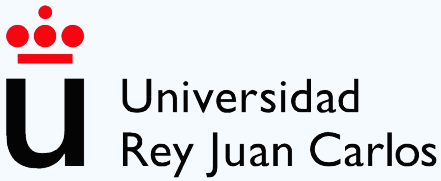
Estadística paramétrica. Intervalos de confianza

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



¿De dónde venimos?

- Usamos estimación puntual para obtener una estimación del parámetro poblacional
  - El número medio de horas/día que pasan los estudiantes en Instagram es 3.7
- Dicha estimación es un único valor. ¡Nuestra mejor apuesta! Pero un único valor
- Los estimadores tienen distribución muestral. Con cada muestra, obtendremos una estimación puntual distinta
- Luego, cuando damos una estimación del parámetro poblacional en base a una muestra, vamos a fallar seguro...basta con coger otra muestra y ver que la estimación cambia/puede cambiar
- ¿Solución? Acompañar la estimación de su variabilidad (error estándar)

Estimación por intervalo

- ¿Solución? Acompañar la estimación de su variabilidad (error estándar)
- Con la estimación puntual y su error estándar —> construiremos **intervalos de confianza** para el parámetro poblacional de interés θ

$$\hat{\theta} \pm \text{error estándar}$$

- Estos intervalos tendrán un **nivel de confianza** $1 - \alpha$ asociado que indicará con qué seguridad el intervalo contiene el verdadero valor del parámetro

Estimación por intervalo

- La estimación puntual proporciona una aproximación razonable para un parámetro de la población, pero no tiene en cuenta la variabilidad debido al tamaño muestral, la variabilidad en la población, el conocimiento de otros parámetros, etc.
- La **estimación por intervalo** es una técnica en estadística que, a diferencia de la estimación puntual que proporciona un único valor, ofrece un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional desconocido con un cierto nivel de confianza. Este rango se denomina **intervalo de confianza (IC)**
- La estimación por intervalos es una herramienta esencial en la inferencia estadística, ya que no solo ofrece una estimación del parámetro poblacional, sino que también proporciona un marco para entender la precisión y confiabilidad de esa estimación. Esto la convierte en una técnica poderosa para hacer inferencias más robustas y útiles basadas en datos muestrales

Conceptos clave

- **Intervalo de Confianza (IC)**: Es un rango de valores calculado a partir de los datos de la muestra, que se utiliza para estimar el parámetro poblacional θ desconocido. Se expresa comúnmente como (Límite Inferior, Límite Superior).
  Dada una muestra aleatoria simple $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de una población X con función de distribución F que depende de un parámetro desconocido θ, diremos que un estimador por intervalos de confianza del parámetro θ con un nivel de confianza de $(1-\alpha) = 100 * (1-\alpha)\%$ es un intervalo de la forma $(T_{inf}(\mathbf{X}), T_{sup}(\mathbf{X}))$ que satisface:

$$P(\theta \in (T_{inf}(\mathbf{X}), T_{sup}(\mathbf{X}))) = 1 - \alpha$$

  Nótese que el parámetro θ es un valor fijo, pero $T_{inf}(\mathbf{X})$ y $T_{sup}(\mathbf{X})$ son cantidades aleatorias que variarán con cada muestra

Conceptos clave

- **Nivel de Confianza**: Es la probabilidad teórica de que el intervalo de confianza contenga el verdadero valor del parámetro poblacional. Se denota como $1 - \alpha$, donde α es el nivel de significancia.

Cuando tomamos una muestra, los valores ya son fijos, por lo que no podemos seguir hablando en términos de probabilidad

Conceptos clave: nivel de confianza

- Un **nivel de confianza** común es el 95%, lo que significa que estamos un 95% seguros de que el intervalo contiene el parámetro verdadero. Si repetimos el experimento N veces, en el 95% de las ocasiones el verdadero valor del parámetro estará incluido en el intervalo proporcionado. Sin embargo es importante señalar que, dado que el experimento solo suele realizarse en una ocasión, no podemos estar seguros de que el verdadero valor del parámetro está incluido en nuestro intervalo. Estará incluido o no estará incluido, pero no podemos saber en qué situación nos encontramos. Estar seguro sería tanto como decir que conocemos el verdadero valor del parámetro. En ese caso, obviamente, no necesitaríamos estimación ninguna.

Conceptos clave

- **Error Estándar (SE)**: Es una medida de la variabilidad de un estimador o estadístico muestral en las distintas muestras
- No debe confundirse con la desviación típica, que se refiere a la variabilidad de las observaciones individuales.
- Se utiliza para calcular los límites del intervalo de confianza.

Cálculo del Intervalo de Confianza

- El cálculo de un intervalo de confianza generalmente sigue la fórmula:

$$\text{Estimación Puntual} \pm (\text{Valor Crítico} \times \text{Error Estándar})$$

- Para alcanzar el intervalo de confianza, generalmente se busca una cantidad (aleatoria) $C(\mathbf{X}, \theta)$ relacionada con el parámetro desconocido θ y con la muestra $\mathbf{X}$, cuya distribución sea conocida y no dependa del valor del parámetro
- Esta cantidad recibe el nombre de *pivote* o *cantidad pivotal* para θ (¡aquí entran en juego las distribuciones muestrales de los estadísticos!)
- El valor crítico dependerá de la distribución teórica

Cálculo del Intervalo de Confianza

Dado que conocemos la distribución del pivote (conocemos la distribución del estimador), podemos usar los cuartiles $1-\alpha/2$ y $\alpha/2$ de dicha distribución, y el error estándar del estimador por intervalos de confianza, para plantear la siguiente ecuación:

$$P(\text{ cuantil}_{1-\alpha/2} < C(\mathbf{X},\theta) < \text{ cuantil}_{\alpha/2}) = 1-\alpha$$

Para obtener los extremos (inferior y superior) del estimador por intervalos de confianza $T_{inf}(\mathbf{X})$ y $T_{sup}(\mathbf{X})$, se resuelve la doble desigualdad en θ. De este modo el intervalo de confianza al $100(1-\alpha)\%$ para θ es $(T_{inf}(\mathbf{X}), T_{sup}(\mathbf{X}))$

IC para la media con var conocida

Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ de tamaño n obtenida de X, siguiendo una distribución normal con parámetros (μ desconocido) y varianza conocida (σ^2), $N(\mu, \sigma^2)$. Queremos obtener un IC para la media μ con un nivel de confianza $1 - \alpha$.

Como ya hemos visto, el estadístico $\bar{X}$ tiene una distribución normal:

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

Además, sabemos que:

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

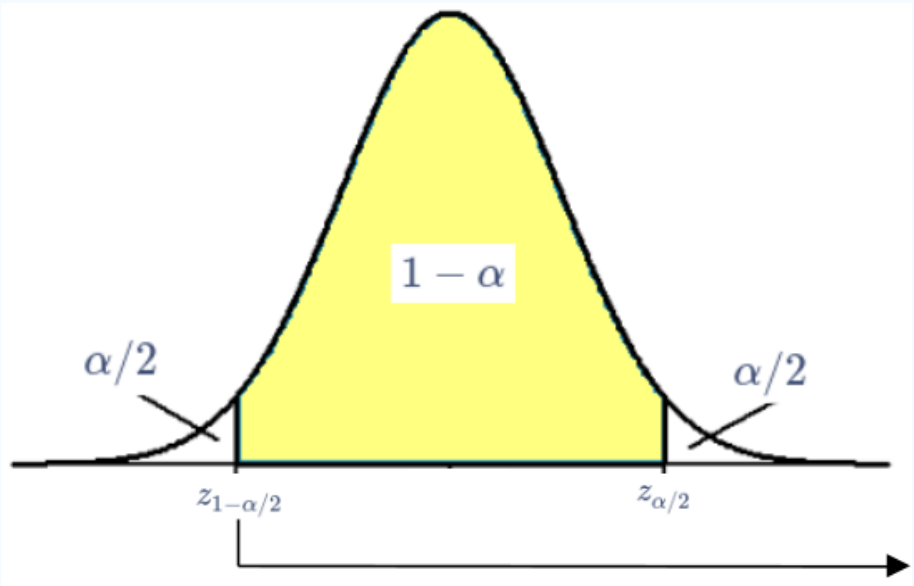
que es la cantidad pivotal para el IC para μ

IC para la media con var conocida

Ahora, si $z_{1-\alpha/2}$ y $z_{\alpha/2}$ son los cuartiles $1-\alpha/2$ y $\alpha/2$ de la distribución $N(0,1)$, entonces tenemos:

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

Nótese que, como la distribución Normal es simétrica: $z_{1-\alpha/2} = -z_{\alpha/2}$



Resolvemos la doble desigualdad para μ:

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

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

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X} < -\mu < -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + \bar{X} > \mu > \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

De modo que el estimador por intervalos de confianza es:

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

y por tanto, el intervalo de confianza para la media se calcula como:

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

donde $\bar{X}$ es la media muestral, $z_{\alpha/2}$ es el valor crítico del estadístico z para el nivel de confianza deseado y $\sigma/\sqrt{n}$ es el error estándar

¿Qué puede afectar al IC?

Pensemos juntos...¿qué factores pueden influir en la longitud intervalo de confianza?

¿Qué puede afectar al IC?

- Tamaño muestral
- Variabilidad
- Nivel de confianza

Aclaración cuantiles

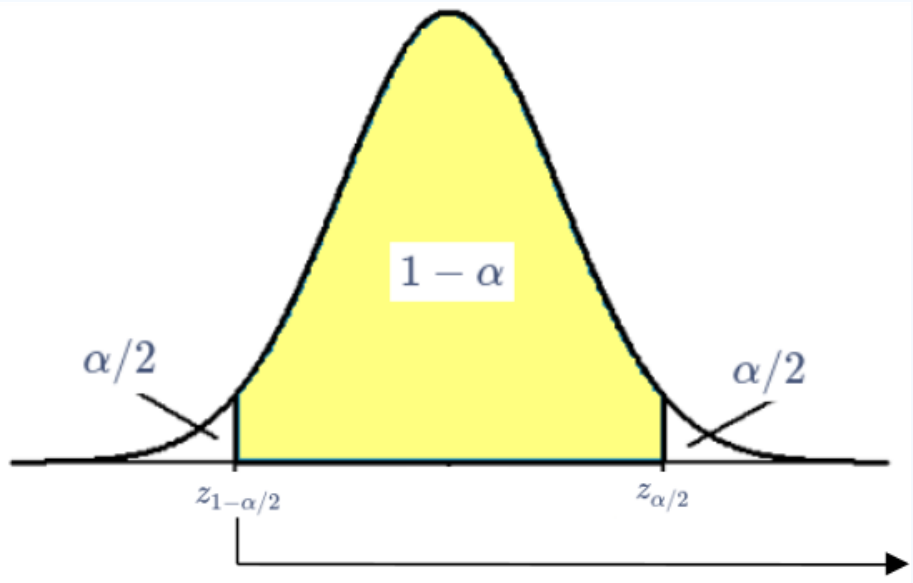


Figure 1: Cola superior

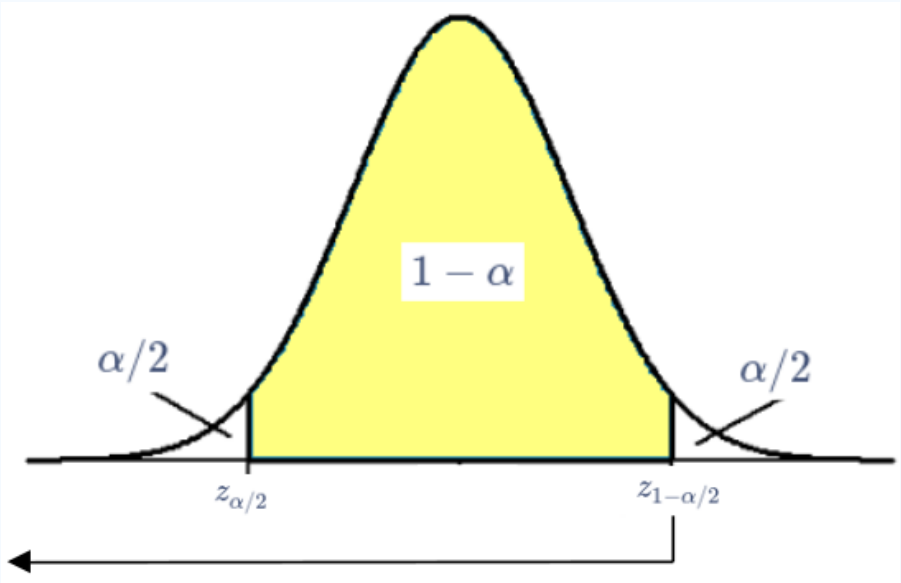


Figure 2: Cola inferior (en R: lower.tail = TRUE)

En R, el parámetro `lower.tail` (en funciones del tipo `pnorm()`, `qnorm()`) nos permite elegir la cola:

- Para la cola superior, `lower.tail = FALSE`
- Para la cola inferior, `lower.tail = TRUE`

IC para la media con var conocida

Se ha probado que la altura de las alumnas de primer curso de la URJC se puede aproximar mediante una variable aleatoria con distribución normal con desviación típica $\sigma = 10$ cm pero la media μ desconocida. En un estudio con 50 alumnas se obtiene una media de 166 cm. Vamos a construir un intervalo de confianza al 95% para μ.

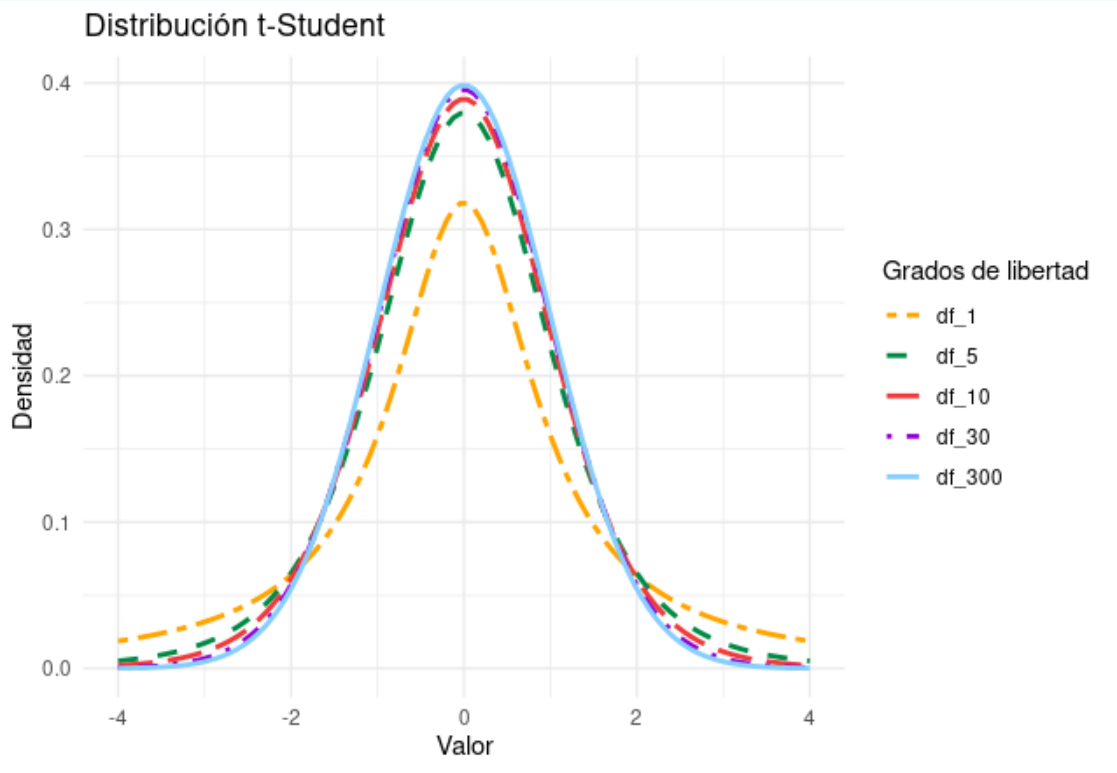
IC para la media con var desconocida

- Concepto previo: Distribución t-Student

Distribución t-Student

La distribución t-student t_n tiene un único parámetro $n > 0$ llamado grados de libertad

- Se parece a la distribución Normal pero con colas "más pesadas"
- Según $n \longrightarrow \infty$, la distribución es una $N(0, 1)$
- Data $T \sim t_n$ para $n > 1$, $E[T] = 0$ y para $n > 2$, $V[T] = \frac{n}{n-2}$



Distribución t-Student

- Sea $X_1, \ldots, X_n$ una muestra aleatoria simple (v.a.i.i.d.) de una población $N(\mu, \sigma^2)$, entonces el estadístico T:

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n - 1}} = \frac{\bar{X} - \mu}{\sqrt{\hat{s}^2/n}} \sim t_{n-1}$$

  sigue la distribución t de Student con $n - 1$ grados de libertad.

- La distribución t de Student se puede definir como

$$T = \frac{N(0,1)}{\sqrt{\chi^2_n/n}}$$

  siendo $N(0,1)$ la normal estándar, χ^2_n es la distribución chi-cuadrado con n grados de libertad, siendo la normal y la chi-cuadrado independientes

Distribución t-Student

En R las funciones de la t-Student son:

- `dt()`: Función de densidad de probabilidad
- `qt()`: Función cuantil de la distribución
- `pt()`: Función de distribución acumulada
- `rt()`: Generación de números pseudoaleatorios

IC para la media con var desconocida

Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ de tamaño n obtenida de X, siguiendo una distribución normal con parámetros μ y varianza σ^2 desconocidos . Queremos obtener un IC para la media μ con un nivel de confianza $1 - \alpha$.

La cantidad pivotal para μ es:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

donde S^2 es la es la cuasi-varianza muestral y t_n es la distribución t de Student con n grados de libertad

IC para la media con var desconocida

Si $t_{n-1;1-\alpha/2}$ y $t_{n-1;\alpha/2}$ son los cuantiles $1-\alpha/2$ y $\alpha/2$ respectivamente de una distribución t de Student con $n-1$ grados de libertad:

$$P(t_{n-1;1-\alpha/2} < T < t_{n-1;\alpha/2}) = 1-\alpha$$

Como la distribución t-Student es simétrica: $t_{n-1;\alpha/2} = t_{n-1;1-\alpha/2}$

$$P(-t_{n-1;\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{n-1;\alpha/2}) = 1-\alpha$$

IC para la media con var desconocida

Se resuelve la doble desigualdad para μ y se obtiene el estimador por intervalos de confianza:

$$\left(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right)$$

Resultando el intervalo de confianza:

$$IC_{1-\alpha}(\mu) = \left(\bar{\mathbf{X}} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \bar{\mathbf{X}} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right)$$

IC para la media con var desconocida

Se ha medido la temperatura media de una muestra aleatoria de 10 soluciones salinas, obteniendo los siguiente resultados:

$$37.2, 34.1, 35.5, 34.5, 32.9, 37.3, 32.0, 33.1, 42.0, 34.8$$

Se nos pide calcular el IC al 90% para la temperatura media, suponiendo que la temperatura de la solución salina se puede aproximar mediante una variable aleatoria con distribución normal

IC para la media con n grande

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria simple (v.a.i.i.d) de tamaño n de una variable aleatoria X. Supongamos que X sigue una distribución (conocida o no) con parámetros μ y σ^2. Además, supongamos que $n \geq 30$. Entonces, por el *Teorema Central del Límite* se tiene que la cantidad pivotal para μ cumple la siguiente propiedad:

$$Z = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0,1)$$

Si $z_{1-\alpha/2}$ y $z_{\alpha/2}$ son los cuantiles $1-\alpha/2$ y $\alpha/2$ de $N(0,1)$, tenemos:

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

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

IC para la media con n grande

Obtenemos el estimador por intervalos de confianza resolviendo la doble desigualdad para μ:

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

El intervalo de confianza es:

$$IC_{1-\alpha}(\mu) = \left(\bar{\mathbf{x}} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{\mathbf{x}} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

IC para la media para la proporción

Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ de tamaño n obtenida de X, siguiendo una distribución de Bernoulli con parámetro p. Esto es:

$$\mu = E[X] = p \qquad \sigma^2 = Var[X] = p(1-p)$$

Además, supongamos que $n \geq 30$. Entonces, por el TCL se tiene que la cantidad pivotal para $\hat{p} = \bar{X}$ cumple la siguiente propiedad:

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

EL intervalo de confianza con nivel de confianza $1-\alpha$ para estimar una proporción poblacional p es:

$$IC_{1-\alpha}(p) = \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

IC para la media para la proporción

- Si el tamaño muestral no es lo suficientemente grande y no se puede aplicar el Teorema Central del Límite –> hay que comprobar la normalidad de los datos
- Si los datos son normales, el intervalo de confianza anterior es apropiado
- En caso contrario, tendrían que buscarse alternativas

IC para la media para la proporción

Supongamos que estamos realizando una encuesta para determinar la proporción de personas que apoyan una nueva política ambiental en una ciudad. Hemos encuestado a 1000 personas, y 560 de ellas han respondido que apoyan la nueva política.

Queremos calcular un intervalo de confianza del 95% para la proporción de apoyo en toda la población.

Interpretación de los IC

- Si calculamos un intervalo de confianza del 95% para la media poblacional y, por ejemplo, obtenemos un intervalo de $(5, 10)$, esto no significa que hay un 95% de probabilidad de que la media poblacional esté en ese intervalo en un caso particular, sino que, si repetimos este procedimiento muchas veces, el 95% de los intervalos construidos contendrán la verdadera media poblacional.
- Podríamos decir que estamos un 95% seguros de que la media poblacional se encuentra entre 5 y 10, pero ¡ojo!, la media poblacional (cuyo valor desconocemos) estará o no estará en ese intervalo.

Interpretación de los IC

Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability - J. Neyman

It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results *will* tend to α.* Consider now the case when a sample, E′, is already drawn and the calculations have given, say, $\underline{\theta}\,(E') = 1$ and $\overline{\theta}\,(E') = 2$. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α ?

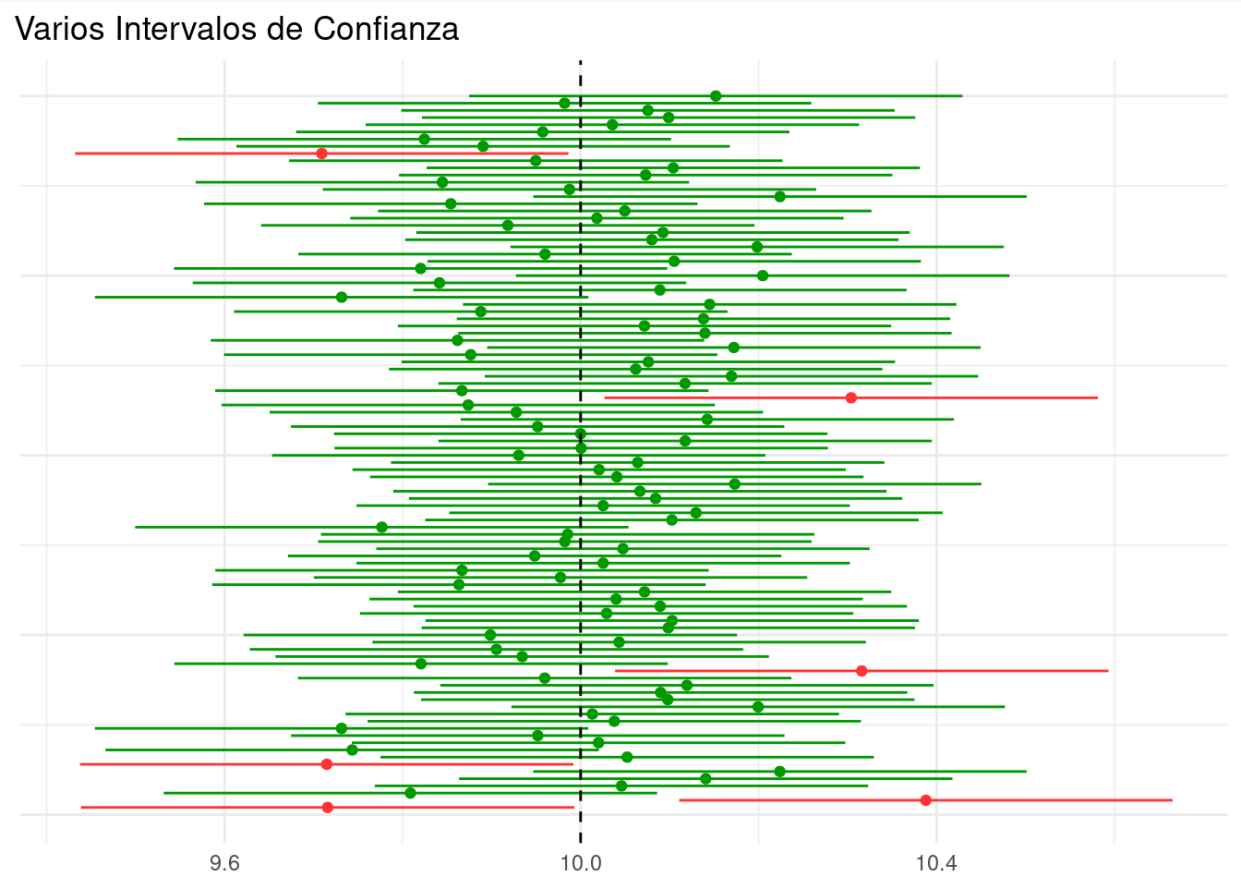
The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made, that is except for

Notación en este artículo: $$P\{\underline{\theta}\,(E) \leq \theta_1{}^0 \leq \overline{\theta}\,(E) | \theta_1{}^0\} = \alpha$$

Simulación para interpretar correctamente el concepto frecuentista de intervalo de confianza

- Generamos 100 muestras de tamaño $n = 50$ de una distribución $X \sim N(\mu = 10, \sigma^2 = 1)$
- Para cada muestra, se construye un IC para la media con $\alpha = 0.05$.
- Representamos todos esos intervalos de confianza en un único gráfico. En verde se pintan los intervalos de confianza que incluyen el verdadero valor del parámetro 10. En rojo los que no

Interpretación de los IC



- ¿Cuántos intervalos de confianza, de entre los 100 contienen al verdadero valor del parámetro?

Interpretación de los IC

- ¿Cómo crees que afecta a la longitud del intervalo de confianza los siguientes aspectos?
  - Tamaño muestral
  - Nivel de confianza

Tamaño muestral

- Un tamaño muestral adecuado es crucial en la inferencia estadística
- Garantiza que los intervalos de confianza sean precisos y que las conclusiones obtenidas sean representativas de la población
- Las técnicas para determinar el tamaño muestral están relacionadas directamente con los intervalos de confianza y se basan en varios factores:
  - Nivel de confianza deseado
  - La precisión (o margen de error) deseada
  - La variabilidad esperada en la población

Tamaño muestral: factores clave

1. **Nivel de confianza** $1-\alpha$: indica el grado de certeza con el que intervalo de confianza contiene al parámetro poblacional. Niveles de confianza comunes son 90%, 95% y 99%. Un nivel de confianza más alto requiere una muestra más grande para asegurar la misma precisión
2. **Margen de Error (E)**: Es la máxima diferencia tolerable entre la estimación muestral y el valor real del parámetro poblacional. Un margen de error más pequeño requiere una muestra más grande para asegurar una estimación precisa
3. **Variabilidad poblacional** σ: La variabilidad en la población, medida por la desviación estándar, afecta directamente al tamaño muestral. Una mayor variabilidad requiere una muestra más grande para obtener una estimación precisa

Tamaño muestral para media poblacional (Normal)

El tamaño muestral n necesario para estimar una media poblacional con un margen de error E y un nivel de confianza $1-\alpha$ se puede calcular usando la fórmula:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

donde:

- $z_{\alpha/2}$ es el valor crítico del estadístico z correspondiente al nivel de confianza deseado
- σ es la desviación estándar de la población (si es desconocida, se puede usar la desviación estándar de la muestra S)

Tamaño muestral para media poblacional (Normal)

¿De dónde sale esta fórmula?

Efectivamente, teníamos que el intervalo de confianza para la media era:

$$\left(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Y buscamos el n para que el margen de error sea menor que E, es decir:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < E$$

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Ejemplo

Supongamos que deseamos estimar la media de una población con un nivel de confianza del 95%, un margen de error de 5 unidades y se estima que la desviación estándar de la población es 15 unidades. ¿Qué tamaño muestral se necesita?

Tamaño muestral para proporción poblacional

El tamaño muestral n necesario para estimar una proporción poblacional p con un margen de error E y un nivel de confianza $1-\alpha$ se puede calcular usando la fórmula:

$$n = \frac{z_{\alpha/2}^2 \cdot p \cdot (1-p)}{E^2}$$

donde

- p es la proporción esperada (si no se conoce, se usa $p = 0.5$ para maximizar el tamaño muestral)
- $z_{\alpha/2}$ es el valor crítico del estadístico z correspondiente al nivel de confianza deseado

Ejemplo

Supongamos que deseamos estimar la proporción de personas que aprueban una nueva ley con un nivel de confianza del 95%, un margen de error del 3% (0.03) y se estima que la proporción esperada es $p = 0.5$. ¿Qué tamaño muestral se necesita?

Referencias

Wasserman, L. (2013). All of Statistics: a Concise Course in Statistical Inference.

Spiegel, M., & Stephens, L. (2009). Estadística–Serie Schaum. *Mc Graw-Hill*.

Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

Canavos, G. C., & Medal, E. G. U. (1987). *Probabilidad y estadística* (p. 651). México: McGraw Hill.