

# Análisis Cuantitativo I

## Pruebas de hipótesis - Distribución t

Carlos Cardona Andrade

Universidad del Rosario

28 de octubre

# Quiz

- Un fabricante de lámparas eléctricas está ensayando un nuevo método de producción que se considerará aceptable si las lámparas obtenidas por este método dan lugar a una población normal de duración media 2400 horas, con una desviación estándar igual a 300. Se toma una muestra de 100 lámparas producidas por este método y esta muestra tiene una duración media de 2320 horas. ¿Se puede aceptar la hipótesis de validez del nuevo proceso de fabricación con un riesgo igual o menor al 5 %?

# Repaso de la clase anterior

1. Se espera que una media muestra  $\bar{X}$  se aproxime a la media poblacional  $\mu$ .
2. El error estándar provee una medida de cuán bien una media muestral se aproxima a la media poblacional.
3. Para cuantificar nuestra inferencia sobre la población, comparemos la media muestra obtenida con la media poblacional de la cual tenemos una hipótesis mediante el cálculo del z-score.
4. El objetivo de la prueba de hipótesis es determinar si la diferencia obtenida entre los datos y la hipótesis es significativamente mayor a la que esperaríamos por azar.

# ¿De dónde provienen los valores objetivo?

- ▶ Los valores poblacionales u objetivos de las pruebas de hipótesis pueden venir de:
  1. Un parámetro poblacional conocido a partir de un grupo de comparación o de un censo poblacional (Representatividad de la muestra).
  2. Parámetros conocidos a partir de un período anterior.
  3. Una idea objetivo (La producción deseada en una empresa).

# El problema del z-score

- ▶ El principal inconveniente de usar el z-score para las pruebas de hipótesis es que su fórmula requiere más información de la que usualmente se tiene.
- ▶ Específicamente, la fórmula necesita que conozcamos la desviación estándar poblacional, la cual se necesita para calcular el error estándar.
- ▶ La paradoja es la siguiente: se quiere utilizar el z-score para inferir información sobre una población desconocida pero, se tiene que conocer la población antes de calcularlo.

## El error estándar estimado

- ▶ Cuando se expusieron las medidas de dispersión, se explicó la diferencia de calcular la desviación estándar poblacional y la muestral.

$$s_X = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

- ▶ Donde  $n - 1$  eran los grados de libertad.
- ▶ Ahora para calcular el error estándar, en vez de utilizar la desviación estándar poblacional, usamos la muestral.

$$s_{\bar{X}} = \frac{s_X}{\sqrt{n}}$$

- ▶ Cabe destacar el cambio de notación, como es un cálculo muestral, el *error estándar estimado* se denota con la letra  $s$  y no con  $\sigma$ .

## El estadístico t

- ▶ El error estándar estimado se puede sustituir en la fórmula del z-score. De ahí resulta el estadístico t:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

- ▶ Este estadístico se utiliza cuando el valor de  $\sigma$  es desconocido.
- ▶ La única diferencia entre el estadístico t y el z-score recae en el hecho que el primero utiliza la desviación estándar muestral mientras que, el segundo utiliza la poblacional.

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \quad z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

# Los grados de libertad

- ▶ Se sabe que el estadístico t es un reemplazo del z-score. La diferencia básica radica en que el primero utiliza la desviación estándar muestral ( $s$ ) mientras que, el segundo utiliza la desviación estándar poblacional ( $\sigma$ ).
- ▶ Para determinar qué tan bien el estadístico t se aproxima al z-score, debemos determinar qué tan bien se aproxima la desviación muestral a la poblacional.

$$s_X = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \quad \sigma_X = \sqrt{\frac{\sum (X - \bar{X})^2}{n}}$$

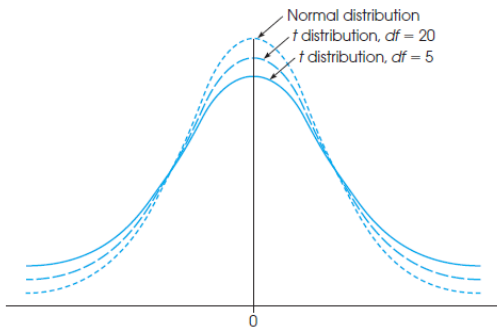
- ▶ Los grados de libertad describen el número de valores en una muestra que son independientes y libres de variar.



# La distribución t

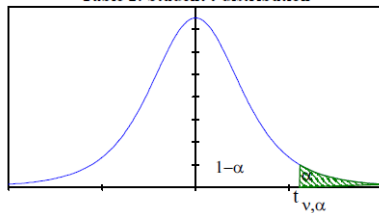
- ▶ Toda muestra de una población puede ser utilizada para calcular tanto el z-score como el t estadístico.
- ▶ Si se seleccionan todas las posibles muestra de tamaño  $n$  y se halla el z-score para cada media muestral, entonces el conjunto de valores z forma una distribución z.
- ▶ De igual manera, se pueden calcular todos los estadísticos t para cada muestra y el conjunto entero de valores t forma una distribución t.
- ▶ Se sabe que la distribución de los z-score tiende a ser una normal si  $n > 30$  o las muestras provienen de una población que se distribuye normal.

- ▶ Bajo esta misma situación, la distribución  $t$  se aproxima a una normal a medida que el estadístico  $t$  se aproxima a un  $z$ -score.
- ▶ Qué tan bien la distribución  $t$  se aproxima a una normal está determinado por los grados de libertad. En general, a medida que  $n$  aumenta,  $n - 1$  también aumenta haciendo que la distribución  $t$  se aproxime de mejor manera a la normal.



# La tabla de la distribución t

Table 2: Student t-distribution



$$\Pr(X > t_{v, \alpha}) = \alpha$$

$\nu$	$\alpha$				
	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977

# Innate and learned perceptual abilities in the newborn infant

- ▶ Los niños, incluso los recién nacidos, prefieren ver personas de caras atractivas.
- ▶ En el estudio de Slater *et al.* (1998), a bebés entre 1 y 6 días de nacidos se les muestra dos fotografías de mujeres.
- ▶ Previamente, un grupo de adultos realiza define uno de los rostros como más atractivo que el otro.
- ▶ Se ubica a los bebés enfrente de una pantalla que presenta ambos rostros durante 20 segundos.
- ▶ Asumamos que el estudio usa una muestra  $n = 9$  y obtienen una media de  $\bar{X} = 13$  segundos observando la cara atractiva con  $s_X = 3$  segundos.

# Prueba de Hipótesis

1. El primer paso consiste en establecer las hipótesis y seleccionar el nivel de significancia de la prueba ( $\alpha$ ). Aún cuando no se tiene información sobre la población, se puede definir una hipótesis lógica acerca del valor  $\mu$ .

Para este caso, la hipótesis nula afirma que no existe preferencia por ningún rostro. Por lo tanto deberían demorarse el mismo tiempo en ambos rostros.

$$H_0 : \mu_{atractivo} = 10$$

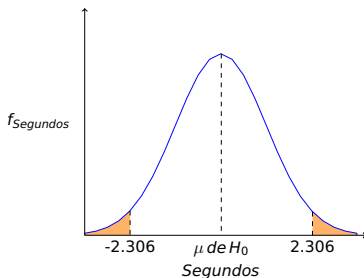
$$H_1 : \mu_{atractivo} \neq 10$$

$$\alpha = 5\%$$

2. El segundo paso es encontrar la región crítica. Para esto calculamos los grados de libertad:

$$n - 1 = 9 - 1 = 8$$

Al consultar la tabla de la distribución t, con un  $\alpha = 5\%$  encontramos que los valores críticos  $t_{\alpha} = \pm 2,306$



3. En el tercer paso se calcula el estadístico t. Para eso calculamos primero el error estándar estimado:

$$s_{\bar{X}} = \frac{s_X}{\sqrt{n}} = \frac{3}{\sqrt{9}} = \frac{3}{3} = 1$$

Luego de esto, se halla el estadístico t:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{13 - 10}{1} = 3$$

4. El último paso consiste en tomar una decisión sobre la  $H_0$ . Como  $t > t_{\alpha}$ , nuestra decisión es rechazar la hipótesis nula y se concluye que los bebés efectivamente muestran una preferencia hacia los rostros atractivos.

## Otro ejemplo

- Un dato revela que los universitarios consumen en promedio 8.8 cervezas al mes. Se toma una muestra de  $n = 28$  estudiantes de una universidad conservadora. Se sospecha que ahí consumen menos cervezas al mes. Probemos esta hipótesis con los siguientes datos:

$$\alpha = 0,1 \quad \bar{X} = 8,5 \quad s_x = 1,8 \text{ cervezas}$$



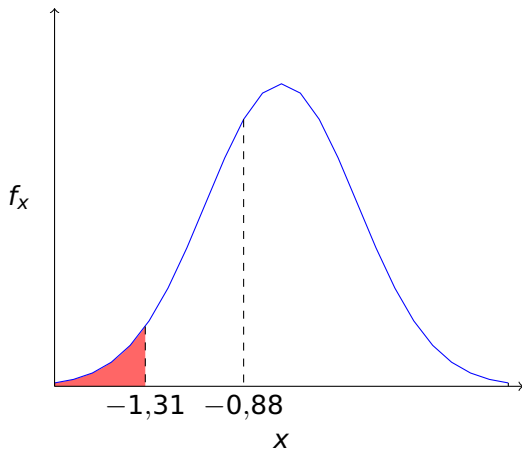
## Otro ejemplo

- Un dato revela que los universitarios consumen en promedio 8.8 cervezas al mes. Se toma una muestra de  $n = 28$  estudiantes de una universidad conservadora. Se sospecha que ahí consumen menos cervezas al mes. Probemos esta hipótesis con los siguientes datos:

$$\alpha = 0,1 \quad \bar{X} = 8,5 \quad s_x = 1,8 \text{ cervezas}$$

$$H_0 : \mu_{\text{muestra}} \geq 8,8 \quad H_1 : \mu_{\text{muestra}} < 8,8$$

$$s_{\bar{X}} = \frac{s_x}{\sqrt{n}} = \frac{1,8}{\sqrt{28}} = 0,34 \quad t = \frac{8,5 - 8,8}{0,34} = -0,88$$



$$t_{\alpha} = -1,31 \rightarrow -1,31 < -0,88 \rightarrow \text{Rechazo } H_0$$

# La representatividad de la muestra

- ▶ La representatividad de una muestra puede llevar a resultados erróneos.
- ▶ Por ejemplo, se quiere conocer la favorabilidad de la creación de un nuevo impuesto. Para estimar esta proporción, se encuestan a personas vía telefónica.
- ▶ Claramente, la encuesta excluye los hogares sin teléfonos, los cuales son habitados en su mayoría por personas con menores recursos.
- ▶ Las personas con mayores ingresos son más probables de poseer la casa, por lo tanto, se ven afectados en mayor medida por el impuesto.

- ▶ Queremos evaluar la representatividad de la encuesta con respecto a la proporción de habitantes bajo la línea de pobreza.
- ▶ Luego de realizar la encuesta, se identifican a 66 de las  $n = 387$  personas por debajo de la línea de pobreza.
- ▶ Es decir, la proporción de personas pobre sería:

$$p_s = \frac{66}{387} = 0,17$$

- ▶ Además, se tiene información del censo de la población, el cual indica que el 22 % de los habitantes son pobres.

# La prueba de hipótesis

- Se plantean las hipótesis:

$$H_0 : P_u = 22 \%$$

$$H_1 : P_u \neq 22 \%$$

$$\alpha = 0,05$$

$$\sigma_{P_S} = \sqrt{\frac{P_u * Q_u}{n}} = \sqrt{\frac{0,22 * 0,78}{387}} = 0,02$$

$$t_{P_S} = \frac{0,17 - 0,22}{0,02} = \frac{-0,05}{0,02} = -2,5 < -1,96 = t_{\alpha}$$

Rechazo  $H_0$  y concluyo que la muestra no es representativa.