# Inferencia estadística

## Propiedades asintóticas

San Andrés

June 2, 2022

# Estimación puntual

## Estimaciones puntuales

Dada una muestra

$$X_1, X_2, \ldots, X_n$$

estimamos el parámetro $\theta_0$ reportando un solo valor $\widehat{\theta}$.

- ¿habremos acertado al valor verdadero $\theta_0$?
- ¿cuán lejos está nuestro estimador de $\theta_0$?
- ¿qué pasa si aumentamos el tamaño de muestra?

## Problema

Es un solo número, no da información sobre la precisión de la estimación. En particular, no incorpora la **variabilidad** del estimador.

# Estimación puntual

**Ejemplo**

Dada la siguiente muestra de una distribución normal con media $\mu_0$

$$4.37 \quad 5.18 \quad 4.16 \quad 6.60 \quad 5.33 \quad 4.18 \quad 5.49 \quad 5.74 \quad 5.58 \quad 4.69$$

Para estimar $\mu$ consideramos la media muestral que nos da

$$\overline{x} = 5.13$$

A partir de esta estimación, y sin saber como fueron generados los datos ¿Es razonable la muestra si $\mu_0 = 5$? ¿ y si $\mu_0 = 7$?

# Intervalos de confianza

## Ejemplo

Para cada muestra el valor de $\overline{X}$ es posiblemente distinto. Una alternativa a la estimación puntual es reportar una estimación del tipo

$$(\overline{X}_n - \textit{margen}; \overline{X}_n + \textit{margen})$$

es decir un **intervalo** donde esperamos hallar al valor de $\mu_0$.

Sabemos que

$$\mu_0 \in (\overline{X}_n - \textit{margen}, \overline{X}_n + \textit{margen})$$

o bien que

$$\mu_0 \notin (\overline{X}_n - \textit{margen}, \overline{X}_n + \textit{margen})$$

¿Cómo nos **aseguramos** que el intervalo siempre contenga a $\mu_0$?

# Intervalos de confianza

## Nivel de confianza

- Si tomamos

$$\textit{margen} = \infty$$

  tenemos certeza pero el intervalo no es informativo. Queremos un balance entre certeza e información.

- Dado un valor $\alpha$ (a elegir) buscamos margen tal que

$$P\left(\mu_0 \in \left(\overline{X}_n - \textit{margen}, \overline{X}_n + \textit{margen}\right)\right) = 1 - \alpha$$

- Llamamos a $1 - \alpha$ el **nivel de confianza**.

# Intervalos de confianza

## Intervalo de confianza

Un **intervalo de confianza** de nivel $(1-\alpha)$ para un parámetro $\theta$ es un intervalo

$$C_n = (A(X_1, \ldots, X_n), B(X_1, \ldots, X_n))$$

donde los bordes son funciones de los datos de manera que para todo $\theta$

$$P(\theta \in C_n) = 1 - \alpha\,.$$

- $C_n$ es aleatorio.
- $\theta$ es un valor (desconocido) fijo.
- Algunos valores frecuentes para $\alpha$ son 0.01 y 0.05.
- Hay un trade-off entre la longitud del intervalo y el nivel de confianza.

# Intervalos de confianza

En general, si el nivel de confianza es alto y el intervalo resultante es angosto, nuestro conocimiento sobre el parámetro es razonablemente preciso.

Un intervalo de confianza ancho, indica un alto nivel de incertidumbre sobre el valor estimado.

# Intervalos de confianza

- Un **estimador puntual** es una estrategia para calcular un valor que, con probabilidad alta si el estimador es consistente, se acerque al valor (desconocido) de un parámetro.
- Un **intervalo de confianza** es una estrategia para construir un intervalo de valores con una alta probabilidad de incluir al valor del parámetro desconocido.

# Intervalos de confianza

## Ejemplo: IC para la media de una normal ($\sigma$ conocido)

Supongamos una población normal con varianza **conocida**:

$$\overline{X}_n \sim \mathcal{N}\left(\mu_0, \sqrt{\frac{\sigma^2}{n}}\right)$$

Equivalentemente,

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

¿Qué significa en este caso que $\mu_0 \in (\overline{X}_n - \textit{margen}, \overline{X}_n + \textit{margen})$?

# Intervalos de confianza

Despejamos,

$$\begin{aligned}
\mu_0 \in (\overline{X}_n - \textit{margen}, \overline{X}_n + \textit{margen}) &\Leftrightarrow \overline{X}_n - \textit{margen} < \mu_0 < \overline{X}_n + \textit{margen} \\
&\Leftrightarrow -\textit{margen} < \mu_0 - \overline{X}_n < +\textit{margen} \\
&\Leftrightarrow -\textit{margen} < \overline{X}_n - \mu_0 < +\textit{margen} \\
&\Leftrightarrow -\frac{\textit{margen}}{\sqrt{\frac{\sigma^2}{n}}} < \frac{\overline{X}_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} < \frac{\textit{margen}}{\sqrt{\frac{\sigma^2}{n}}}
\end{aligned}$$

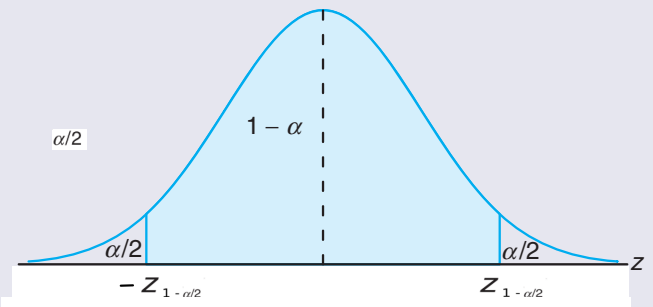
Queremos elegir margen para que

$$P\left(\mu_0 \in (\overline{X}_n - \textit{margen}, \overline{X}_n + \textit{margen})\right) = 1 - \alpha.$$

# Intervalos de confianza

## Notación

Llamamos $z_{1-\alpha/2}$ al cuantil $1-\alpha/2$ de la normal standard.



# Intervalos de confianza

Obtenemos que

$$\frac{\textit{margen}}{\sqrt{\frac{\sigma^2}{n}}} = z_{1-\alpha/2} \Leftrightarrow \textit{margen} = z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

## IC para la media de una población normal con varianza conocida

Dada una muestra aleatoria tamaño $n$ de una población normal con varianza $\sigma^2$ conocida, un intervalo de confianza de nivel $(1-\alpha)100\%$ para $\mu_0$ está dado por

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

# Intervalos de confianza

**Ejemplo**

Supongamos que se mide el peso de 36 personas adultas de Argentina elegidas al azar y se obtiene un peso promedio de 71kg. Supongamos que los pesos de la población adulta de Argentina están distribuidos como una $N(\mu, 100)$. Calcular los intervalos de confianza de niveles 95% y 99% basados en los datos observados.

- ¿Cuál de los intervalos es más amplio? ¿Si queremos estimar $\mu$ con mayor confianza, ¿El intervalo debe ser más ancho o mas angosto?
- Si el borde derecho del intervalo fuera 73 ¿Con qué confianza fue construido?

# Intervalos de confianza

**¿Qué significa tener un intervalo de confianza del 90% para $\mu_0$?**

- Simulemos $N = 1000$ muestras de tamaño $n = 10$ de una variable aleatoria normal con media 0 y desvío 0.5 y para cada una de esas muestras calculemos el IC del 90% para $\mu_0$.
- Obtendremos 1000 IC. Contemos cuántos contienen a $\mu_0 = 0$, la verdadera media de la población (podemos hacerlo porque estamos simulando y por lo tanto conocemos $\mu_0$).

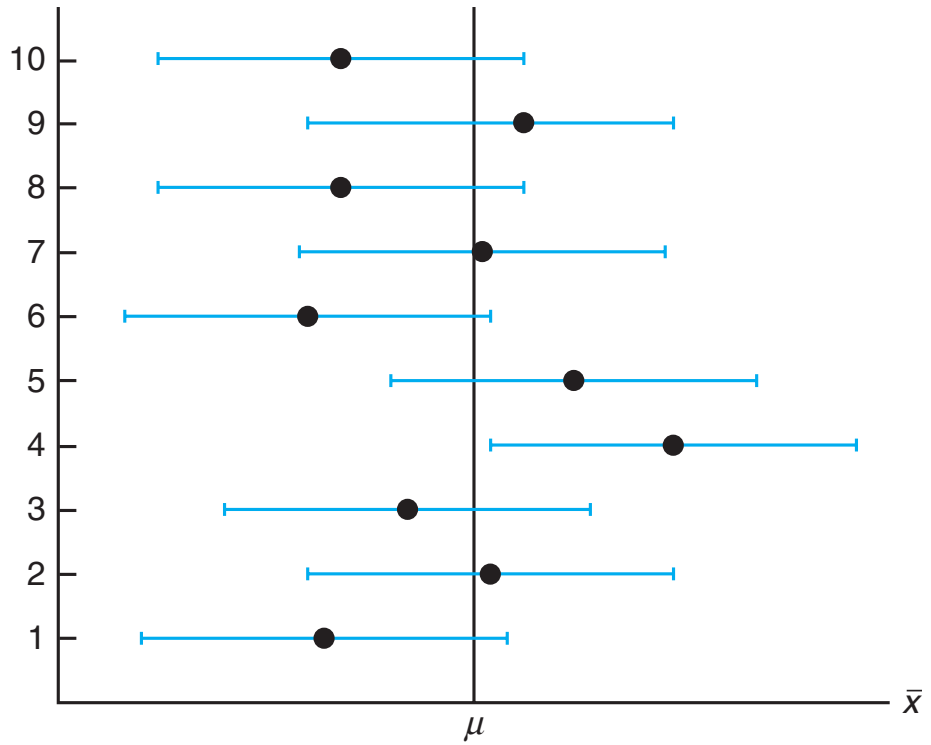
# Intervalos de confianza

```
# Definimos los parametros
nreps = 1000
mu = 0
n = 10
sigma = 0.5
alfa = 0.1
z = qnorm(1-alfa/2)
# Calculamos los IC
linf = lsup = numeric(nreps)
for (iter in 1:nreps){
  set.seed(iter)  # fijamos la semilla
  muestra = rnorm(n, mu, sigma)
  media = mean(muestra)
  linf[iter] = media - z * sigma/sqrt(n)
  lsup[iter] = media + z * sigma/sqrt(n)
}
# Proporcion que el IC contiene al mu verdadero
mean((mu<lsup) * (mu>linf))
```

# Intervalos de confianza

## Interpretación

Si repetimos muchas veces el experimento de tomar una muestra de tamaño $n$ y calcular el IC del 90% para $\mu_0$, aproximadamente el 90% de las veces la verdadera media $\mu_0$ pertenecerá al intervalo calculado (¡y el 10% no!)



# Intervalos de confianza

## Longitud del intervalo

Dada una muestra aleatoria tamaño $n$ de una población normal con varianza $\sigma^2$ conocida, un intervalo de confianza de nivel $(1-\alpha)100\%$ para $\mu_0$ está dado por

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

La longitud se define como el borde derecho menos el izquierdo. En este caso:

$$\text{longitud} = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

1. Disminuye si aumenta $n$.
2. Aumenta si aumenta la confianza $1-\alpha$.
3. Aumenta si aumenta el desvío $\sigma$ (que no controlamos...).

# Intervalos de confianza

## Error de estimación

Si usamos $\overline{X}_n$ como estimación de $\mu_0$, podemos tener una confianza de $(1-\alpha)$ de que el error no excederá

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

# Intervalos de confianza

## Tamaño de muestra

Si usamos $\overline{X}_n$ como estimación de $\mu_0$ podemos tener una confianza de $(1-\alpha)100\%$ de que el error no será mayor a

$$error = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Con esto, **si conocemos o podemos acotar** $\sigma$, podemos despejar un valor de $n$ para que el error sea menor que algún valor dado, con un nivel de confianza $1-\alpha$.

# Intervalos de confianza

**Ejemplo (peso)**

Halle el tamaño de muestra si queremos tener el 95% de confianza de que nuestra estimación difiera de $\mu_0$ por menos de medio kilo?

$$n = \left[\frac{1.96 \cdot 10}{0.5}\right]^2 \approx 1540$$

# Intervalos de confianza

**Ejemplo (peso)**

Halle el tamaño de muestra si queremos tener el 95% de confianza de que nuestra estimación difiera de $\mu_0$ por menos de medio kilo?

$$n = \left[\frac{1.96 \cdot 10}{0.5}\right]^2 \approx 1540$$

# ¿Qué pasa cuando la varianza es desconocida?

**Experimento**

Realizar una simulación para ver que sucede con la cobertura de los intervalos vistos.

# ¿Qué pasa cuando la varianza es desconocida?

- Si la varianza es conocida usamos que

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

- Cuando no conocemos $\sigma^2$ podemos estimarla mediante $S^2$ pero

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{S^2}{n}}} \sim F$$

no sabemos la distribución $F$ en este caso. Dos opciones:

Opción 1: buscamos la **distribución exacta**.
Opción 2: buscamos una **aproximación asintótica**.

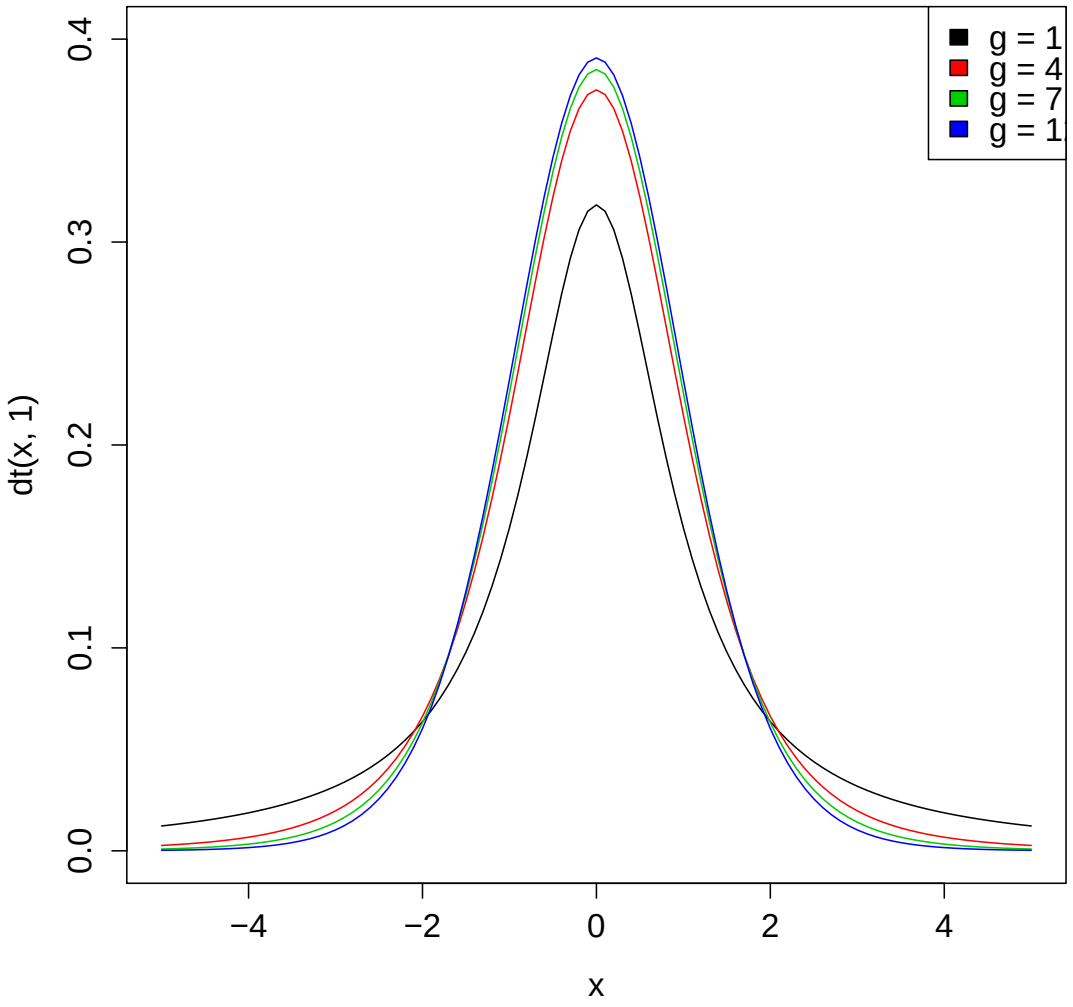
# Distribución t-Student

Si bien nos enfocaremos en las aproximaciones asintóticas mencionaremos brevemente el caso de la distribución t-Student.

La distribución t-Student depende de su parámetro **grados de libertad**.

```
grados = c(1, 4, 7, 12)
grilla = seq(-5, 5, length = 1000)
curvas = sapply(grados, dt, x=grilla)
matplot(grilla, curvas, type = 'l')
legend("topright", paste("g= ", grados),
       lty = 1:4, col = 1:4)
```

# Distribución t-Student



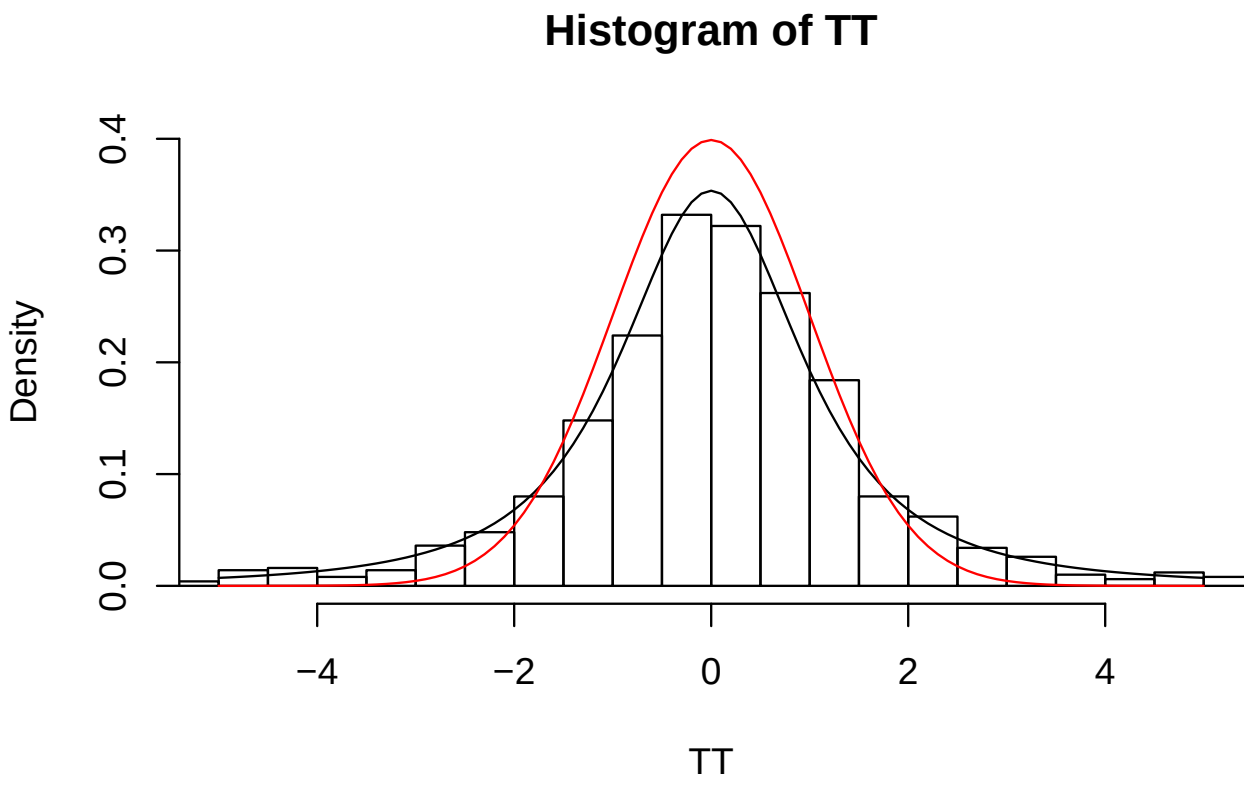
# Distribución de $\sqrt{n}(\overline{X}_n - \mu)/S$

```
# Parametros
n = 3
mu = 10
sigma = 2
N = 1000

# Medias muestrales estandarizadas
for(i in 1:N){
  muestra = rnorm(n, mu, sigma)
  xbarra = mean(muestra)
  s = sd(muestra)
  TT[i] = (xbarra-mu)/(s/sqrt(n))}

# Grafico
hist(TT, freq = FALSE, nclass = 100,
     xlim = c(-5, 5), ylim = c(0, 0.4))
curve(dt(x, n-1), add = TRUE)
curve(dnorm(x), add = TRUE, col = 'red')
```

# Distribución de $\sqrt{n}(\overline{X}_n - \mu_0)/S$




¿Cuál ajusta mejor?

# Intervalos de confianza

- La estimación por intervalo de confianza de $\mu_0$, cuando $\sigma$ es desconocido, utiliza el estadístico

$$\frac{\overline{X}_n - \mu_0}{\frac{S}{\sqrt{n}}}$$

el cual **bajo normalidad**, tiene distribución $t_{n-1}$.

- Entonces el intervalo de confianza de $(1-\alpha)100\%$ para $\mu_0$ está dado por

$$\overline{X}_n - t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} < \mu_0 < \overline{X}_n + t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$$

donde $t_{1-\frac{\alpha}{2},n-1}$ es el cuantil $1-\alpha/2$ de la distribución $t$ de Student con $n-1$ grados de libertad.

# Intervalos de confianza

```
nreps = 10000
mu = 0
n = 10
sig = 0.5
alfa = 0.05
t = qt(1-alfa/2, n-1)

linf = lsup = numeric(nreps)
for (i in 1:nreps) {
  set.seed(1)
  muestra = rnorm(n, mu, sig)
  media = mean(muestra)
  s = sd(muestra)
  linf[i] = media - t*s/sqrt(n)
  lsup[i] = media + t*s/sqrt(n)
}

# La proporcion de veces que contiene a mu
mean((mu<lsup) * (mu>linf))
```

# Intervalos de confianza

**Interpretación**

Si repetimos muchas veces el experimento de tomar una muestra de tamaño $n$ y hacer el IC del 95% para $\mu_0$, aproximadamente el 95% de las veces la verdadera media $\mu_0$ pertenecerá al intervalo (y el 5% no!)

**"Así, un intervalo de confianza de 95% nos indica que dentro del rango dado se encuentra el valor real de un parámetro con 95% de certeza"**

# Intervalos asintóticos

**Example**

¿Que sucede cuando la muestra no sigue una distribución normal? Realizar una simulación con una muestra con distribución $Exp(\lambda)$ con $\lambda = 2$. Estimar el verdadero nivel de confianza para un intervalo de confianza 95% deducido bajo normalidad. Hacer esto para $n = 5, 10, 50, 100$ y 200.
¿Qué observa? ¿Qué explicación encuentra a esto?

# Intervalos asintóticos

Encontrar la distribución exacta del estadístico puede ser muy difícil. Una alternativa es buscar aproximaciones a través de **resultados asintóticos**.

Si queremos un IC para la media de una distribución **cualquiera** cuando $\sigma$ es desconocido estimamos $\sigma$ con su estimador consistente $S$ y obtenemos que

$$\frac{\overline{X}_n - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{\sigma}{S} \cdot \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

que converge en distribución (TCL + Slutzky) a una normal estándar.

# Intervalos de confianza

## Intervalo de confianza

Un **intervalo de confianza** de nivel **asintótico o aproximado** $(1-\alpha)100\%$ para un parámetro $\theta$ es un intervalo

$$C_n = (A(X_1, \ldots, X_n), B(X_1, \ldots, X_n))$$

donde los bordes son funciones de los datos de manera que para todo $\theta$

$$P(\theta \in C_n) \to 1 - \alpha\,,$$

cuando $n \to \infty$.

# Intervalos asintóticos

**IC asintótico para la media de una población con varianza finita**

Dada una muestra aleatoria tamaño $n$ de una población varianza $\sigma^2$ finita pero desconocida, un intervalo de confianza de nivel aproximado $(1-\alpha)100\%$ para $\mu_0$ está dado por

$$\left(\overline{X}_n - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right).$$

# Intervalo de confianza para una proporción

## Ejemplo

Dada una muestra aleatoria $X_1, \ldots, X_n$ tal que $X \sim Be(p)$, consideramos el estimador

$$\widehat{p} = \sum_{i=1}^{n} X_i / n$$

Observemos que

$$\sum_{i=1}^{n} X_i \sim Bin\,(n, p)$$

# Intervalo de confianza para una proporción

## Ejemplo

Luego, si $n$ es suficientemente grande, la distribución muestral de $\widehat{p}$ se aproxima a una normal con media $p$ y varianza $p(1-p)/n$:

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

Repetimos el procedimiento anterior para construir un intervalo de confianza para la proporción $p$:

$$p \in (\hat{p} - \mathit{margen}, \hat{p} + \mathit{margen}) \Leftrightarrow -\frac{\mathit{margen}}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\mathit{margen}}{\sqrt{\frac{p(1-p)}{n}}}$$

Entonces tomamos $\frac{\mathit{margen}}{\sqrt{\frac{p(1-p)}{n}}} \approx z_{1-\alpha/2}$ o $\mathit{margen} \approx z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

# Intervalo de confianza para una proporción

## Ejemplo

Luego,

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

Los bordes dependen del parámetro que queremos estimar, usamos su estimación $\widehat{p}$. **Por qué es válido esto?**

## IC asintótico para una proporción

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño $n$, un intervalo de confianza aproximado de $(1-\alpha)100\%$ para $p$ está dado por

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

# Intervalos de confianza asintóticos

**Ejemplo**

En una muestra aleatoria de $n = 500$ familias que tienen televisores en la ciudad de Hamilton, Canadá, se encuentra que $X = 340$ están suscriptas a HBO. Encuentre un intervalo de confianza del 95% para la proporción real de familias en esta ciudad que están suscriptas a HBO.

# Intervalos de confianza asintóticos

```
# En R:
prop.test(340, 500)

1-sample proportions test with continuity correction
data:  340 out of 500, null probability 0.5
X-squared = 64.082, df = 1, p-value = 1.193e-15
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.6368473 0.7203411
sample estimates:
   p
0.68
```

# Intervalos de confianza asintóticos

## Intervalos asintóticos

Supongamos ahora más generalmente que $\widehat{\theta}_n$ es un estimador de $\theta$ que es asintóticamente normal:

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, V(\theta)\right),$$

donde $V(\theta)$ es la varianza asintótica del estimador. Supongamos que tenemos un estimador consistente de $V(\theta)$. Por ejemplo, supongamos que $V(\widehat{\theta}_n)$ es consistente para $V(\theta)$. Entonces, por Slutzky

$$\sqrt{n}\frac{\left(\widehat{\theta}_n - \theta\right)}{\sqrt{V(\widehat{\theta}_n)}} \xrightarrow{d} \mathcal{N}\left(0, 1\right),$$

Luego, dado un $\alpha > 0$, vale que, para $n$ grande

$$P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\sqrt{V(\widehat{\theta}_n)}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

# Intervalos de confianza asintóticos

Un intervalo de confianza de nivel aproximado $(1-\alpha)100\%$ es entonces

$$\widehat{\theta}_n \pm z_{1-\alpha/2}\sqrt{\frac{V(\widehat{\theta}_n)}{n}}$$

# Intervalos de confianza basados en máxima verosimilitud

## Distribución asintótica del EMV

Vimos que bajo ciertas condiciones de regularidad el estimador de máxima verosimilitud es asintóticamente normal:

$$\sqrt{n}\left(\frac{\widehat{\theta}_{MV}-\theta_0}{\sqrt{\frac{1}{I_1(\theta_0)}}}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

## IC asintótico para el EMV

Supongamos que $I_1(\widehat{\theta}_{MV})$ es consistente para $I_1(\theta)$. Si definimos el intervalo

$$C_n = \left(\widehat{\theta}_{MV} - z_{1-\alpha/2}\sqrt{\frac{I_1\left(\widehat{\theta}_{MV}\right)^{-1}}{n}}; \widehat{\theta}_{MV} + z_{1-\alpha/2}\sqrt{\frac{I_1\left(\widehat{\theta}_{MV}\right)^{-1}}{n}}\right)$$

Es un intervalo de nivel asintótico $1-\alpha$.

# Intervalos de confianza basados en máxima verosimilitud

La longitud del intervalo resultante es

$$2z_{1-\alpha/2}\sqrt{\frac{I_1^{-1}(\widehat{\theta}_{MV})}{n}} \approx 2z_{1-\alpha/2}\sqrt{\frac{I_1^{-1}(\theta_0)}{n}}.$$

Luego, con probabilidad aproximadamente $1-\alpha$ (si $n$ es 'grande'), el error de estimación es a lo sumo

$$2z_{1-\alpha/2}\sqrt{\frac{I_1^{-1}(\theta_0)}{n}}.$$

# Intervalos de confianza

**Observación**

Notar que ningún estimador asíntoticamente normal puede dar lugar a un intervalo de confianza de nivel $(1 - \alpha)$ con longitud menor que el que da máxima verosimilitud. **Por qué?**

Usando máxima verosimilitud, para un mismo nivel dado, obtenemos resultados más precisos, con menos incertidumbre, que con cualquier otro estimador.

# Intervalos de confianza basados en máxima verosimilitud

## Ejemplo

Sea $X_1, \ldots, X_n$ una muestra aleatoria con distribución Poisson de parámetro $\lambda$. El estimador de máxima verosimilitud es

$$\widehat{\lambda}_{MV} = \overline{X}_n$$

y el número de información de Fisher es

$$I_1(\lambda) = 1/\lambda$$

entonces

$$I_1(\widehat{\lambda}_{MV}) = 1/\widehat{\lambda}_{MV}$$

es consistente para $I_1(\lambda)$. Luego, un intervalo de confianza de nivel asintótico $1 - \alpha$ para $\lambda$ es

$$\widehat{\lambda}_{MV} \pm z_{1-\alpha/2}\sqrt{\widehat{\lambda}_{MV}/n}$$

# Intervalos de confianza basados en máxima verosimilitud

## Ejemplo

Dadas $X_1, X_2, \ldots, X_n \sim Be(p)$ y $\psi = g(p) = \log(p/(1-p))$.
Por invarianza, EMV de $\psi$ es $\widehat{\psi}_{MV} = \log \widehat{p}_n/(1-\widehat{p}_n)$. Por el método delta, la varianza asintótica de $\widehat{\psi}$ es

$$\frac{1}{p(1-p)}.$$

Esto lo podemos estimar consistentemente usando

$$\frac{1}{\widehat{p}_n(1-\widehat{p}_n)}.$$

Un intervalo de confianza de nivel aproximado 95% es entonces

$$\widehat{\psi}_{MV} \pm \frac{1.96}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}$$

# Intervalo de confianza para una estimación no paramétrica

Supongamos que queremos estimar la distribución del ingreso por hogar en Argentina. Tenemos $X_1, \ldots, X_n$ el ingreso mensual de $n$ hogares elegidos al azar.

- Cómo damos un intervalo de confianza para

  $F(x_0) = P(X \leq x_0) =$ fracción de hogares que ganan a los sumo $x_0$,

  para un $x_0$ fijo?

# Intervalo de confianza para una estimación no paramétrica

Dado $x_0$, si llamamos $Y_i = I(X_i \leq x_0)$, vemos que $Y_i \sim Ber(F(x_0))$. El problema se reduce a estimar el parámetro de una Bernoulli.

## Distribución empírica

Dada $X_1, \ldots, X_n$ una muestra aleatoria con distribución $F$ definida sobre la recta real, estimamos la función de distribucion acumulada $F$ como

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

$\widehat{F}_n(x)$ es simplemente la media muestral de $I(X_1 \leq x), \ldots, I(X_n \leq x)$. Usando esto, se puede probar que:

- $E(\widehat{F}_n(x)) = F(x)$
- $Var(\widehat{F}_n(x)) = F(x)(1 - F(x))/n$
- $\widehat{F}_n(x) \xrightarrow{P} F(x)$

# Intervalo de confianza para una estimación no paramétrica

Dado $x_0$, si llamamos $Y_i = I(X_i \leq x_0)$, vemos que $Y_i \sim Ber(F(x_0))$. El problema se reduce a estimar el parámetro de una Bernoulli.

## Distribución empírica

Dada $X_1, \ldots, X_n$ una muestra aleatoria con distribución $F$ definida sobre la recta real, estimamos la función de distribucion acumulada $F$ como

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

$\widehat{F}_n(x)$ es simplemente la media muestral de $I(X_1 \leq x), \ldots, I(X_n \leq x)$. Usando esto, se puede probar que:

- $E(\widehat{F}_n(x)) = F(x)$
- $Var(\widehat{F}_n(x)) = F(x)(1 - F(x))/n$
- $\widehat{F}_n(x) \xrightarrow{P} F(x)$

# Intervalo de confianza para una estimación no paramétrica

Además,

$$\widehat{F}_n(x_0) \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{F}_n(x_0)(1-\widehat{F}_n(x_0))}{n}}$$

es un intervalo de confianza de nivel asintótico 95% para $F(x_0)$.