

Estadística en Analítica

2023-2

Pablo A. Saldarriaga
psaldar2@eafit.edu.co

UNIVERSIDAD
EAFIT

Relación entre variables

Covarianza:

$$COV(X, Y) = E[XY] - E[X]E[Y]$$

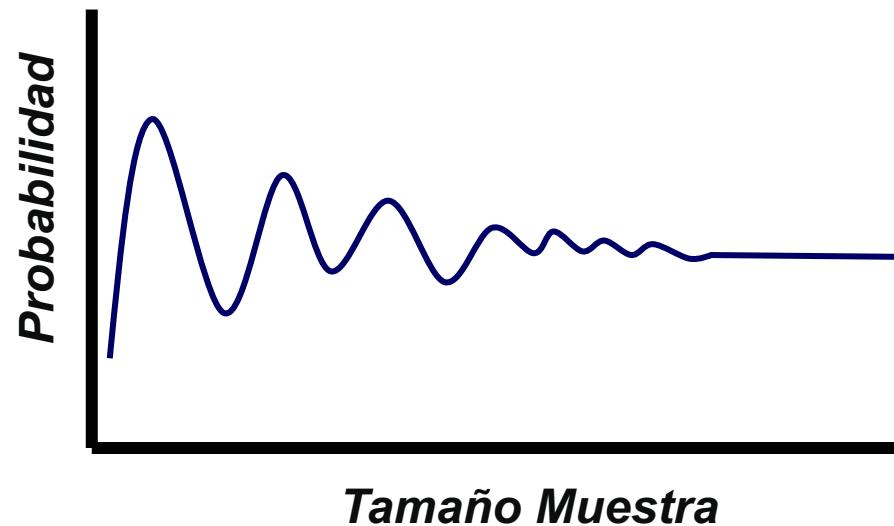
Correlación*:

$$Corr(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

*Esta medida de correlación es conocida como el índice de correlación de Pearson

Ley de los grandes números

La frecuencia relativa de los resultados de un cierto experimento aleatorio, *tienden a estabilizarse en cierto número, que es precisamente la probabilidad*, cuando el experimento se realiza muchas veces.



Teorema del límite central

Sea x_1, x_2, \dots, x_n una muestra aleatoria, tal que $x_i \sim (\mu, \sigma)$, así se cumple que:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Estimadores y Propiedades

Un estimador ($\hat{\theta}$) es una regla que nos indica como utilizar la información muestral para calcular el valor de un parámetro poblacional

$$E[\hat{\theta}] = \theta$$

Insesgados

$$\lim_{n \rightarrow \infty} P(|\widehat{\theta}_n - \theta| < \varepsilon) = 1$$

Consistencia

Distribuciones Muestrales

Media

Sea x_1, x_2, \dots, x_n una muestra aleatoria, tal que $x_i \sim N(\mu, \sigma)$, así se cumple que:

$$\bar{X} = \sum_{i=1}^n x_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

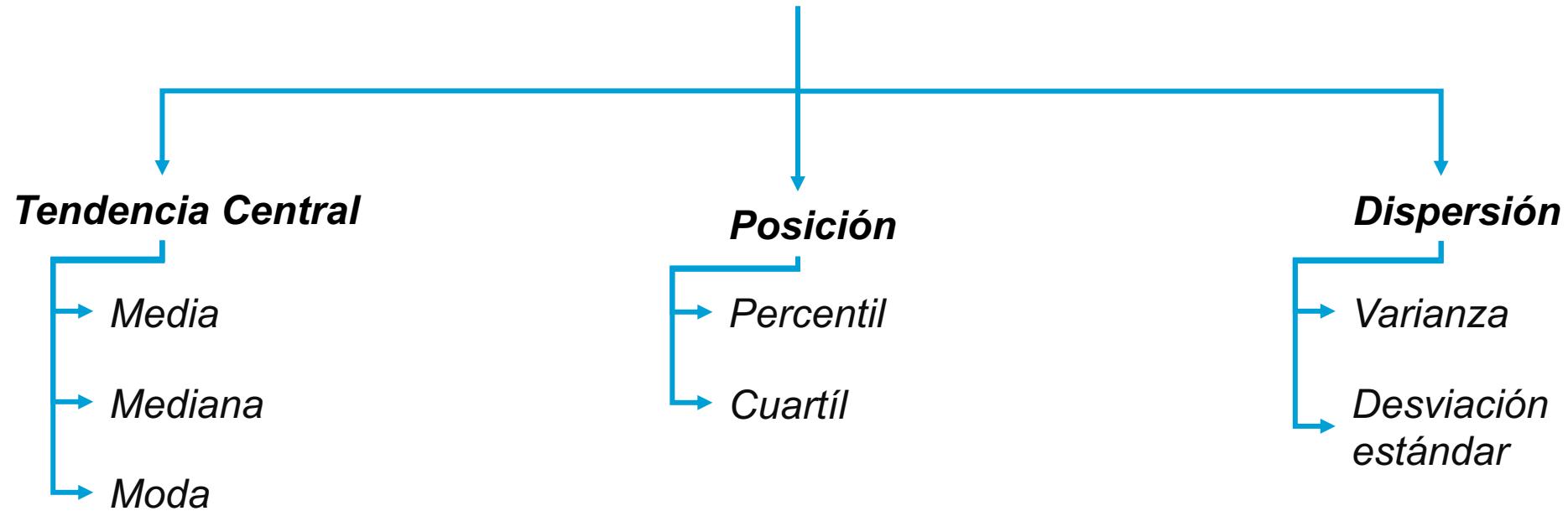
Distribuciones Muestrales

Varianza

Sea x_1, x_2, \dots, x_n una muestra aleatoria, tal que $x_i \sim N(\mu, \sigma)$, así se cumple que:

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

Medidas Estadísticas



Intervalos de Confianza

$$P(L_b \leq \theta \leq U_b) = 1 - \alpha$$

Cota Inferior **Cota Superior** **Confianza** **Significancia**

Para muestras grandes:

$$[\hat{\theta} - Z_{\alpha/2} \sigma_{\hat{\theta}} ; \hat{\theta} + Z_{\alpha/2} \sigma_{\hat{\theta}}]$$

Intervalos de Confianza

Sea x_1, \dots, x_n una m.a. de una población con media μ desconocida y varianza σ^2 conocida (o desconocida). Entonces, un I.C. al $(1 - \alpha)100\%$ para μ está dado por:

Caso	Intervalo de confianza
I: Si $n \in \mathbb{Z}^+$ y $X \sim N(\mu, \sigma^2)$ con σ^2 conocida	$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$
II: Si $n < 30$ y $X \sim N(\mu, \sigma^2)$ con σ^2 desconocida	$\left(\bar{x} - t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{s}{\sqrt{n}}, \quad \bar{x} + t_{\left(\frac{\alpha}{2}, n-1\right)} \frac{s}{\sqrt{n}} \right)$
III: Si $n \geq 30$ y X tiene una distribución con media μ y varianza σ^2 conocida (o desconocida)	$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$ Nota: Si σ^2 es desconocida se reemplaza σ por s

Intervalos de Confianza

Suponga que la distribución del gasto mensual en consumo de bienes alimenticios, por cada persona entre los 16 y 20 años en una ciudad, sigue una distribución normal. Al encuestar a 25 personas con edades entre los 16 y 20 años, y preguntarles por sus gastos de alimentación se obtiene un gasto medio de \$158.950 y una desviación estándar de \$23.800. ¿Se puede afirmar con un nivel de confianza del 96% que el gasto medio mensual en consumo de bienes alimenticos, de una persona con edad entre 16 y 20 años de esa ciudad, es inferior a \$180.000?

Intervalos de Confianza

Muestras grandes ($n_1, n_2 \geq 30$): Suponga que x_1, \dots, x_{n_1} es una m.a. de una población $N(\mu_1, \sigma_1^2)$. Sea x_1, \dots, x_{n_2} otra m.a. *independiente* de la anterior de una población $N(\mu_2, \sigma_2^2)$. Un I.C. al $(1 - \alpha)100\%$ para $\mu_1 - \mu_2$ es:

Caso	Intervalo de confianza
I: Si las varianzas σ_1^2 y σ_2^2 son conocidas o desconocidas	$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ Nota: Si las varianzas σ_1^2 y σ_2^2 son desconocidas, se reemplazan σ_1^2 y σ_2^2 por s_1^2 y s_2^2 .

Intervalos de Confianza

Muestras pequeñas ($n_1, n_2 < 30$): Suponga que x_1, \dots, x_{n_1} es una m.a. de una población $N(\mu_1, \sigma_1^2)$. Sea x_1, \dots, x_{n_2} otra m.a. *independiente* de la anterior de una población $N(\mu_2, \sigma_2^2)$.
Un I.C. al $(1 - \alpha)100\%$ para $\mu_1 - \mu_2$ es:

Caso	Intervalo de confianza
II: Si las varianzas σ_1^2 y σ_2^2 son conocidas	$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
III: Si las varianzas σ_1^2 y σ_2^2 son desconocidas, pero $\sigma_1^2 = \sigma_2^2 = \sigma^2$	$(\bar{x}_1 - \bar{x}_2) \pm t_{(\frac{\alpha}{2}, n_1+n_2-2)} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$ <p>donde $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$</p>
IV: Si las varianzas σ_1^2 y σ_2^2 son desconocidas, pero $\sigma_1^2 \neq \sigma_2^2$	$(\bar{x}_1 - \bar{x}_2) \pm t_{(\frac{\alpha}{2}, v)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ <p>donde el número de grados de libertad está dado por</p> $v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(s_1^2 \right)^2}{n_1-1} + \frac{\left(s_2^2 \right)^2}{n_2-1}}$

Intervalos de Confianza

La confederación colombiana de consumidores está interesada en estudiar la duración de las pilas E y D, por lo que prueba el número de horas que duran las pilas, tomando una muestra de 21 pilas de cada una con los siguientes resultados:

	Pilas E	Pilas D
Media	111.6 hrs	115.8 hrs
Desviación estándar	10 hrs	15 hrs

Con base en la evidencia muestral la confederación colombiana de consumidores, ¿puede concluir que no existe una diferencia significativa entre las duraciones promedios de las pilas D y E?.

Intervalos de Confianza

Recordemos que el estimador de p es la v.a. $\hat{p} = \frac{x}{n}$, donde $x \sim B(n, p)$. Un I.C. aproximado al $(1 - \alpha)100\%$ para p viene dado por:

Caso	Intervalo de confianza aproximado
I: Si $n < 30$	$\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}$ $\left(1 + \frac{z_{\alpha/2}^2}{n}\right)$
II: Si $n \geq 30$	$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right)$

Intervalos de Confianza

Es importante que las máscaras utilizadas por bomberos sean capaces de soportar altas temperaturas porque los bomberos comúnmente trabajan en temperaturas de $200 - 500^{\circ}F$. El fabricante de un tipo de máscaras asegura que más del 80% de sus máscaras soportan (sin sufrir daño alguno) temperaturas superiores a $250^{\circ}F$. Si en una prueba de ese tipo de máscara, a 11 de 55 máscaras se les desprendió la mica a $250^{\circ}F$, ¿respalda esta evidencia la afirmación del fabricante con un nivel de confianza del 90%? Justifique su respuesta.

Intervalos de Confianza

Sean p_1 y p_2 dos proporciones de interés para dos poblaciones independientes. Se puede mostrar que si $n_1, n_2 \geq 30$, entonces un I.C. aproximado al nivel $(1 - \alpha)100\%$ para $p_1 - p_2$ es:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Intervalos de Confianza

Los recientes incidentes de contaminación de los alimentos han causado gran preocupación entre los consumidores. En un reciente estudio se informó que 35 de los 80 pollos seleccionados al azar de la marca P dieron positivo, ya sea para campylobacter o salmonella (o ambos), las principales causas bacterianas de enfermedades transmitidas por los alimentos; mientras que 46 de 80 pollos de la marca T dieron positivo. ¿Parece que la verdadera proporción de los pollos P no contaminados difiere significativamente de aquella de la marca T? Construya un I.C. usando un nivel de confianza del 99%. Justifique su respuesta.

Pruebas de Hipótesis

Una **prueba de hipótesis** es un procedimiento para tomar una decisión, bajo incertidumbre, sobre la validez de la hipótesis nula usando la evidencia de los datos.

Componentes:

1. Dos hipótesis, una nula y otra alternativa
2. Estadístico de prueba
3. Región de rechazo
4. Tomar la decisión

Pruebas de Hipótesis

Cuadro de decisiones y errores

Posibles decisiones	Situaciones posibles	
	H_0 es cierta	H_0 es falsa
Rechazar H_0	Error tipo I	Decisión correcta
No rechazar H_0	Decisión correcta	Error tipo II

Nivel de significancia (α): $P(\text{Error tipo I})$

- ✓ Es deseable que estas las probabilidades de error sean pequeñas.
- ✓ El nivel de significancia de la prueba (o tamaño de la región crítica) debe ser prefijado de antemano.
- ✓ Al realizar la prueba se toma en cuenta el error de tipo I. Por lo tanto, la prueba es significativa si se rechaza la hipótesis nula, pues en este caso se conoce la probabilidad de haber cometido un error.

Pruebas de Hipótesis

1. Plantear H_0 y H_1

Si θ es un parámetro de interés y θ_0 es un valor fijo, entonces se puede proponer cualquiera de las siguientes parejas de hipótesis:

$$\begin{cases} H_o^{(1)}: \theta \leq \theta_0 \\ H_1^{(1)}: \theta > \theta_0 \end{cases}$$

Prueba unilateral
de una cola a **derecha**

$$\begin{cases} H_o^{(2)}: \theta \geq \theta_0 \\ H_1^{(2)}: \theta < \theta_0 \end{cases}$$

Prueba unilateral
de una cola a **izquierda**

$$\begin{cases} H_o^{(3)}: \theta = \theta_0 \\ H_1^{(3)}: \theta \neq \theta_0 \end{cases}$$

Prueba bilateral
o de **dos colas**

2. Seleccionar el nivel de significancia α

En la práctica, es frecuente un nivel de significancia de 0.05 ó 0.01, si bien se pueden elegir otros valores.

Pruebas de Hipótesis

3. Calcular el estadístico de prueba.

- Es un valor determinado a partir de la información muestral, que se utiliza para determinar si se rechaza la hipótesis nula.
- Existen muchos estadísticos de prueba. Sin embargo, para nuestro caso utilizaremos los estadísticos z_c , t_c y F_c .
- Generalmente, los estadísticos z_c y t_c tienen la siguiente estructura:

$$\frac{\text{estimador puntual} - \text{valor nulo}}{\text{error estándar}} = \frac{\text{estimador puntual} - \theta_0}{\text{error estándar}}$$

- La distribución de referencia en la prueba de hipótesis de un parámetro θ , es la distribución que sigue el estadístico cuando $\theta = \theta_0$.

Pruebas de Hipótesis

4. Decisión estadística

Se establecen las condiciones específicas en las que se rechaza H_0 y las condiciones en las que no se rechaza H_0 .

Basada en la región crítica

Prueba de una cola a derecha	Prueba de una cola a izquierda	Prueba de dos colas
$H_o^{(1)}: \theta \leq \theta_0$ vs. $H_1^{(1)}: \theta > \theta_0$	$H_o^{(2)}: \theta \geq \theta_0$ vs. $H_1^{(2)}: \theta < \theta_0$	$H_o^{(3)}: \theta = \theta_0$ vs. $H_1^{(3)}: \theta \neq \theta_0$

El signo
en H_1
indica la
región
crítica

Si $\alpha > p - value$ se rechaza H_0

Pruebas de Hipótesis - Media

Sea x_1, \dots, x_n una m.a. de una población, con media μ desconocida y varianza σ^2 conocida (o desconocida según el caso). Sea μ_0 un valor de interés para μ .

1. Para probar:

$$\begin{cases} H_0^{(1)} : \mu \leq \mu_0 \\ H_1^{(1)} : \mu > \mu_0 \end{cases} \quad \text{ó} \quad \begin{cases} H_0^{(2)} : \mu \geq \mu_0 \\ H_1^{(2)} : \mu < \mu_0 \end{cases} \quad \text{ó} \quad \begin{cases} H_0^{(3)} : \mu = \mu_0 \\ H_1^{(3)} : \mu \neq \mu_0 \end{cases}$$

2. Se elige un nivel de significancia $\alpha : 1\%, 5\%, 10\%$, etc.
3. Se calcula el estadístico de prueba, y
4. Se toma la decisión estadística:

Nota: Las condiciones expuestas a continuación, basados en la región crítica y el valor p , son para **rechazar H_0** .

Pruebas de Hipótesis - Media

Caso	3. Estadístico de prueba	4. Decisión estadística			
		Basada en	$H_1^{(1)}: \mu > \mu_0$	$H_1^{(2)}: \mu < \mu_0$	$H_1^{(3)}: \mu \neq \mu_0$
I: Si $n \in \mathbb{Z}^+$ y $X \sim N(\mu, \sigma^2)$ con σ^2 conocida III: Si $n \geq 30$ y X tiene una distribución con media μ y varianza σ^2 conocida (o desconocida).	$z_c = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ <p>Nota: En el caso III, si σ^2 es desconocida se reemplaza σ por s</p>	Región crítica	$z_c > z_\alpha$	$z_c < -z_\alpha$	$ z_c > z_{\alpha/2}$
		Valor p	$P(Z > z_c) \leq \alpha$	$P(Z < z_c) \leq \alpha$	$P(Z > z_c) \leq \alpha$
II: Si $n < 30$ y $X \sim N(\mu, \sigma^2)$ con σ^2 desconocida	$t_c = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	Región crítica	$t_c > t_{(\alpha, n-1)}$	$t_c < -t_{(\alpha, n-1)}$	$ t_c > t_{(\alpha/2, n-1)}$
		Valor p	$P(t_{(n-1)} > t_c) \leq \alpha$	$P(t_{(n-1)} < t_c) \leq \alpha$	$P(t_{(n-1)} > t_c) \leq \alpha$

Pruebas de Hipótesis - Media

Suponga que la distribución del gasto mensual en consumo de bienes alimenticios, por cada persona entre los 16 y 20 años en una ciudad, sigue una distribución normal. Al encuestar a 25 personas con edades entre los 16 y 20 años, y preguntarles por sus gastos de alimentación se obtiene un gasto medio de \$158.950 y una desviación estándar de \$23.800. ¿Se puede afirmar con un nivel de significancia del 2%, que el gasto medio mensual en consumo de bienes alimenticos, de una persona con edad entre 16 y 20 años de esa ciudad, es inferior a \$180.000?

Pruebas de Hipótesis – Diferencia de Medias

Suponga que x_1, \dots, x_{n_1} es una m.a. de una población $N(\mu_1, \sigma_1^2)$. Sea x_1, \dots, x_{n_2} otra m.a. *independiente* de la anterior de una población $N(\mu_2, \sigma_2^2)$.

- 1 Para probar:

$$\begin{cases} H_0^{(1)} : \mu_1 - \mu_2 \leq \mu_0 \\ H_1^{(1)} : \mu_1 - \mu_2 > \mu_0 \end{cases} \quad \text{ó} \quad \begin{cases} H_0^{(2)} : \mu_1 - \mu_2 \geq \mu_0 \\ H_1^{(2)} : \mu_1 - \mu_2 < \mu_0 \end{cases} \quad \text{ó}$$

$$\begin{cases} H_0^{(3)} : \mu_1 - \mu_2 = \mu_0 \\ H_1^{(3)} : \mu_1 - \mu_2 \neq \mu_0 \end{cases}$$

donde μ_0 es una cantidad fija.

- 2 Se elige un nivel de significancia $\alpha : 1\%, 5\%, 10\%$, etc.

Pruebas de Hipótesis – Diferencia de Medias

- ③ Se calcula el estadístico de prueba:

Caso I: Si $n_1, n_2 \geq 30$ y σ_1^2 y σ_2^2 son ambas conocidas (o desconocidas).

Caso II: Si $n_1, n_2 < 30$, X_1 y X_2 son normales, y σ_1^2 y σ_2^2 son conocidas.

$$z_c = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

- ④ Decisión estadística (sistema creado para rechazar H_0):

	$H_1^{(1)} : \mu_1 - \mu_2 > \mu_0$	$H_1^{(2)} : \mu_1 - \mu_2 < \mu_0$	$H_1^{(3)} : \mu_1 - \mu_2 \neq \mu_0$
Casos I y II	$z_c > z_\alpha$	$z_c < -z_\alpha$	$ z_c > z_{\alpha/2}$
Caso III	$t_c > t_{(\alpha, n_1+n_2-2)}$	$t_c < -t_{(\alpha, n_1+n_2-2)}$	$ t_c > t_{(\alpha/2, n_1+n_2-2)}$
Caso IV	$t_c > t_{(\alpha, \nu)}$	$t_c < -t_{(\alpha, \nu)}$	$ t_c > t_{(\alpha/2, \nu)}$

- ⑤ Conclusión.

Pruebas de Hipótesis – Diferencia de Medias

(Ejercicio) Un estudio de dos tipos de equipo de fotocopiado demuestra que 60 fallas del primer tipo de equipo tardaron un promedio de 80.7 minutos en ser reparadas, con una desviación estándar de 19.4 minutos; mientras tanto, 50 fallas del segundo tipo de equipo tardaron en promedio 88.1 minutos en repararse con una desviación estándar de 18.8 minutos. ¿La evidencia muestral indica que las fallas del segundo tipo de equipo requieren un mayor tiempo para ser reparadas?. Use $\alpha = 0.01$.

Prueba de localización

Considere las muestras aleatorias independientes X_1, X_2, \dots, X_{n_1} y Y_1, Y_2, \dots, Y_{n_2} , se desea probar si ambas provienen de la misma distribución. Para esto se plantea:

$$\begin{aligned} H_0: \mu_x - \mu_y &= 0 \\ H_1: \mu_x - \mu_y &\neq 0 \end{aligned}$$

En el caso paramétrico, se asume normalidad en las muestras, además de tener la misma media y varianza

En la estadística paramétrica se utiliza la prueba t para contrastar la hipótesis



Pruebas de Hipótesis – Proporción

Sea x_1, \dots, x_n una m.a. de una distribución Bernoulli de parámetro p . Recordemos que el estimador de éste parámetro es la v.a. $\hat{p} = \frac{X}{n}$, donde $X \sim B(n, p)$. Además, el Teorema Central del Límite garantiza que si $n \geq 30$

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

Para probar:

$$\begin{cases} H_0^{(1)} : p \leq p_0 \\ H_1^{(1)} : p > p_0 \end{cases} \quad \text{ó} \quad \begin{cases} H_0^{(2)} : p \geq p_0 \\ H_1^{(2)} : p < p_0 \end{cases} \quad \text{ó} \quad \begin{cases} H_0^{(3)} : p = p_0 \\ H_1^{(2)} : p < p_0 \end{cases}$$

con p_0 fijo.

Pruebas de Hipótesis – Proporción

Se calcula el estadístico de prueba (se requiere que $n \geq 30$):

$$z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1)$$

donde, $\hat{p} = \frac{x}{n}$, $X \sim Bin(n, p)$.

Decisión estadística (sistema recreado para rechazar H_0):

$H_1^{(1)} : p > p_0$	$H_1^{(2)} : p < p_0$	$H_1^{(2)} : p < p_0$
$z_c > z_\alpha$	$z_c < -z_\alpha$	$ z_c > z_{\alpha/2}$

Pruebas de Hipótesis – Proporción

La proporción de fumadores adultos en la ciudad A en el año 1992 era de 25.5%. En el año 1993 el departamento de salud decide hacer una encuesta para saber si ha bajado la proporción. Se realizó una encuesta telefónica a 2400 adultos, y se obtuvo que la proporción de fumadores en la encuesta era de 25%. ¿Se puede afirmar que la proporción real de fumadores ha bajado significativamente? Use $\alpha = 0.05$

Pruebas de Hipótesis – Diferencia de Proporciones

Sea x_1, \dots, x_{n_1} una m.a. de una distribución Bernoulli de parámetro p_1 , cuyo estimador es la v.a. $\hat{p}_1 = \frac{X_1}{n_1}$, donde $X_1 \sim B(n_1, p_1)$. Suponga que x_1, \dots, x_{n_2} es una m.a. (independiente de la anterior) de una distribución Bernoulli de parámetro p_2 , cuyo estimador es la v.a. $\hat{p}_2 = \frac{X_2}{n_2}$, donde $X_2 \sim B(n_2, p_2)$.

- ① Para probar:

$$\begin{cases} H_0^{(1)} : p_1 - p_2 \leq p_D \\ H_1^{(1)} : p_1 - p_2 > p_D \end{cases} \quad \text{ó} \quad \begin{cases} H_0^{(2)} : p_1 - p_2 \geq p_D \\ H_1^{(2)} : p_1 - p_2 < p_D \end{cases} \quad \text{ó}$$
$$\begin{cases} H_0^{(3)} : p_1 - p_2 = p_D \\ H_1^{(3)} : p_1 - p_2 \neq p_D \end{cases}$$

donde p_D es una cantidad fija.

- ② Se elige un nivel de significancia $\alpha : 1\%, 5\%, 10\%$, etc.

Pruebas de Hipótesis – Diferencia de Proporciones

- ③ Se calcula el estadístico de prueba (con $n_1, n_2 \geq 30$):

- Si $p_D = 0$:

$$z_c = \frac{(\hat{p}_1 - \hat{p}_2) - p_D}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$$

donde, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ y $X_i \sim Bin(n_i, p_i)$, $i = 1, 2$.

- Si $p_D \neq 0$:

$$z_c = \frac{(\hat{p}_1 - \hat{p}_2) - p_D}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} \sim N(0, 1)$$

donde, $\hat{p}_i = \frac{x_i}{n_i}$ y $X_i \sim Bin(n_i, p_i)$, $i = 1, 2$.

Pruebas de Hipótesis – Diferencia de Proporciones

- ④ Decisión estadística (sistema recreado para rechazar H_0):

$H_1^{(1)} : p_1 - p_2 > p_D$	$H_1^{(2)} : p_1 - p_2 < p_D$	$H_1^{(3)} : p_1 - p_2 \neq p_D$
$z_c > z_\alpha$	$z_c < -z_\alpha$	$ z_c > z_{\alpha/2}$

- ⑤ Conclusión.

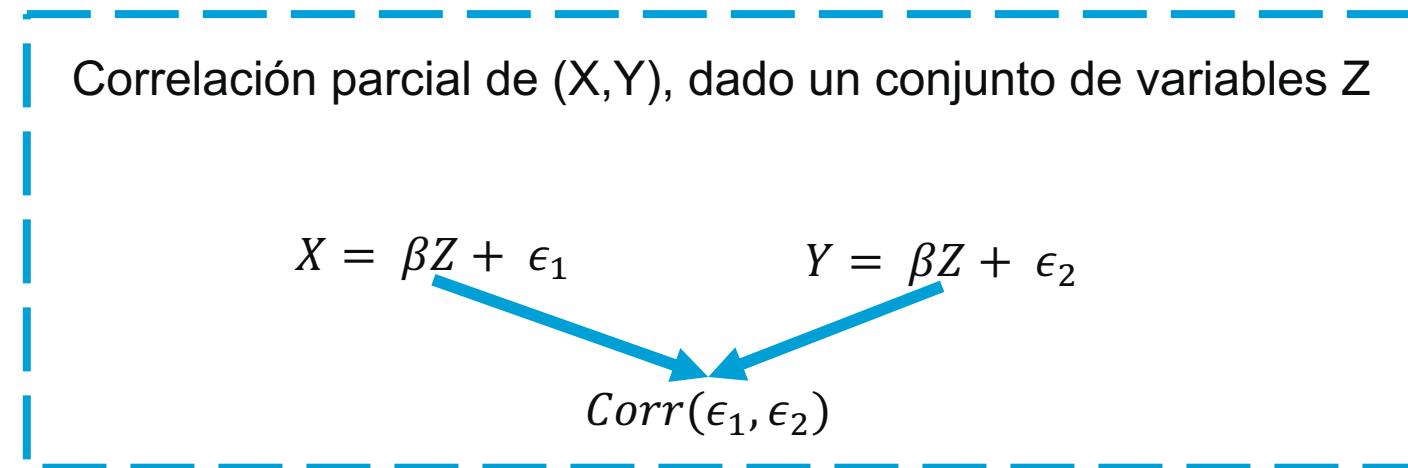
Análisis de Correlación

Correlación (Pearson)

- ✓ Indicador que permite medir la “fuerza” o “intensidad” de relación lineal entre dos variables
- ✓ Debido a la relación lineal que se encuentra, se puede interpretar como una relación de directa o inversa proporcionalidad
- ✓ Se considera como un primer método/filtro de variables dentro del proceso de modelamiento

Correlación Parcial

Corresponde a la correlación existente entre 2 variables *eliminando el efecto de las variables* adicionales en consideración



Índice de Correlación Múltiple

Este índice corresponde al *porcentaje de variabilidad* que puede ser explicada una variable en términos de las otras

$$1 - \frac{1}{\text{diag}(\Sigma)\text{diag}(\Sigma^{-1})}$$

Σ corresponde a la matriz de varianzas y covarianzas de los datos

**Inspira
Crea
Transforma**