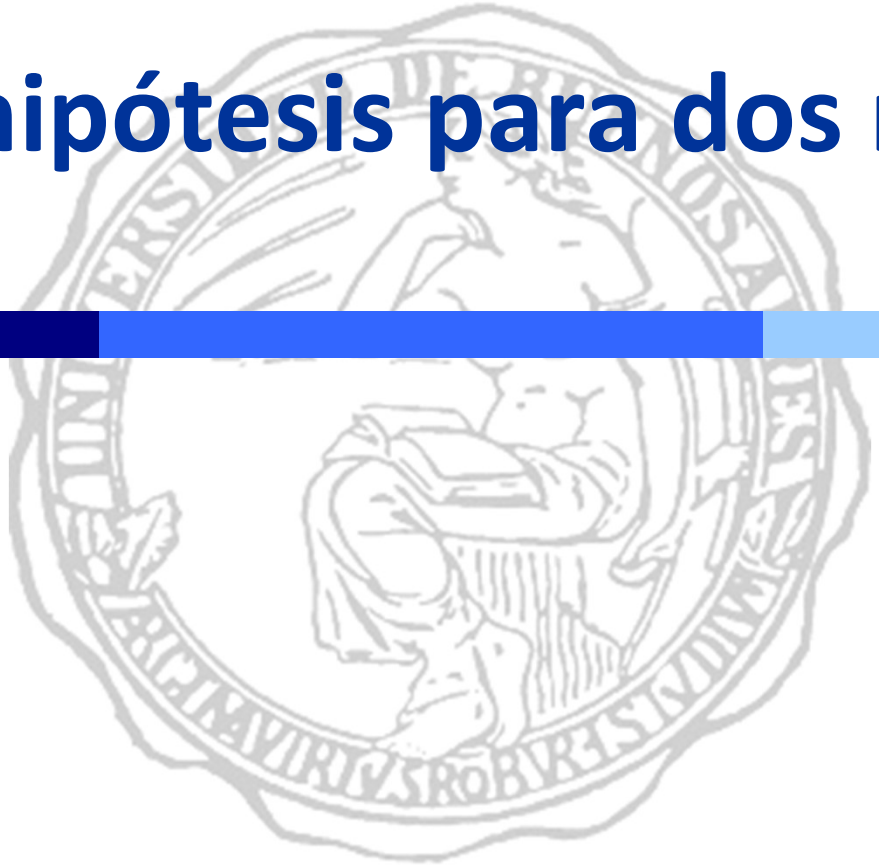
# Test de hipótesis para dos muestras



# Test de hipótesis para dos muestras

- Hasta ahora estuvimos trabajando con una muestra referenciada a una población. Ahora nos enfocaremos en dos poblaciones y por ende dos muestras.
- Pensemos en la materia de Estadística de la FCE UBA. Existe la creencia que la mujeres son en general más aplicadas que los hombres. Podríamos tomar una muestra de mujeres y una muestra de hombres que hayan cursado Estadística, y preguntarnos **si el rendimiento medio de las mujeres es superior al rendimiento medio de los hombres**.
- Es el m2 promedio del barrio de Palermo más caro que el m2 promedio de Recoleta?
- Es el tiempo de vida de las lámparas llamadas “larga vida” superior al de las lámparas comunes?

# Test para la diferencia de dos medias poblaciones, muestras grandes e independientes

- Se asume que se dispone de muestras independientes. La primera población tiene media $\mu_x$ y varianza $\sigma_x$ . El tamaño de la muestra aleatoria simple es $n_x$.
- La segunda población tiene media $\mu_y$ y varianza $\sigma_y$ . El tamaño de la muestra aleatoria simple es $n_y$.
- Sean $\bar{x}$ e $\bar{y}$ las medias muestrales de las respectivas muestras.
- Entonces $\bar{x}$ - $\bar{y}$ es el estimador puntual de $\mu_x$ - $\mu_y$.
- La distribución muestral de $\bar{x}$ - $\bar{y}$ tiene $E(\bar{x} - \bar{y}) = \mu_x$ - $\mu_y$ y

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

- Entonces **si ambas muestras son grandes en tamaño, la distribución muestral de $\bar{x}$ - $\bar{y}$ se aproxima con la normal**.

# Test para la diferencia de dos medias poblaciones, muestras grandes e independientes

- El estadístico es Z y tiene distribución límite normal

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

- El interés radica en evaluar si las medias son:

$H_0: \mu_x \leq \mu_y$ ó $\mu_x - \mu_y \leq 0$ vs. $H_1: \mu_x > \mu_y$ ó $\mu_x - \mu_y > 0$

$H_0: \mu_x \geq \mu_y$ ó $\mu_x - \mu_y \geq 0$ vs. $H_1: \mu_x < \mu_y$ ó $\mu_x - \mu_y < 0$

$H_0: \mu_x = \mu_y$ ó $\mu_x - \mu_y = 0$ vs. $H_1: \mu_x \neq \mu_y$ ó $\mu_x - \mu_y \neq 0$

# Test para la diferencia de dos medias poblaciones, muestras grandes e independientes

- Por varianzas poblacionales conocidas asumimos procesos que se han mantenido estables en el tiempo y hemos obtenido estimaciones similares de las varianzas para cada población a lo largo del tiempo.

- Tener presente que se asume muestras grandes.

# Test para la diferencia de 2 medias poblacionales: muestras independientes (y grandes): regla de decisión

- $H_0: \mu_x - \mu_y \leq 0 \;\; vs. \;\; H_1: \mu_x - \mu_y > 0$

Rechazo $H_0$ si $\dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > Z_\alpha$

- $H_0: \mu_x - \mu_y \geq 0 \;\; vs. \;\; H_1: \mu_x - \mu_y < 0$

Rechazo $H_0$ si $\dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < -Z_\alpha$

- $H_0: \mu_x - \mu_y = 0 \;\; vs. \;\; H_1: \mu_x - \mu_y \neq 0$

Rechazo $H_0$ si $\dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} < -Z_{\alpha/2}$ o $\dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} > Z_{\alpha/2}$

# Ejemplo

- El equipo de diseño de una red social tiene como objetivo incrementar el tiempo que los usuarios pasan en ella ya que los ingresos dependen de la cantidad de publicidad que les puedan mostrar. Con el fin de incrementar sus ingresos, deciden evaluar un cambio en la interfaz gráfica de la aplicación.
- Del total de los usuarios activos, se toman dos muestras al azar de perfiles de usuarios similares, a los que se denomina grupo A y grupo B. En principio no informan cuales de los grupos ve la interfaz nueva.
- En el grupo A se asignaron 189 usuarios que arrojaron un tiempo de conexión a la red social promedio de 2580 minutos con un desvió estándar 680 minutos, mientras que el grupo B, conformado por 203 usuarios arrojó un tiempo promedio de conexión de 2708 minutos y desvío estándar de 730 minutos.

# Ejemplo

- Ud. está a cargo de la investigación y es el responsable de tomar la decisión de implementar o no el cambio en la interfaz propuesto por el equipo de diseño.
- Especifique los supuestos que deberían cumplirse. Plantee el test de hipótesis adecuado para el objetivo de la red social y determine la regla de decisión.
- ¿Cuál es el valor  p asociado al estadístico?
- ¿Cuál sería su conclusión si deciden trabajar al 5% se significación?

# Dos medias, poblaciones independientes, varianzas desconocidas que se supone iguales

- En los casos en que no se conocen las varianzas poblacionales y el tamaño de muestra es pequeño (inferior a 100), se debe utilizar la ***t*** de ***student.***
- Por lo tanto debemos asumir que ambas poblaciones son normales
- Hay problemas teóricos si las varianzas poblacionales son distintas.
- Pero **si asumimos la misma varianza para ambas poblaciones**, un estimador agrupado común de la varianza poblacional a partir de las varianzas muestrales $s_x^2$ y $s_y^2$, es :

$$S_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

# Dos medias, poblaciones independientes, varianzas desconocidas que se supone iguales

- El estadístico para el caso de muestra pequeña es

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$$

- Y la distribución ***t*** tendrá distribución ***t de student*** con $\boldsymbol{n_x + n_y - 2}$ grados de libertad
- Tener presente que $\boldsymbol{s_p^2}$ es la media ponderada de las varianzas muestrales $\boldsymbol{s_x^2}$ y $\boldsymbol{s_y^2}$.
- Comprobar que si $\boldsymbol{n_x = n_y}$ entonces $s_p^2 = \frac{s_x^2 + s_y^2}{2}$

# Dos medias, poblaciones independientes, varianzas desconocidas que se supone iguales : regla de decisión

- $H_0: \mu_x - \mu_y \leq 0$ vs. $H_1: \mu_x - \mu_y > 0$

Rechazo $H_0$ si $t = \dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\dfrac{s_p^2}{n_x}+\dfrac{s_p^2}{n_y}}} > t_{n_x+n_y-2,\alpha}$

- $H_0: \mu_x - \mu_y \geq 0$ vs. $H_1: \mu_x - \mu_y < 0$

Rechazo $H_0$ si $t = \dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\dfrac{s_p^2}{n_x}+\dfrac{s_p^2}{n_y}}} < -t_{n_x+n_y-2,\alpha}$

# Dos medias, poblaciones independientes, varianzas desconocidas que se supone iguales: regla de decisión

- $H_0: \mu_x - \mu_y = 0$ vs. $H_1: \mu_x - \mu_y \neq 0$

Rechazo $H_0$ si

$$t = \frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} < -t_{n_x+n_y-2,\alpha/2}$$

ó

$$t = \frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} > t_{n_x+n_y-2,\alpha/2}$$

# Ejemplo

- De acuerdo con los especialistas en marketing, más del 70% de la decisisión de compra se da cuando el cliente está frente al producto. De ahí que, las estrategias de ubicación de los productos en la góndola, son claves para el éxito de venta.
- Cierta empresa de venta de bebidas colas encarga un estudio de marketing para evaluar los beneficios que podría brindarle ubicar su producto en punta de góndola vs. la ubicación estándar que suele tener en la góndola.
- El experimento consiste en elegir una cadena de supermercados y estudiar la ventas semanales promedio para las distintas ubicaciones en góndola una semana del mes. La primera población son las ventas registradas, cuando la ubicación es la estándar en góndola, mientras que la segunda población son la ventas correspondientes a la ubicación en la punta de góndola.

# Ejemplo

- La tabla a continuación muestra las ventas semanales (en número de unidades) para diferentes ubicaciones en gondola

| Estándar | | | | | Punta de góndola | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 22 | 34 | 52 | 62 | 30 | 52 | 71 | 76 | 54 | 67 |
| 40 | 64 | 84 | 56 | 59 | 83 | 66 | 90 | 77 | 84 |

- Especifique los supuestos que debería asumir para poder realizar el test que evalua si las ventas promedio semanal de la bebida cola difiere por su ubicación en góndola?
- Cual sería su conclusión si decide trabajar al 1% se significación?

# Dos medias, datos apareados o pareados

- Supongan que se elige una muestra al azar de individuos y se les solicita que informen el tiempo destinado a la lectura y el tiempo destinado a mirar series/ películas, durante la última semana.
- La información brindada serán pares ordenados de (tiempo de lectura, tiempo de mirar series/pelis) para cada uno de los individuos de la muestra. Por esa razón se llaman **muestras apareadas/ pareadas**.
- La principal diferencia en este tipo de experimento con los planteados anteriormente radica en que se elimina una fuente de variación muestral.
- En los experimentos anteriores disponíamos de 2 muestras. En este caso, se tiene una sola muestra a la que se le hacen dos preguntas.
- Por lo tanto la variación muestral debida al diseño de muestra (dos muestras vs. una muestra) es menor en el caso de muestras apareadas. Por esa razón siempre que se pueda se prefiere el diseño de muestras apareadas.

# Dos medias, datos apareados o pareados

- El interés va a radicar en la diferencia $d = x - y \; para \; cada \; par \; (x, y)$.
- Supongamos que se dispone de una muestra de n pares de observaciones con media $\mu_x \; y \; \mu_y$.
- Sea $\mu_d = \mu_x - \mu_y$ la diferencia de la medias poblacionales.
- Sean $\bar{d} \; y \; S_{\bar{d}}$ la media muestral y la desviación típica muestral observadas de las **n** diferencias observadas $d_i = x_i - y_i$.
- El estadístico para muestras apareadas tiene una distribución ***t de student*** y es

$$t = \frac{\bar{d} - \mu_{\bar{d}}}{S_{\bar{d}} / \sqrt{n}}$$

donde $\bar{d} = \frac{\sum_i d_i}{n}$ y $S_{\bar{d}} = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$

# Dos medias, datos apareados o pareados: regla de decisión

- $H_0$: $\mu_d = 0$ o $\mu_d \leq 0$ vs. $H_1: \mu_d > 0$

Rechazo $H_0$ si $t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}/\sqrt{n}} > t_{n-1,\alpha}$

- $H_0: \mu_d = 0$ o $\mu_d \geq 0$ vs. $H_1: \mu_d < 0$

Rechazo $H_0$ si $t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}/\sqrt{n}} < -t_{n-1,\alpha}$

# Dos medias, datos apareados o pareados: regla de decisión

- $H_0: \mu_d = 0$ vs. $H_1: \mu_d \neq 0$

Rechazo $H_0$ si $t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}/\sqrt{n}} > t_{n-1,\alpha/2}$ o si

Rechazo $H_0$ si $t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}/\sqrt{n}} < -t_{n-1,\alpha/2}$

donde la $P(t_{n-1} > t_{n-1,\alpha}) = \alpha$

*t* de student con *n-1* grados de libertad

# Ejercicio: test de muestras apareadas

- Las autoridades públicas estan interesadas en conocer si hay diferencias en productos de una canasta básica de alimentos antes y después de la implementación de cierto Programa de precios.
- Para una muestra de 15 productos, se relevan los precios durante la semana previa a la implementación del Programa y una semana después de su implementación. El valor promedio de los productos la semana previa vs. el valor promedio de los productos la semana posterior a la implementación del Programa se muestran en la tabla a continuación.
- Plantee el test de hipótesis que considere adecuado y comente si existe alguna diferencia significativa al 10% en favor de la implementación del Programa.

# Ejemplo: muestras apareadas

| Producto | Semana previa a la implementación del Programa - valores en $- | Semana posterior a la implementación del Programa - valores en $- |
|---|---|---|
| 1 | 64.5 | 61.2 |
| 2 | 345 | 235 |
| 3 | 120 | 153 |
| 4 | 95 | 89 |
| 5 | 49 | 57 |
| 6 | 190 | 189 |
| 7 | 134 | 129 |
| 8 | 189 | 169 |
| 9 | 160 | 179 |
| 10 | 89 | 122 |
| 11 | 95 | 105 |
| 12 | 92 | 75 |
| 13 | 108 | 99 |
| 14 | 59 | 65 |
| 15 | 157 | 139 |

# Test para la diferencia entre dos proporciones (muestras independientes)

- El interés radica en evaluar si la proporción de individuos que utilizan Spotify es la misma que utiliza Deezer, en Argentina.
- Supongasé que se tienen una nuestra aleatoria de $n_x$ observaciones procedente de una población con probabilidad $p_x$ de éxito y una segunda muestra aleatoria independiente de $n_y$ observaciones procedentes de una población con probabilidad $p_y$ de éxito.
- Sabemos, por lo visto en las clases anteriores que cuando se cumplen ciertas condiciones, la distribución normal es una buena aproximación de las proporciones.

# Test para la diferencia entre dos proporciones (muestras grandes e independientes)

- Por lo tanto

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}}$$

Sigue una distribución normal estándar.

- Se quiere evaluar la hipótesis de que las proporciones poblacionales $p_x$ $y$ $p_y$ son iguales, i.e.

$$H_0\colon p_x - p_y = 0 \qquad o \qquad H_0\colon p_x = p_y$$

# Test para la diferencia entre dos proporciones (muestras grandes e independientes)

- Sea $p_0$ el valor común. Entonces partiendo de esta hipótesis

$$Z = \frac{\left(\hat{p}_x - \hat{p}_y\right)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}}$$

Sigue una distribución normal estándar.

- Finalmente la proporción desconocida $p_0$ se puede estimar por medio de un estimador agrupado

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

- En este test, la hipótesis nula asume que las proporciones poblacionales son iguales. Si $H_0$ es verdadera, entonces se puede obtener un estimador insesgado y eficiente de $p_0$ combinando las dos muestras aleatorias para calcular $\hat{p}_0$, y sustituir el $p_0$ desconocido por el $\hat{p}_0$ estimado.

# Test para la diferencia entre dos proporciones (muestras grandes e independientes): regla de decisión

- $H_0: p_x - p_y = 0 \quad \text{o} \quad p_x - p_y \leq 0 \quad \text{vs.} \quad H_1: p_x - p_y > 0$

Rechazo $H_0$ si $\dfrac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_\alpha$

- $H_0: p_x - p_y = 0 \quad \text{o} \quad p_x - p_y \geq 0 \quad \text{vs.} \quad H_1: p_x - p_y < 0$

Rechazo $H_0$ si $\dfrac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < -z_\alpha$

# Test para la diferencia entre dos proporciones (muestras grandes e independientes): regla de decisión

- $H_0: p_x - p_y = 0$ vs. $H_1: p_x - p_y \neq 0$

Rechazo $H_0$ si $\dfrac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < -z_{\alpha/2}$

- $o$

Rechazo $H_0$ si $\dfrac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{\alpha/2}$

# Ejercicio: test de diferencias de proporciones

- Se dispone de dos muestras de hombres y mujeres para comparar la cantidad de tareas domésticas realizadas por hombres y mujeres en matrimonios donde ambos aportan económicamente al hogar.
- De la muestra de 1645 hombres, el 67.5% de los hombres consideraba que la division de tareas era justa, mientras que de las 1691 mujeres muestreadas, el 60.8% sentían que la división de tareas era justa (American Journal of Sociology, septiembre 1994).
- ¿Es mayor la proporción de hombres que sentían que la division del trabajo doméstico era justa que la proporción correspondiente de mujeres? Respalde su conclusión con una prueba estadística con error de tipo I del 5%.