# “Introducción a la Estadística, Probabilidad e Inferencia”

Maestría en Estadística Aplicada

Facultad de Ciencias Económicas y Estadística

UNR

# Unidad 6

- Estimación puntual y por intervalos de confianza.
- Contrastes de hipótesis.
- Probabilidades de error tipo I y II.
- Determinación del tamaño muestral.
- Algunos casos particulares.

En esta unidad trabajaremos con **Estadística Inferencial**: realizaremos inferencia sobre parámetros (poblacionales) utilizando datos muestrales.

Los métodos inferenciales permiten predecir cuán cerca se encuentran los valores de las estadísticas con respecto a los parámetros.

Lo visto en las unidades previas será fundamental:

- Los métodos de estadística inferencial utilizan en sus cálculos probabilidades y asumen que los datos fueron obtenidos mediante un muestreo o un experimento aleatorio.
- Los cálculos de probabilidades se refieren a la distribución en el muestreo de una estadística.

Hay dos tipos de métodos en la inferencia estadística:

- **estimación de parámetros** y
- **contrastes o pruebas (test) de hipótesis** sobre dichos parámetros.

Recordar…

- Una **estadística** es un valor que describe una muestra y que se calcula con valores muestrales.
- Un **parámetro** es un valor que describe una población.

# Estimación de parámetros, puntual y por intervalos de confianza

La estimación es un procedimiento para inferir el valor de un parámetro. Puede realizarse de dos formas distintas.

- Un **estimador puntual** utiliza los datos muestrales para calcular un único valor que sirva como estimación de un parámetro; por sí mismo no es suficiente ya que no informa cuán cerca es probable que esté del parámetro desconocido.
- Un **estimador de intervalo** utiliza los datos muestrales para calcular dos valores que definan un intervalo que, con alguna probabilidad previamente determinada, contendrá al parámetro. Se lo considera un método de estimación más informativo al incorporar un margen de error, de manera que permite juzgar la precisión de la estimación puntual.

# Estimación puntual

Una vez que se tiene un conjunto de datos, es directa la estimación del valor de interés; solamente debe elegirse la estadística apropiada.

Por ejemplo, para estimar un promedio poblacional (μ) puede utilizarse el promedio muestral; para estimar una proporción (p), la proporción muestral.

Al momento de elegir la estadística a emplear será necesario considerar los métodos de estimación y las propiedades de los estimadores (Unidad 5).

# Estimación por intervalo de confianza

Una estimación por intervalo incorpora en cierta medida información sobre la precisión, al dar un intervalo de valores alrededor de la estimación puntual.

El intervalo se forma con los valores que se creen que son los más probables para el valor desconocido, en base a los datos observados.

Se construye de manera de contener al parámetro con cierta probabilidad elegida previamente, la cual se indica con $1-\alpha$.

Dado que el intervalo contiene al parámetro con cierto grado de confianza, se denominan *intervalos de confianza*.

# Estimación por intervalo de confianza

- Un **intervalo de confianza** (**IC**) es un intervalo que contiene los valores más probables para un determinado parámetro.
- La probabilidad, <u>antes del muestreo</u>, de que el intervalo contenga al parámetro se denomina **coeficiente de confianza**. Es un valor que se elige cercano a 1, frecuentemente 0,95 pero puede elegirse 0,90 o 0,99.

*¿Cómo se construye un IC?*

Tomaremos como punto de partida la distribución en el muestreo de las estadísticas.

# Distribuciones en el muestreo

- La **distribución en el muestreo** de una estadística especifica los valores posibles que la estadística puede tomar y con qué probabilidad. Permite conocer la probabilidad de que la estimación puntual se encuentre a determinada distancia del parámetro.

La **desviación estándar de una distribución muestral** describe la variabilidad en los valores posibles de una estadística para un determinado tamaño de muestra; indica cuánto se espera que los valores de dicha estadística varíen de muestra en muestra (para un mismo *n*).
Se la denomina **error estándar** de la estadística.

# Distribuciones en el muestreo

- de la proporción muestral
- de la media muestral
- del desvío estándar de una distribución Normal.

# Distribución en el muestreo de la proporción muestral

Sea *p* la proporción de veces que un determinado suceso ocurre, es decir la proporción de "éxitos".
Si una muestra simple al azar de tamaño *n* se extrae de una población donde la proporción de éxito es *p* y *n* es grande, entonces:

$$\hat{p} \underset{n\to\infty}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

# Distribución en el muestreo de la proporción muestral

La distribución es aproximadamente Normal para muestras grandes, con $np \geq 15$ y $n(1-p) \geq 15$.

Es decir, que se deberían tener al menos 15 éxitos y 15 fracasos del resultado binario de interés.

En algunos textos se indica 5 o 10 en lugar de 15 pero algunas investigaciones indican que son muy bajos esos valores.

(L. Brown et al., *Statistical Science*, vol. 16, pp 101-113, 2001)

# Distribución en el muestreo de la media muestral

Se extrae una muestra simple al azar de tamaño *n* de una población de media *μ* y desvío estándar *s*, entonces:

- Si la población original tiene distribución Normal:

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

- Si la población original no tiene distribución Normal, pero *n* es lo suficientemente grande, entonces la distribución es aproximadamente Normal (TCL):

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

# Distribución en el muestreo del desvío estándar de una distribución Normal

Se extrae una muestra simple al azar de tamaño *n* de una población Normal con desvío estándar $s$, entonces:

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

# Construcción de los intervalos de confianza

Una forma de determinar un IC para un parámetro $\theta$ es considerar una *estadística pivote* que sea función de los valores de la muestra y del parámetro de interés.

Sin perder generalidad, supongamos que $\hat{\theta}$ es una estadística con una distribución en el muestreo aproximadamente Normal para muestras grandes con $E(\hat{\theta}) = \theta$ y error estándar $\sigma_{\hat{\theta}}$ .

Luego, la estadística pivote queda definida como:

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0,1)$$

# Construcción de los intervalos de confianza

Se necesita una expresión de probabilidad para la estadística pivote. Si se trabaja con un *coeficiente de confianza (1-$\alpha$)*, se localiza el valor $z_{\alpha/2}$ (ubica una probabilidad igual a $\alpha/2$ en la distribución de Z). Por la simetría en dicha distribución se tiene que:

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

Trabajando sobre esta desigualdad se obtiene que:

$$P\left(\underbrace{\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}}_{\downarrow} \leq \theta \leq \underbrace{\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}}_{\downarrow}\right) = 1 - \alpha$$

Límite de confianza inferior          Límite de confianza superior

# Estimación por intervalo de confianza para una proporción

Un IC para una proporción poblacional *p*, utilizando la proporción muestral $\hat{p}$ como estadística, para un tamaño muestral *n* lo suficientemente grande, viene dado por:

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

En este caso se asume:
la distribución es aproximadamente Normal para muestras grandes, con $np \geq 15$ y $n(1-p) \geq 15$.

*Desarrollo en clase.*

# Estimación por intervalo de confianza

- El **margen de error** mide con qué precisión la estimación puntual estima al parámetro. Se calcula como un múltiplo del error estándar del estimador.

- Si se construye un IC, el margen de error es su semiamplitud.

# Estimación por intervalo de confianza

**Efecto del tamaño muestral y del coeficiente de confianza sobre la amplitud del IC.**

Esperamos que la estimación sea más precisa con un tamaño de muestra mayor; con más datos, se tiene más información acerca de la población.

El margen de error en el caso de la proporción es:

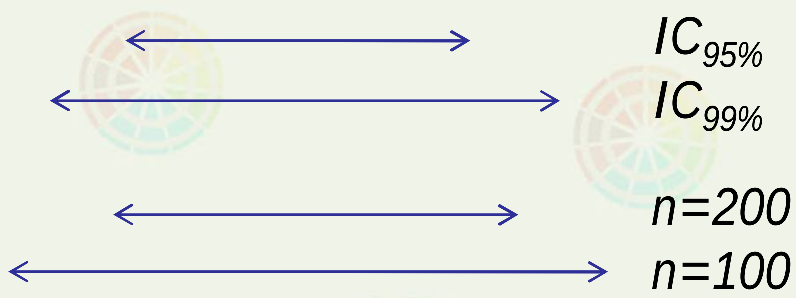
$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

# Estimación por intervalo de confianza

Este margen:

- Disminuye a medida que *n* se incrementa; cuanto mayor sea *n*, más angosto será el intervalo para un valor dado de la estadística y del coeficiente de confianza.

- Disminuye a medida que el coeficiente de confianza *(1-$\alpha$)* también disminuye; para un valor fijo de *n* y de la estadística.




*Esto ocurre para cualquier IC, no sólo el de la proporción.*

# Estimación por intervalo de confianza

**Interpretación del coeficiente de confianza.**

El significado del *(1-$\alpha$)%* de confianza hace referencia a cómo funciona el método cuando es utilizado una y otra vez con muchas muestras aleatorias diferentes.

Si utilizáramos el estimador de intervalo repetidamente, el *(1-$\alpha$)%* de los intervalos formados contendrían al parámetro de interés, es decir que darían buenos resultados.

Esto ocurre porque el *(1-$\alpha$)%* de las proporciones muestrales, por ejemplo, se encontrarían dentro de $z_{\alpha/2}$*(error estándar) de la proporción muestral.

# Estimación por intervalo de confianza

**Interpretación del coeficiente de confianza.**

Un intervalo particular, por ejemplo

$IC_{\mu;(1-\alpha)\%}$: (límite inf.; límite sup),

o contiene a $\mu$ o no lo contiene.

La probabilidad de que $\mu$ pertenezca al intervalo es 0 o 1; sin embargo, tenemos un $(1-\alpha)\%$ de confianza de que sí pertenezca porque se construye el intervalo con un método tal que si se extraen reiteradas muestras, todas de igual tamaño y para cada una de ellas se construye el mencionado IC, es de esperar que el $(1-\alpha)\%$ de los intervalos cubran a $\mu$.

# Estimación por intervalo de confianza

**Interpretación del coeficiente de confianza.**

En la práctica, el coeficiente de confianza más habitual es el del 95%, pero algunas aplicaciones requieren de mayor confianza.

Por ejemplo, en investigaciones médicas.

Tal como se comentó, para incrementar la probabilidad de que la inferencia sea correcta, se puede utilizar un coeficiente de confianza mayor, tal como 99%.

Se debe encontrar un equilibrio entre la precisión deseada y la confianza deseada de obtener una inferencia correcta; cuando una mejora, la otra empeora.

# Estimación por intervalo de confianza para una media ($\mu$) con s conocido

Un IC para media poblacional $\mu$, utilizando la media muestral $\overline{X}$ como estadística, para un tamaño muestral *n* lo suficientemente grande y **<u>cuando s es conocido</u>**, viene dado por:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Supuestos: si *n* es lo suficientemente grande, ninguno. En otro caso debe ser $X \sim N(\mu, s)$.

*Desarrollo en clase.*

# Estimación por intervalo de confianza para una media ($\mu$) con s desconocido.

Un IC para media poblacional $\mu$, utilizando la media muestral $\overline{X}$ como estadística, para un tamaño muestral *n* lo suficientemente grande y **cuando s es desconocido**, viene dado por:

$$\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}$$

Supuestos: si *n* es lo suficientemente grande, ninguno. En otro caso debe ser $X \sim N(\mu, s)$.

*Desarrollo en clase.*

# Estimación por intervalo de confianza para la variancia ($s^2$) de una población Normal

Un IC para la variancia poblacional $s^2$ de una distribución Normal, viene dado por:

$$\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}};\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}$$

*Desarrollo en clase.*

# Estimación por intervalo de confianza

**En relación al cumplimiento de los supuestos**

Los IC para la media que utilizan la distribución *t* son robustos (es decir, se desempeñan adecuadamente) frente a la violación del supuesto de distribución Normal.

El caso más importante en el cual el IC basado en la distribución *t* no funciona bien es cuando los datos contienen *outliers*. Por lo tanto, es importante verificar que no haya *outliers* que puedan afectar la validez de la media o de su IC.

# Determinación del tamaño muestral *n*

El tamaño muestral depende de cuánta precisión se requiere en la estimación, medida a través del margen de error.

*A menor margen de error, mayor será el n necesario*
*(para un coeficiente de confianza dado)*

A modo de ejemplo veremos el cálculo para el caso de una proporción y se propone como ejercitación trabajar sobre los restantes casos.

# Determinación del tamaño muestral *n*

## en el caso de la estimación de una proporción

- Se debe decidir el margen de error deseado, es decir, cuán cercano se desea que se encuentre al proporción muestral del valor poblacional.
- Se debe elegir el coeficiente de confianza para alcanzar dicho margen de error.

Sea *m* el margen de error: $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

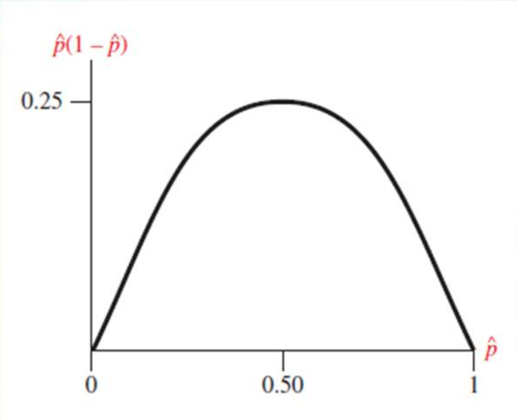
# Determinación del tamaño muestral *n*

## en el caso de la estimación de una proporción

Entonces, el mínimo tamaño de muestra requerido para el cual el $IC_{p;(1-\alpha)\%}$ tiene margen de error *m* es:

$$n = \frac{\hat{p}(1-\hat{p})z^2_{\alpha/2}}{m^2}$$

Si no se tiene información acerca de cuál puede ser el valor de la proporción estimada (que también depende de *n*), se puede considerar el valor más conservador: 0,50.



# Determinación del tamaño muestral *n*

## en el caso de la estimación de una media

**Ejercicio**

- ¿Cuál es el tamaño de la muestra aleatoria para el cual el $IC_{\mu;(1-\alpha)\%}$ tiene margen de error $m$?

# Determinación del tamaño muestral *n*

Tal como se observó, varios factores tienen implicancia en la determinación del tamaño muestral:

- La precisión deseada, medida a través del margen de error.
- El coeficiente de confianza.
- La variabilidad de los datos. Cuanto mayor sea el valor esperado para la desviación estándar s, mayor será el tamaño muestral *n* .

# Contrastes de hipótesis

# Contrastes de hipótesis

Junto con la estimación por IC, los contrastes de hipótesis son un método para realizar inferencia estadística acerca de una población.

Ambos métodos utilizan la probabilidad para proveer una forma de cuantificar cuán posibles son los valores de un parámetro mientras se controla la probabilidad de una inferencia incorrecta.

# Contrastes de hipótesis

El principal objetivo de muchas investigaciones es evaluar si los datos soportan o no ciertas afirmaciones o conjeturas.

Estas afirmaciones son hipótesis acerca de una población y en general están expresadas en términos de parámetros poblacionales para las variables medidas en el estudio.

(*En la Unidad 1 se presentaron algunos conceptos básicos en relación a este tema*)

# Contrastes de hipótesis

## Algunas definiciones…

**Hipótesis Estadísticas:**

$H_0$) **Hipótesis nula**: es la afirmación de que nada está sucediendo, que no existe diferencia, que no hay cambios en la población.
$H_1$) **Hipótesis alternativa**: es la afirmación que el investigador espera que sea cierta; representa el cambio en la población que el investigador está buscando.

# Contrastes de hipótesis

## Algunas definiciones…

**Error tipo I ($e_I$)**: error que se comete cuando se rechaza la Ho siendo cierta.

**Error tipo II ($e_{II}$)**: error que se comete cuando no se rechaza la $H_0$ siendo cierta la $H_1$.

| Decisión basada en los datos | $H_0$ es cierta | $H_1$ es cierta |
|---|---|---|
| No rechazar $H_0$ | No hay error | $e_{II}$ |
| Rechazar $H_0$ | $e_I$ | No hay error |

# Contrastes de hipótesis

## Algunas definiciones…

- **Región de rechazo** o **región crítica**: conjunto de valores para los cuales rechazaríamos $H_0$ (valores que son contradictorios a la $H_0$ y favorables a la $H_1$).
- **Región de no rechazo**: conjunto de valores para los cuales no rechazaríamos la $H_0$.
- **Valor crítico**: aquel que marca el "punto inicial" del conjunto de valores que comprenden la región de rechazo.

# Contrastes de hipótesis

## Algunas definiciones…

- **Nivel de significación**: un valor $0 < \alpha < 1$, tal que:

$$P(\text{rechazar } H_0 / H_0 \text{ es cierta}) \leq \alpha.$$

- **Potencia del test**: el valor $(1- \beta)$ siendo $\beta=P(e_{II})$.
- **Probabilidad asociada** a un resultado (*p*): probabilidad de obtener el resultado observado o uno más extremo (en dirección a la $H_1$), suponiendo que la $H_0$ es la verdadera.
  - Si $\mathbf{p} \leq \boldsymbol{\alpha} \Rightarrow$ se **rechaza** $H_0$; los resultados *son estadísticamente significativos*.
  - Si $\mathbf{p} > \boldsymbol{\alpha} \Rightarrow$ no se **rechaza** $H_0$; los resultados *no son estadísticamente significativos*.

# Pasos de los contrastes de hipótesis

En la construcción de un test de hipótesis pueden considerarse los siguientes pasos:

1. **Supuestos**: en cada test se realizan ciertos supuestos o se imponen ciertas condiciones bajo las cuales éste se aplica. En primer lugar, se supone que los datos fueron obtenidos mediante un procedimiento aleatorio; luego, puede ser necesario hacer supuestos acerca del tamaño de la muestra o acerca de la distribución de la población.

2. **Hipótesis**: cada test tiene dos hipótesis referidas al parámetro poblacional (nula y alternativa).

# Pasos de los contrastes de hipótesis

3. **Estadística de prueba**: el parámetro al cual hacen referencia las hipótesis tiene un estimador puntual y éste una determinada distribución en el muestreo. Una forma de construir los test de hipótesis es considerando dicha distribución. Así, puede describirse cuán lejos la estimación puntual se encuentra del valor del parámetro dado en la hipótesis nula.

4. **Probabilidad asociada, *p***: para interpretar el valor de la estadística del test, puede utilizarse una medida de probabilidad que resume la fuerza de la evidencia contra la $H_0$.

# Pasos de los contrastes de hipótesis

4. **(*cont.*) P**ara calcularla, se asume que $H_0$ es cierta, de manera que "la carga de la prueba" recae sobre la $H_{1.}$ Luego, se consideran los valores que se esperarían obtener para la estadística del test, de acuerdo a su distribución en el muestreo (asumiendo $H_0$ cierta).
   Si la estadística del test se encuentra lejos en la cola de la distribución quiere decir que si $H_0$ fuera cierta, tal valor sería inusual.
   Se resume cuán lejos en la cola cae la estadística del test a través del valor de la probabilidad asociada *p*.
5. **Conclusión**: se informa y se interpreta el valor de *p* en el contexto del estudio; se toma una decisión acerca de $H_0$**.**

# Contraste para una proporción

- Se cuenta con datos provenientes de una m.s.a. de tamaño *n*, lo suficientemente grande.
- Se plantea: $H_0)\ p=p_0$ *vs* $H_1)\ p>p_0$ o bien
  $H_1)\ p<p_0$ o bien $H_1)\ p\neq p_0$
- La estadística de prueba se define como:

$$Z=\frac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}\underset{H_0}{\sim} N(0,1)$$

- Si $p \leq \alpha \Rightarrow$ se rechaza $H_0$.

*Desarrollo en clase.*

# Contraste para una media con s conocido

- Se cuenta con datos provenientes de una m.s.a. de tamaño *n*, de una población normal con desvío estándar conocido. El supuesto de distribución normal no es crucial si *n* es lo suficientemente grande (TCL).
- Se plantea: $H_0)\ \mu=\mu_0$ *vs* $H_1)\ \mu>\mu_0$ o bien
  $H_1)\ \mu<\mu_0$ o bien $H_1)\ \mu\neq\mu_0$
- La estadística de prueba se define como:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \underset{H_0}{\sim} N(0,1)$$

- Si $p \leq \alpha \Rightarrow$ se rechaza $H_0$.

*Desarrollo en clase.*

# Contraste para una media con s desconocido

- Se cuenta con datos provenientes de una m.s.a. de tamaño *n*, de una población normal con desvío estándar desconocido. El supuesto de distribución normal no es crucial si *n* es lo suficientemente grande (TCL).
- Se plantea: $H_0)\ \mu=\mu_0$ *vs* $H_1)\ \mu>\mu_0$ o bien
  $H_1)\ \mu<\mu_0$ o bien $H_1)\ \mu\neq\mu_0$
- La estadística de prueba se define como:

$$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\underset{H_0}{\sim}t_{n-1}$$

- Si $p \leq \alpha \Rightarrow$ se rechaza $H_0$.

*Desarrollo en clase.*

# Contraste para la variancia de una población normal

- Se plantea: $H_0)\ \sigma^2 = \sigma^2_0$ *vs* $H_1)\ \sigma^2 > \sigma^2_0$ o bien
  $H_1)\ \sigma^2 < \sigma^2_0$ o bien $H_1)\ \sigma^2 \neq \sigma^2_0$

- La estadística de prueba se define como:

$$Y = \frac{(n-1)S^2}{\sigma_0^2} \underset{H_0}{\sim} \chi^2_{n-1}$$

- Si $p \leq \alpha \Rightarrow$ se rechaza $H_0$.

*Desarrollo en clase.*

# Relación entre contrastes de hipótesis e IC

Las conclusiones acerca de los parámetros utilizando test bilaterales son consistentes con las obtenidas a través de los IC.

Por ejemplo, en un test bilateral los valores hipotéticos de $\mu_0$ que no serían rechazados serían:

$$\bar{x} - z_{1-\alpha/2}\sigma / \sqrt{n} < \mu_0 < \bar{x} + z_{1-\alpha/2}\sigma / \sqrt{n} \qquad (*)$$

Es decir que los valores de $\mu_0$ que pertenecen a la "región de no rechazo" son los mismos que el $IC_{\mu;\,(1-\alpha)\%}$.

Uno puede llevar a cabo un test $H_0$) $\mu=\mu_0$ vs. $H_0$) $\mu\neq\mu_0$ calculando el intervalo de confianza y rechazando $H_0$ si el intervalo no contiene a $\mu_0$.

(*) *Demostrar.*

# Comentarios generales…

# Bibliografía


- Agresti A., Franklin C. (2009) The art and science of learning from data. 2º Ed. New Jersey. Pearson Prentice Hall.

- Ruggieri M. (2010) Métodos estadísticos I. Reimpresión. Rosario. UNR editora.

- Mendenhall W., Sincich T. (1997) Probabilidad y estadística para ingeniería y ciencias. Pearson Educación.