los entrevistados de acuerdo con los últimos datos censales. **Tamaño total de la muestra:** 1800 casos. **Error estadístico:** +/- 2.53% para un nivel de confianza del 95%. **Fecha finalización de campo:** del 14 al 17 de abril 2015 / LA NACION

Reformulando la pregunta del ejemplo de las constructoras, ¿cree usted que con una confianza del 95% hay evidencia suficiente para decir que la proporción de la población que votará a Larreta es mayor que la que votará a Michetti?

Más sobre esto en *Tests de Hipótesis*.