

Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS
MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 8. Contrastes de hipótesis

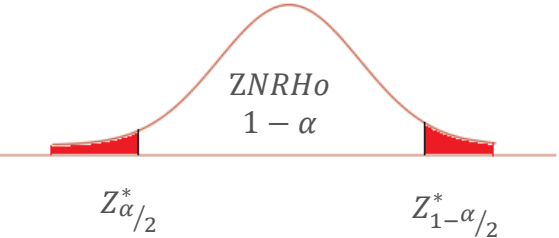
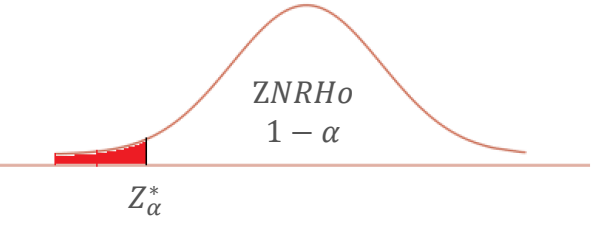
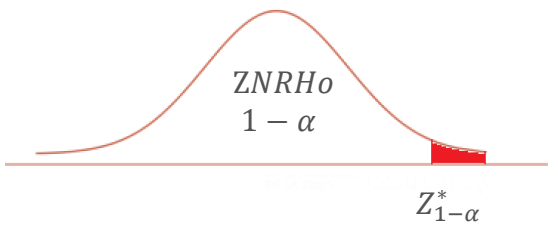
Tabla de contenido

□ Tema 8: Contrastes de hipótesis_Parte II

- Contrastes de hipótesis sobre la varianza.
- Contrastes paramétricos para dos muestras.

ISSN 0013-735X

Prueba de hipótesis para la proporción (P)

Prueba Bilateral	Prueba Unilateral	
$H_0: P = P_0$ $H_1: P \neq P_0$	$H_0: P \geq P_0$ $H_1: P < P_0$	$H_0: P \leq P_0$ $H_1: P > P_0$
		
$P - value = 2 * P(Z > Z_c)$	$P - value = P(Z < Z_c)$	$P - value = P(Z > Z_c)$
Rechazo H_o si $ Z_c > Z_{1-\alpha/2}^*$ O si $P - value < \alpha$	Rechazo H_o si $Z_c < Z_{\alpha}^*$ O si $P - value < \alpha$	Rechazo H_o si $Z_c > Z_{1-\alpha}^*$ O si $P - value < \alpha$

Valor calculado Z →

$$z_c = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}} \sim N(0,1)$$

Prueba de hipótesis para la proporción (P)

Ejemplo:

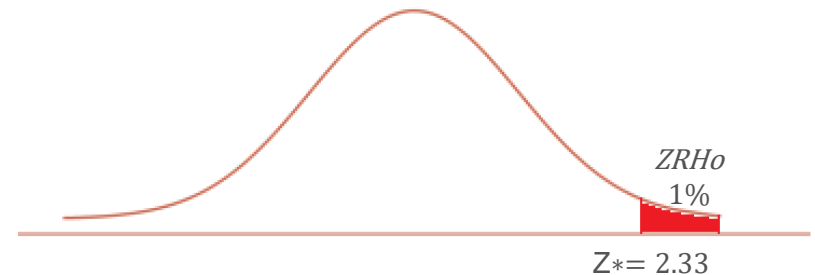
Un fabricante asegura a una compañía que le compra un producto en forma regular, que el porcentaje de productos defectuosos no es mayor que el 5%. La compañía decide comprobar la afirmación del fabricante seleccionando de su inventario 200 unidades de este producto y probándolas. Deberá sospechar la compañía de la afirmación del fabricante si se descubren un total de 19 unidades defectuosas en la muestra? (use un nivel de significancia del 1%).

$$H_0: P \leq 5\%$$

$$H_1: P > 5\%$$

$$\hat{p} = \frac{19}{200} = 0.095$$

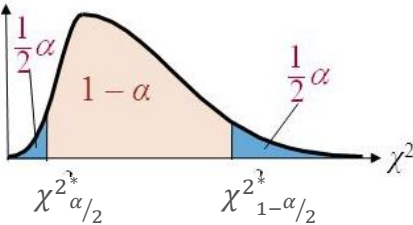
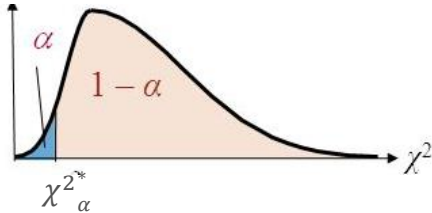
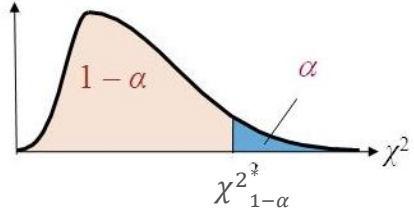
$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}} = \frac{9.5\% - 5\%}{\sqrt{\frac{5\% * (100\% - 5\%)}{200}}} = 2.92$$



$$P_value < \alpha$$

Hay evidencia estadística para decir que el porcentaje de piezas defectuosas es superior al 5%

Prueba de hipótesis para la varianza (σ^2)

Prueba Bilateral	Prueba Unilateral	
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$H_0: \sigma^2 \geq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$
		
$P - value = 2 * \min\{P(\chi^2 > \chi_c^2), P(\chi^2 < \chi_c^2)\}$	$P - value = P(\chi^2 < \chi_c^2)$	$P - value = P(\chi^2 > \chi_c^2)$

$$\text{Valor calculado } \chi^2 \rightarrow \chi_c^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Bajo el supuesto de normalidad de la variable

Prueba de hipótesis para la varianza (σ^2)

Ejemplo

¿Están cumpliendo las embotelladoras españolas de Coca-Cola?

La marca Coca Cola impone a sus empresas embotelladoras un riguroso control de calidad, para que sean capaces de embotellar el producto con una varianza mínima de líquido contenido en cada botella, cantidad que establecen en los contratos pertinentes. En este caso la marca registrada Coca Cola España® ha acordado con la embotelladora *Paco's bottle S.A.* (sita en Cáceres) embotellar el producto con una desviación típica $\sigma=0,151$ cl.

Para ver si realmente está cumpliendo lo acordado vamos a contrastar:

47,26	47,26	47,29	47,26	47,02	47,38	47,29	47,11
47,38	47,38	47,38	47,20	47,35	47,32	47,29	47,17
47,17	47,20	47,20	47,38	47,29	47,53	47,11	47,47

Prueba de hipótesis para la varianza (σ^2)

Ejemplo

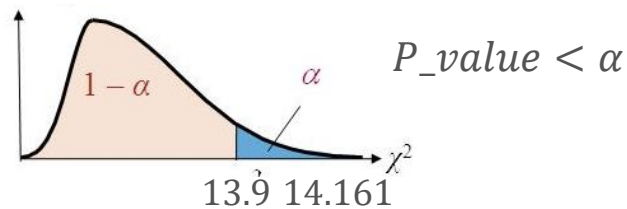
$$H_0: \sigma^2 \leq 0.023$$

$$H_1: \sigma^2 > 0.023$$

Bajo el supuesto de normalidad de la variable

$$\chi_c^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(24-1)0.119^2}{0.023} = 14.161$$

$$\chi_{23;0.99}^2 = 13.9$$



Se rechaza H_0 y se concluye que la embotelladora no está cumpliendo su contrato con CocaCola. (Nivel de significancia del 1%)

Prueba de hipótesis para diferencia de medias (2 muestras)

Para dos poblaciones (muestras) independientes



- ¿el peso promedio de los jugadores es el mismo?
- ¿la estatura promedio de los jugadores es la misma?
- ¿la edad promedio de los jugadores es la misma?
- ¿el ingreso promedio de los jugadores es el mismo?
- ¿La dispersión del ingreso es la misma?
- ¿las habilidades de los jugadores (score) es el mismo?

Prueba de hipótesis para diferencia de medias (2 muestras)

Para dos poblaciones (muestras) independientes

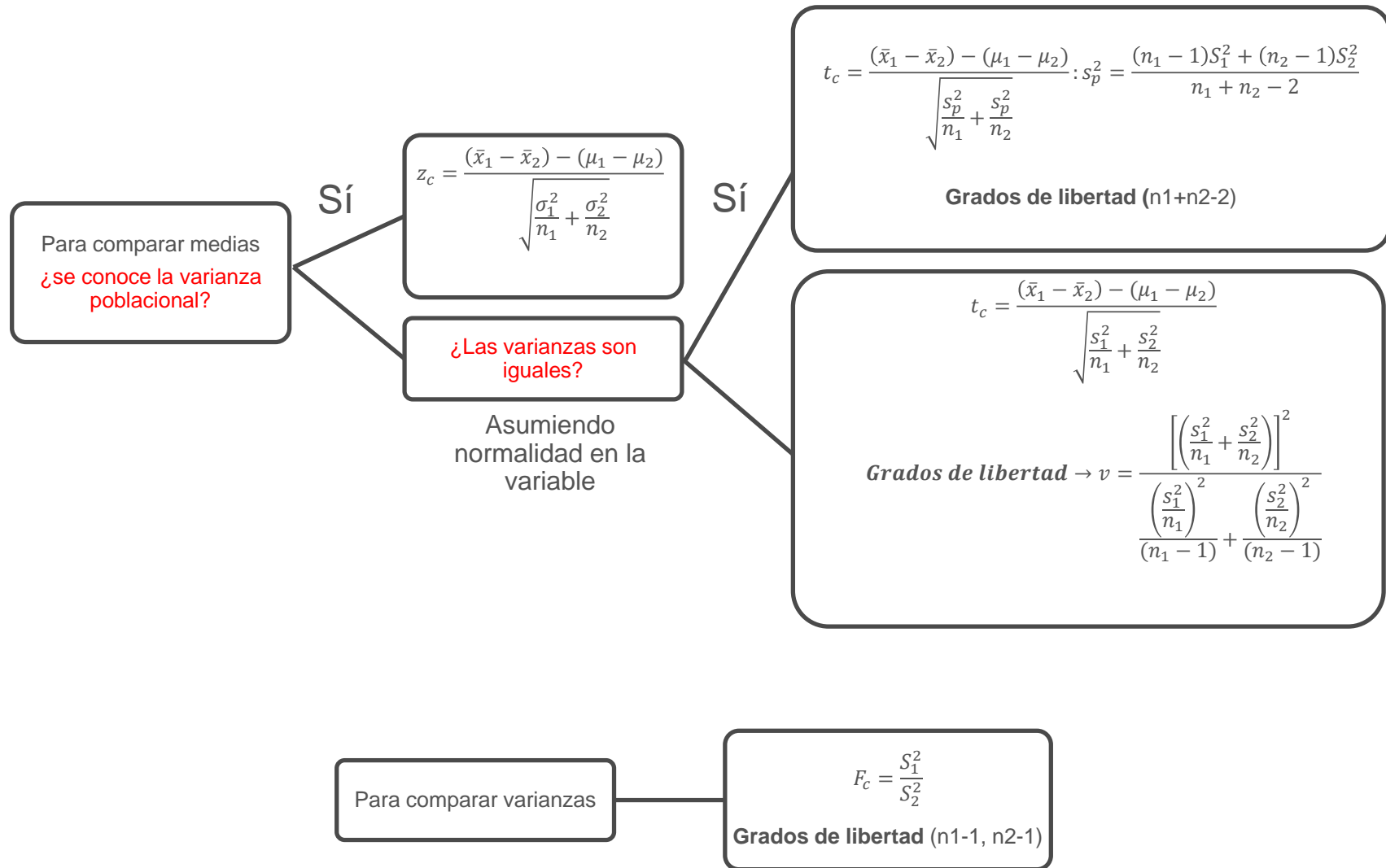
¿Existen diferencias entre los CI de personas con trastornos psicóticos y las que presentan rasgos paranoides?

¿Existen diferencias entre las cantidades de aceite diario producidas en las diferentes cooperativas de aceite de la provincia de Castellón?

¿Son igual de longevas dos especies de hormigas?

¿Dos fármacos son igualmente efectivos para tratar determinada patología?

Prueba de hipótesis para diferencia de medias (2 muestras)



Prueba de hipótesis para diferencia de medias (2 muestras)

Con varianzas desconocidas homogéneas.
Bajo normalidad de la variable de interés.

$$H_o : \mu_1 - \mu_2 = 0$$

$$H_o : \mu_1 - \mu_2 \geq 0$$

$$H_o : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

$$H_a : \mu_1 - \mu_2 < 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

Estadístico de prueba

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}; s_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Estadístico de prueba t con $n_1 + n_2 - 2$ grados de libertad

Prueba de hipótesis para diferencia de medias (2 muestras)

Con varianzas desconocidas heterogéneas (Welch- Satterthwaite).
Bajo normalidad de la variable de interés.

$$H_o : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

$$H_o : \mu_1 - \mu_2 \geq 0$$

$$H_a : \mu_1 - \mu_2 < 0$$

$$H_o : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

Estadístico de prueba

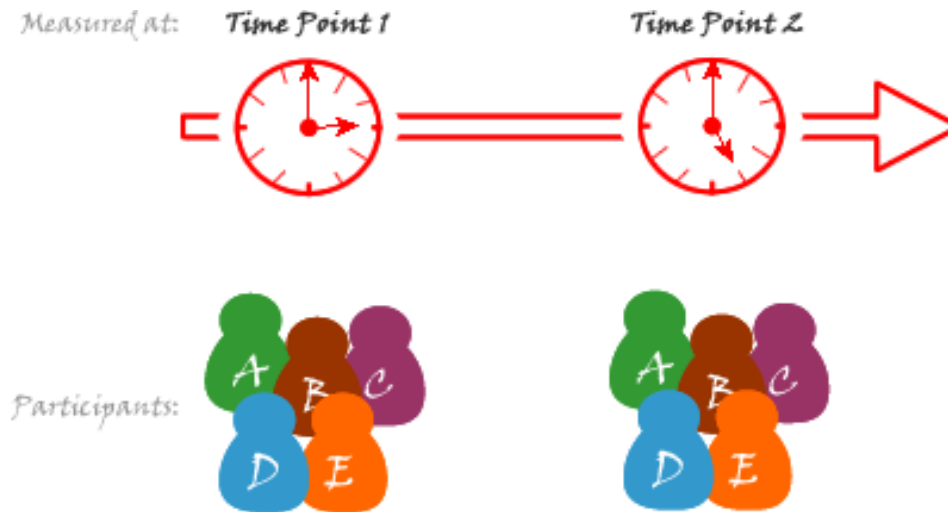
$$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Estadístico de
prueba t con v
grados de libertad

$$v = \frac{\left[\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right) \right]^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{(n_1 - 1)} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{(n_2 - 1)}}$$

Prueba t de hipótesis (2 muestras relacionadas)

Bajo normalidad de la variable de interés.
2 muestras dependientes o relacionadas.



es una

Ej. Hay algún efecto del ejercicio físico en el nivel de colesterol en plasma. Experimento: Muestras de sangre antes y después del ejercicio

Ej. Hay algún efecto de la educación financiera en el nivel de ahorro de personas de bajos recursos. Experimento: Evaluación del nivel de ahorro antes y después de una intervención de EF.

1. Same Participants
2. Same Dependent Variable
3. Same condition/treatment

Prueba t de hipótesis (2 muestras relacionadas)

$$H_0: \mu_d = \mu_{d0}$$

$$H_1: \mu_d \neq \mu_{d0}$$

$$H_0: \mu_d \geq \mu_{d0}$$

$$H_1: \mu_d < \mu_{d0}$$

$$H_0: \mu_d \leq \mu_{d0}$$

$$H_1: \mu_d > \mu_{d0}$$

$$d = X_{G1} - X_{G2}$$

El estadístico de prueba es:

$$t_c = \frac{\bar{d} - \mu_{d0}}{s_d / \sqrt{n}} \sim t_{n_d - 1} \text{ donde } \bar{d} = \frac{\sum d_i}{n}$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n_d - 1}} \text{ desviación estándar observada de las } n \text{ diferencias}$$

Prueba z de hipótesis (diferencia de proporciones)

$$H_0: P_1 - P_2 = P_0$$

$$H_1: P_1 - P_2 \neq P_0$$

$$H_0: P_1 - P_2 \geq P_0$$

$$H_1: P_1 - P_2 < P_0$$

$$H_0: P_1 - P_2 \leq P_0$$

$$H_1: P_1 - P_2 > P_0$$

El estadístico de prueba asumiendo $P_0 = 0$ es:

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{P}_0(1 - \hat{P}_0)}{n_1} + \frac{\hat{P}_0(1 - \hat{P}_0)}{n_2}}} \quad \text{Con } \hat{P}_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

El estadístico de prueba asumiendo $P_1 \neq P_2$ es:

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2) - (P_0)}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$$

Como supuesto para garantizar normalidad; $np > 5$ y $n(1 - p) > 5$ para ambas poblaciones

Contraste de hipótesis dos poblaciones (Ejemplos)

¿Hubo discriminación por la estatura?

Un grupo de aspirantes a un puesto de vigilante de seguridad que no fueron elegidos por una compañía subcontratada por el Metro de Madrid y piensan en poner una denuncia, pues estiman que hubo discriminación. Creen que seleccionaron a los más altos en estatura, cuando la empresa en las condiciones no decía nada de ese punto.

Gracias a la estadística podemos resolver este asunto y descubrir si realmente hubo discriminación y, por consiguiente, su demanda podría tener un buen apoyo de cara a una resolución judicial favorable.

Las estaturas en centímetros de los seleccionados:

169	172	172	173	176	177	178	179	179	180
180	180	181	183	184	188	188	189	189	190
191	192	195							

Las estaturas en centímetros de los no seleccionados:

162	168	171	172	173	173	174	177	177	178
178	179	179	179	180	180	180	180	181	181
182	182	183	183	184	184	186	186	187	189

Contraste de hipótesis dos poblaciones (Ejemplos)

¿Hubo discriminación por la estatura?

$$H_0: \mu_{seleccionados} - \mu_{No_seleccionados} \leq 0 \quad \text{Variable: Estatura en cm}$$

$$H_1: \mu_{seleccionados} - \mu_{No_seleccionados} > 0$$

Supuestos:

- La estatura sigue una distribución normal (se puede validar con una PH)
- Varianzas desconocidas pero iguales (se puede validar con una PH)

$$\bar{x} = 181.96 \text{ cm}, n_x = 23, s_x = 7.22$$

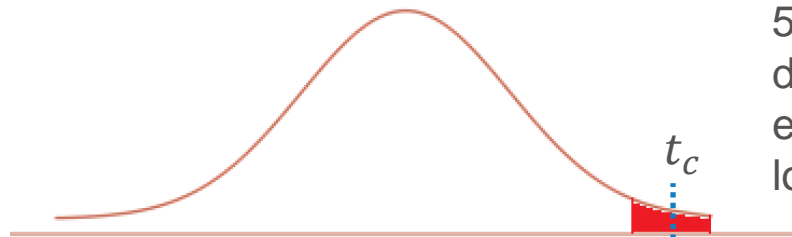
$$\bar{y} = 178.93 \text{ cm}, n_y = 30, s_y = 5.88$$

$$s_p = 6.492$$

$$GL = 23 + 30 - 2 = 51$$

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = 1.684$$

Valor crítico:



Se rechaza H_0 . Hay evidencia (con 5% de significancia) de discriminación por estatura de la empresa subcontratada para elegir a los vigilantes

$$t_{51;0.95} = 1.675$$

$$P_{value} < 0.05$$

Contraste de hipótesis dos poblaciones (Ejemplos)

Se sospecha que añadiendo al tratamiento habitual para la curación de una enfermedad un medicamento A, se consigue mayor número de curaciones. Tomamos dos grupos de enfermos de 100 individuos cada uno. A un grupo se le suministra el medicamento A y se curan 60 enfermos y al otro no se le suministra, curándose 55 enfermos. ¿Es efectivo el tratamiento A en la curación de la enfermedad?

$$H_0: P_A - P_{NA} \leq 0$$

$$H_1: P_A - P_{NA} > 0$$

$$X = \begin{cases} 1 & \text{si se cura} \\ 0 & \text{en otro caso} \end{cases}$$

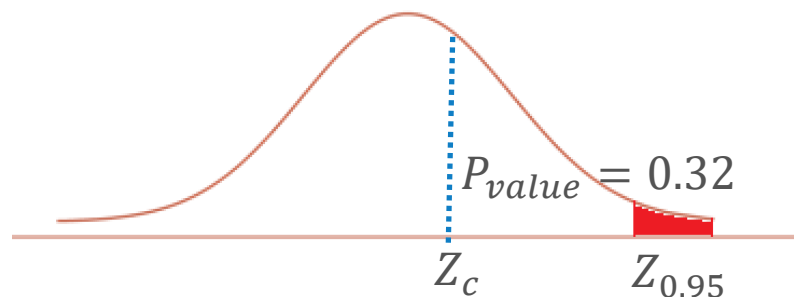
$$\hat{P}_A = 0.6$$

$$\hat{P}_{NA} = 0.55$$

$$n_A = n_{NA} = 100$$

$$\hat{P}_0 = \frac{100 * 0.6 + 100 * 0.55}{200} = 0.575$$

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{P}_0(1 - \hat{P}_0)}{n_1} + \frac{\hat{P}_0(1 - \hat{P}_0)}{n_2}}} = 0.466$$



No se rechaza H_0 . No hay evidencia (con 5% de significancia) para decir que le medicamento A garantice un mayor tasa de reuperación

Contraste de hipótesis dos poblaciones (Ejemplos)

Se realiza un estudio, en el que participan 10 individuos, para investigar el efecto del ejercicio físico (que se espera positivo) en el nivel de colesterol en plasma. Antes del ejercicio se tomaron muestras de sangre para determinar el nivel de colesterol de cada individuo. Después, los participantes fueron sometidos a un programa de ejercicios. Al final de los ejercicios se tomaron nuevamente muestras de sangre y se obtuvo una segunda lectura del nivel de colesterol. Los resultados se muestran a continuación.

Nivel previo: 182; 230; 160; 200; 160; 240; 260; 480; 263; 240

Nivel posterior: 190; 220; 166; 150; 140; 220; 156; 312; 240; 250 (*Variable: Nivel de colesterol en plasma*)

$$H_0: \mu_d \leq 0$$

Previo - Posterior

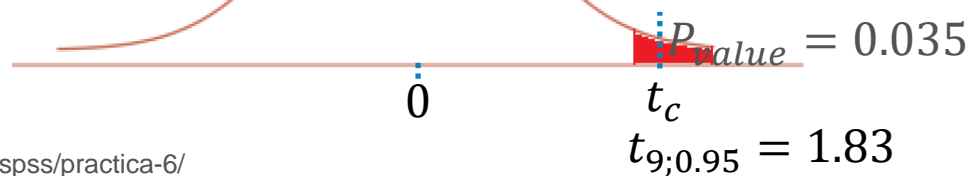
$$H_1: \mu_d > 0$$

$$d_i = \{182 - 190, 230 - 220, \dots, 240 - 250\}$$

$$\bar{d} = 37, n_d = 10, s_d = 57.16$$

$$t_c = \frac{\bar{d} - \mu_{d0}}{s_d / \sqrt{n}} = 2.053$$

Se rechaza H_0 . Hay evidencia (con 5% de significancia) de un efecto positivo en el nivel colesterol en sangre con el ejercicio



Tema 9: Regresión

- El modelo de regresión simple.
- Contrastando la regresión.
- Contrastando la regresión con el programa.
- La regresión como suma de cuadrados.

Explicación del examen.



www.unir.net