Elementos de Probabilidad y Estadística

Diplomatura en Ciencias de Datos

2024

Entrega 3

Dr. Matías Hisgen – Lic. Celine Cabás – Lic. Fernando Álvarez

FACENA - UNNE

Media poblacional

- Anteriormente definimos a μ_y como la *media* poblacional de la variable aleatoria *y*. Dicha *media* puede ser vista como el *Valor Esperado* o *Esperanza* de *y:*

$$E(y) = \mu_y$$

Así, es posible escribir *y* como:

$$y = \mu_y + u,$$

en donde $u = (y - \mu_y)$ son las desviaciones respecto de la media.

Estimación: Media poblacional

- Dada una muestra aleatoria de tamaño n de la población $\{(y_i)$: $i=1, \ldots, n\}$, podemos escribir cada observación de la muestra como

$$y_i = \mu_y + u_i$$

La idea básica es estimar el parámetro poblacional μ_y usando la muestra, para obtener

$$y_i = \hat{\mu}_y + \hat{u}_i$$

Mínimos Cuadrados Ordinarios

- El residuo $\hat{u}_i$ es un estimador del término de error u_i y es la diferencia entre la media estimada y la *i-esima* observación muestral.
- Intuitivamente, MCO consiste en seleccionar un valor del estimador de tal forma que la suma de los residuos ($\hat{u}_i$) elevados al cuadrado sea tan pequeña como fuese posible, de allí el término “mínimos cuadrados”

El problema de minimización

- Dada la idea intuitiva, podemos establecer ahora un problema formal de minimización
- Esto es, queremos elegir los parámetros de tal forma que se minimice la siguiente expresión:

$$\sum_{i=1}^{n}(\hat{u}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\mu}_y\right)^2$$

El problema de minimización

- Resolviendo el problema de minimización para el único parámetro, obtenemos la condición de primer orden:

$\sum_{i=1}^{n}(y_i - \hat{\mu}_y) = 0,$ que es igual a

$\sum_{i=1}^{n} \hat{u}_i = 0,$ la suma de residuos es cero en la muestra muestra

Estimador MCO: media muestral

- Dada las propiedades de la sumatoria, podemos reescribir la primera condición para obtener el estimador de la media poblacional:

$\sum_{i=1}^{n}(y_i) - n\,\hat{\mu}_y = 0,$ quedando

$\sum_{i=1}^{n}(y_i)/n = y = \hat{\mu}_y,$ el promedio muestral

Insesgamiento

El estimador "promedio muestral" es ***insesgado***, en ***muestras repetidas***:

$$E\left(\hat{\mu}_y\right) = E\left[\sum_{i=1}^{n} (y_i)/n\right] = \mu_y,$$

Recordar que insesgamiento es una propiedad del ***estimador*** – en una muestra dada podemos estar "cerca" o "lejos" del verdadero valor del parámetro.

Varianza de la media muestral

Para resumir el error de estimación, la varianza del estimador es:

$$Var(\hat{\mu}_y) = Var\left[\sum_{i=1}^{n} (y_i)/n\right] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

$$= {}^{n\,\sigma^2}/_{n^2} = {}^{\sigma^2}/_{n}$$

Componentes de la Varianza

- A mayor varianza de *y*, σ^2, mayor varianza del estimador.
- Un mayor tamaño de la muestra hace disminuir la varianza del estimador de la pendiente
- Problema: σ^2 es desconocida.
- Podemos estimarla con la varianza muestral S_y^2

Error estándar de la media muestral

- Sustituyendo μ_y por su estimador y tomando raíz cuadrada tenemos el *error estándar*

$$ee(\bar{y}) = \sqrt{\frac{{S_y}^2}{n}} = \sqrt{\frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)/(n-1)}{n}} = \frac{S_y}{\sqrt{n}}$$

Es fácil notar que la variabilidad del estimador "media muestral" depende *directamente* de la varianza de *y* en la muestra, e *inversamente* del tamaño muestral.

Intervalo de Confianza para la Media

- La media muestral es un estadístico descriptivo, no un estimador de la media poblacional μ_y
- Un estimador de μ_y debe incorporar el error muestral de estimación, así surge el *intervalo de estimación* o *intervalo de confianza*.
- Tal *Intervalo de Confianza* (IC) contendrá a la media μ_y , en muestras repetidas, una proporción prefijada de veces. Dicha proporción se denomina *Nivel de Confianza* (NC).

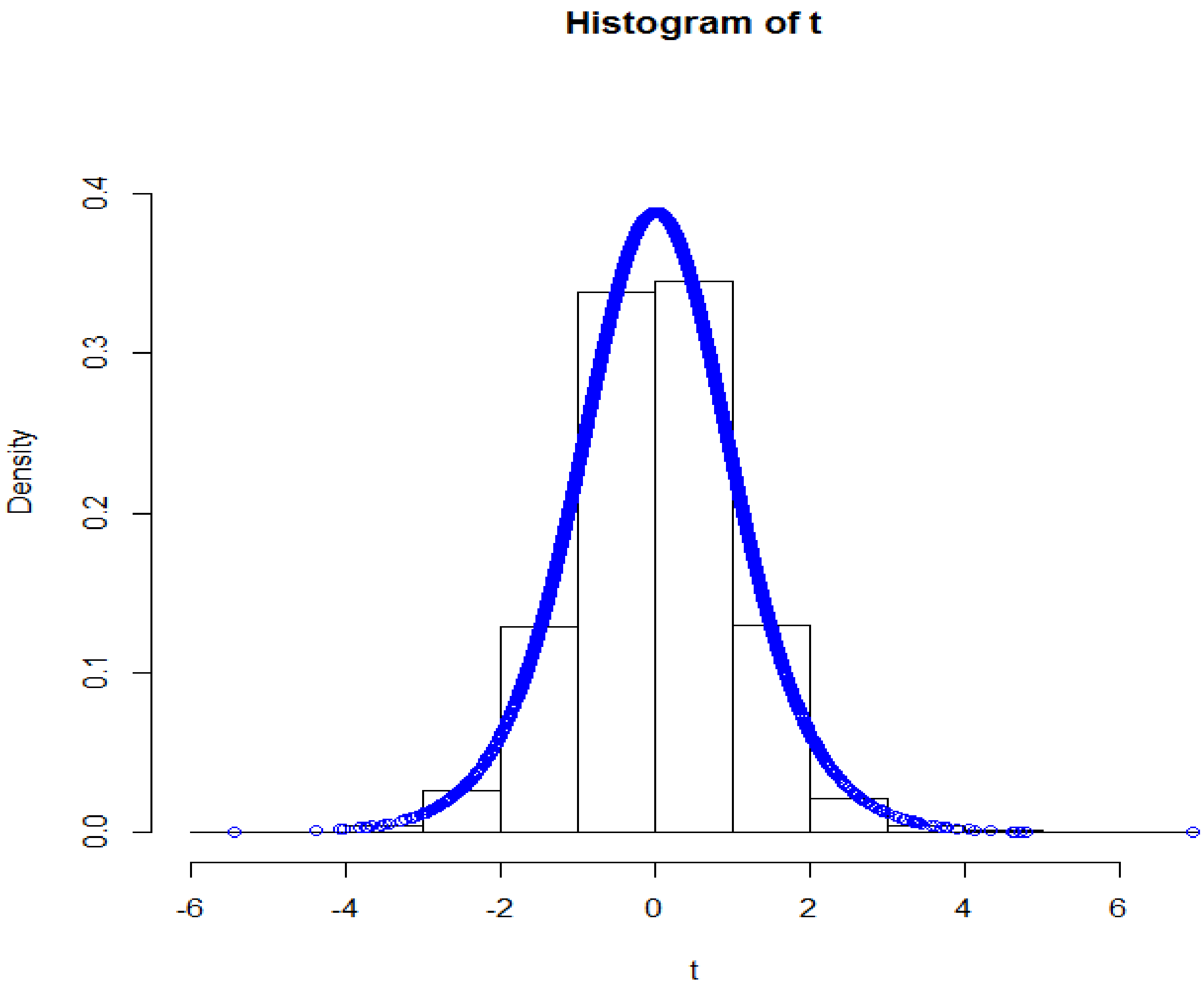
Intervalo de Conf. para la Media

- Para construir un intervalo de confianza (IC) para la media *poblacional*, necesitamos saber cómo la media *muestral* se distribuye en muestras repetidas.
- Si el error u se distribuye Normal(0, σ^2), se cumple que la media muestral estandarizada se distribuye como una t-Student con n-1 grados de libertad:

$$t = \frac{(\bar{y} - \mu_y)}{ee(\bar{y})} \approx t \text{ de } Student(n-1) = t_{n-1}$$

Gráficamente

- Histograma de *t* en 5000 muestras con n=10, junto con la densidad de una distribución t-Student(9)



Intervalo de Conf. para la Media (cont.)

- Ahora necesitamos definir el *Nivel de Confianza* (NC) deseado, que es la proporción de veces que queremos que la verdadera media μ_y quede *incluida dentro* del intervalo, en muestreo repetido.
- Lo llamamos NC = $(1 - \alpha)$, siendo α la proporción de veces que la media μ_y quedará *fuera* del intervalo.
- Ello equivale a definir un intervalo $(-t_c , t_c)$, de la distribución $t_{n\text{-}1}$, que contenga una proporción $(1 - \alpha)$ de todos los valores posibles de t en muestras repetidas. Llamamos *valor crítico* al t_c definido.

Intervalo de Conf. para la Media (cont. II)

- Una vez definido el NC y su t_c correspondiente, el IC para la media muestral estandarizada viene dado por

$$-t_c \leq \frac{(\bar{y}-\mu_y)}{ee(\bar{y})} \leq t_c$$

- Despejando, queda el IC para la media poblacional desconocida

$$\bar{y}-t_c.ee(\bar{y}) \leq \mu_y \leq \bar{y}+t_c.ee(\bar{y})$$

Pruebas de Hipótesis sobre la Media

- Los elementos definidos para construir un IC para μ_y se pueden usar para probar hipótesis sobre μ_y.
- Por ejemplo, podemos estar interesados en saber si μ_y es *mayor*, *menor* o solo *distinto* a un valor hipotetizado μ_0.
- Concretamente podemos definir 2 hipótesis:

<u>Hipótesis nula</u>: $H_0: \mu_y = \mu_0$ versus

<u>Hipótesis alternativa</u>, que pueden ser 3 casos:

$H_1: \mu_y \neq \mu_0$ o $H_1: \mu_y > \mu_0$ o $H_1: \mu_y < \mu_0$

Hipótesis Nula y estadístico de prueba

- Notar que si la nula, H_0: $\mu_y = \mu_0$, es cierta, entonces el estadístico de prueba *t* cumple con:

$$t = \frac{(\bar{y} - \mu_0)}{ee(\bar{y})} \approx t \text{ de } Student(n-1) = t_{n-1}$$

- De éste modo, dada una muestra, es posible calcular el estadístico *t* previo, que implica que H_0 es verdadera, y compararlo con la distribución teórica que debería poseer, la *t* de *Student*(*n*-1), para saber si su valor es *plausible* o NO lo es.

Prueba bilateral

- La prueba bilateral viene dada por

$$H_0: \mu_y = \mu_0 \text{ versus } H_1: \mu_y \neq \mu_0 ,$$

- Establecida H_0 , con la muestra se calcula el estadístico de prueba *t*:

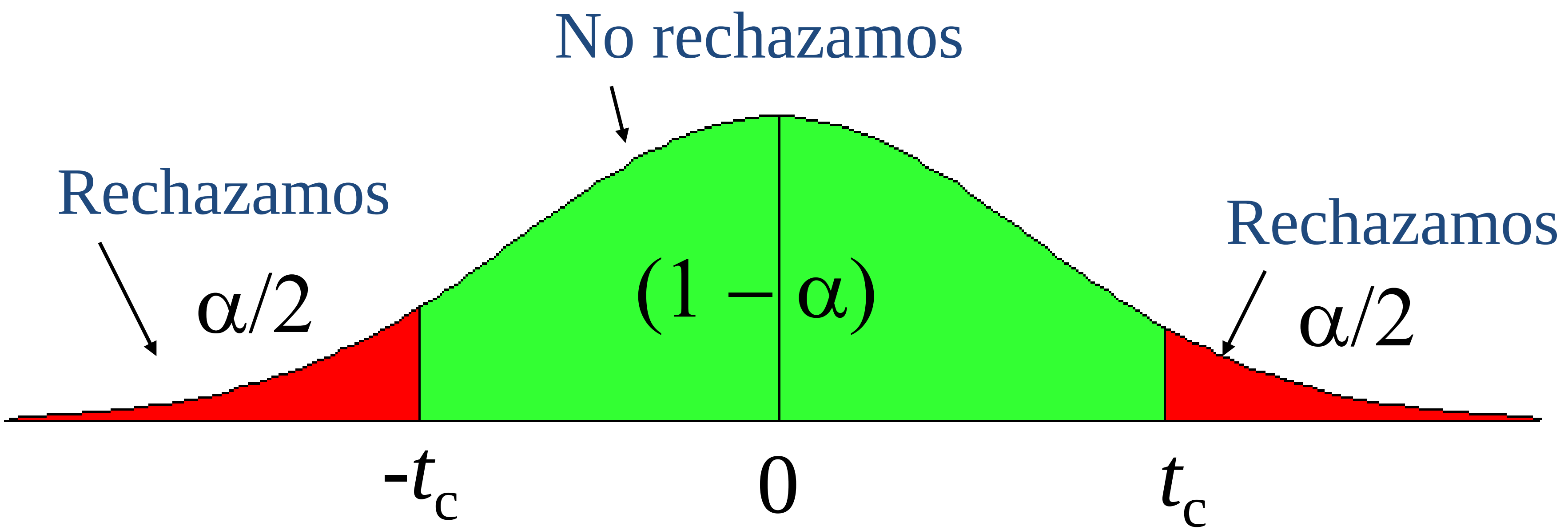
$$t = \frac{(\bar{y} - \mu_0)}{ee(\bar{y})}$$

- Es necesario definir un Nivel de Confianza (NC) para la prueba. En éste caso, el NC será la proporción (1- α) de veces que la prueba ***NO*** *RECHAZARÁ* H_0 siendo ésta *VERDADERA*.

Prueba bilateral II

Definido un NC, y por tanto un valor crítico t_c , la regla de rechazo de la prueba es:

Si $|t| > t_c$ se Rechaza H_0 y se Acepta H_1



Interpretación

- Es importante notar que el rechazo de H_0 implica que existe *suficiente evidencia muestral en contra* de H_0 , por lo que se puede aceptar H_1.

- En cambio, si No se Rechaza H_0 , no se puede decir que *hay suficiente evidencia a favor* de H_0, por eso está mal decir que Aceptamos H_0. Solo podemos decir que no la podemos rechazar a favor de H_1.

Interpretación II

- H_0 sería como la "presunción de inocencia" en un juicio. Entonces, solo se declara "culpable" (es decir, se acepta H_1) si se reúne la evidencia suficiente para culparlo.

- En este sentido, una prueba clásica solo aporta evidencia empírica de una hipótesis (H_1) cuando se rechaza H_0. Por ello, en H_1 se debe plantear la hipótesis que nos interese poner a prueba.

Prueba sobre diferencia de medias

- Para probar si 2 medias (de 2 poblaciones o grupos) son *diferentes*, las hipótesis son:

$$H_0: \mu^1{}_y = \mu^2{}_y \text{ vs. } H_1: \mu^1{}_y \neq \mu^2{}_y,$$

o $$H_0: \mu^1{}_y - \mu^2{}_y = 0 \text{ vs. } H_1: \mu^1{}_y - \mu^2{}_y \neq 0,$$

- Con las dos muestras se calculan las dos medias muestrales y el estadístico de prueba t

$$t = \frac{(\bar{y}^1 - \bar{y}^2) - 0}{ee(\bar{y}^1 - \bar{y}^2 - 0)} = \frac{(\bar{y}^1 - \bar{y}^2)}{ee(\bar{y}^1 - \bar{y}^2)}$$

y se aplica la regla de rechazo de la prueba bilateral.

Cómputo de los *p*-valores

- Para no tener que hacer varias pruebas para probar diferentes NC (90%, 95%, 99%, etc), se calcula el *p-valor*, que implica calcular el *Máximo* NC al cual H_0 puede ser *Rechazada*.
- Entonces, calculamos el estadístico *t*, y luego miramos en qué *percentil* de la distribución *t-Student*(*n*-1) se encuentra.
- Así se tiene: p-valor = 1- [percentil / 100]
- El p-valor estará siempre entre 0 y 1. A *menor* p-valor, *mayor* NC al cual se puede rechazar H_0.

Cómputo de los *p*-valores

- Por ejemplo, un p-valor de 0,05 implica que se puede rechazar H_0 hasta al 95% de confianza.
- Un p-valor de 0,01 implica rechazar H_0 hasta con un 99% de confianza, etc.
- A menor p-valor mayor es la evidencia en contra de H_0 .
- Entonces, el *p*-valor se interpreta como la mínima probabilidad de cometer el *error* de "Rechazar H_0 cuando ésta es verdadera"