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

LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

Contenido

CONTRASTES DE HIPÓTESIS

Introducción

H_0 = Hipótesis nula
H_1 = Hipótesis alternativa

Contraste bilateral

$$H_0: \theta = \theta_0 \qquad H_1: \theta \neq \theta_0$$

Contrastes unilaterales

$$H_0: \theta \geq \theta_0 \qquad H_1: \theta < \theta_0$$

$$H_0: \theta \leq \theta_0 \qquad H_1: \theta > \theta_0$$

Dos tipos de error

P(rechazar H_0 | siendo H_0 verdadera) = α

P(aceptar H_0 | siendo H_0 falsa) = β

P-valor

Si $p\ valor > \alpha \rightarrow Aceptamos\ H_0$

Si $p\ valor \leq \alpha \rightarrow Rechazamos\ H_0$

		Cuando realmente es cierto	
		H_0	H_1
Se concluye que es cierto	H_0	Acierto 1-$\boldsymbol{\alpha}$	Error de tipo II $\boldsymbol{\beta}$
	H_1	Error de tipo I $\boldsymbol{\alpha}$	Acierto (potencia = 1-β)

Contrastes una población

Ejemplo: caso bilateral con μ y σ conocida

$$H_0: \mu = \mu_0 \qquad H_1: \mu \neq \mu_0$$

Aceptamos si

$$-z_{\alpha/2} \leq z_{exp} \leq z_{\alpha/2}$$

Rechazamos si

$$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

$$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

Si σ desconocida empleamos T-Student

Contrastes de dos poblaciones

Ejemplo: caso bilateral con μ_1 y μ_2

Caso $\boldsymbol{\sigma}$ desconocido o realista

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$$

Aceptamos si

$$t_{exp} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

Rechazamos si

$$-t_{n+m-2;\,\alpha/2} \leq t_{exp} \leq t_{n+m-2;\,\alpha/2}$$

$$t_{exp} \leq -t_{n+m-2;\,\alpha/2} \text{ o } t_{exp} \geq t_{n+m-2;\,\alpha/2}$$

Introducción

INFERENCIA

Sobre Parámetros,
relaciones,
distribuciones…entre otros

- ☑ Intervalos de confianza
- ☑ Pruebas de hipótesis

Procedimiento estadístico para evaluar y decidir si una afirmación se puede considerar valida (verosímil) estadísticamente a partir de los datos

Introducción

- Una hipótesis estadística es una **proposición** con respecto a alguna característica desconocida de una población de interés.

- La esencia de probar una hipótesis estadística es decidir si la afirmación se encuentra apoyada por la evidencia experimental que se obtiene de una muestra aleatoria (datos).

✓ Se establece que una afirmación de partida a contrastar, que es la llamada **hipótesis nula:** $\boldsymbol{H_0}$

✓ y otra de negación, que se da en caso de no ocurrir la primera (o mejor dicho, de ser rechazada la $\boldsymbol{H_0}$)**,** que es la **hipótesis alternativa: H_1.**

Ejemplo:

¿es cierto que "la media de edad de los fallecidos por covid-19 en todos los países supera los 70 años"?

- La media es mayor a 70 años
- La media no es mayor a 70 años

¿Quién es $\boldsymbol{H_0}$ y $\boldsymbol{H_1}$?

Introducción

$\boldsymbol{H_0}$:

- Es lo que se piensa, **es en principio lo que ya está establecido, lo conservador**, bien porque cierta teoría lo apoya, o bien porque empíricamente está consolidado. También puede establecerse porque tenemos una fuerte intuición de que es cierto algo o así funciona, etc.

- Determina las distribuciones de probabilidad para decidir sobre la validez de las hipótesis.

$\boldsymbol{H_1}$:

Se plantea como lo novedoso, lo que «rompe» con algo establecido o conservador, algo que puede ser extraordinario, aquello que se pretende que sea demostrado. Proposición que el investigador espera probar o sobre la que se espera encontrar evidencia**.**

Introducción

Ejemplo:

¿es cierto que "la media de edad de los fallecidos por covid-19 en todos los países supera los 70 años"?

- La media es mayor a 70 años
- La media no es mayor a 70 años

¿Quién es $\boldsymbol{H_0}$ y $\boldsymbol{H_1}$?

$\boldsymbol{H_0}$:La media no es mayor a 70 años:$\mu \leq 70$

$\boldsymbol{H_1}$:La media es mayor a 70 años: $\mu > 70$

Otras claves:

La igualdad siempre se asocia con la hipótesis nula, por lo que las desigualdades se fijan comúnmente en la hipótesis alterna.

Introducción

Ejemplo:

Si quisiéramos contrastar si las antenas de repetición pueden provocar mayores probabilidades de tener cáncer a las personas que viven en su cercanía, entonces:

- La tasa de incidencia de cáncer en los bloques de viviendas con antenas de repetición instaladas es la misma que en aquellos donde no están instaladas.

- La tasa de incidencia de cáncer en los bloques de viviendas con antenas de repetición instaladas no es la misma. Incluso, puede ser mayor.

Introducción

Ejemplo:

Si quisiéramos contrastar si las antenas de repetición pueden provocar mayores probabilidades de tener cáncer a las personas que viven en su cercanía, entonces:

H_0: La tasa de incidencia de cáncer en los bloques de viviendas con antenas de repetición instaladas **es la misma** que en aquellos donde no están instaladas.

H_1: La tasa de incidencia de cáncer en los bloques de viviendas con antenas de repetición instaladas **no es la misma**. Incluso, **puede ser mayor.**

Introducción

Tipos de maneras de plantear las hipótesis:

A	$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$	Bilateral. Dos colas
B	$H_0: \theta \geq \theta_0$ $H_1: \theta < \theta_0$	Unilateral a izquierda. Cola a izquierda
C	$H_0: \theta \leq \theta_0$ $H_1: \theta > \theta_0$	Unilateral a derecha. Cola a derecha
D	$H_0: \theta_1 \leq \theta \leq \theta_2$ $H_1: \theta < \theta_1 \ o \ \theta > \theta_2$	Bilateral. Dos colas

Introducción

Tipos de error en las pruebas de hipótesis:

Hipótesis Nula: No esta embarazada

Hipótesis Alternativa: esta embarazada

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

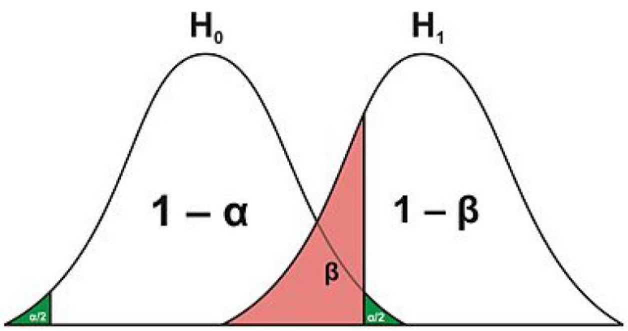
Falso positivo o error tipo I consiste en rechazar la hipótesis nula cuando esta es verdadera. En este caso se dice que esta embarazado cuando no lo esta.

Falso negativo o error tipo II consiste en NO rechazar la hipótesis nula cuando esta es falsa. En este caso se dice que no esta embarazada cuando si lo esta.

Introducción

Tipos de error en las pruebas de hipótesis:

	SITUACION REAL	
	H_0 cierta	H_0 Falsa
Se Rechazo H_0	X Error de Tipo I α	Decisión Correcta $(1-\beta)$
NO Rechazo H_0	Decisión Correcta $(1-\alpha)$	X Error de Tipo II B



- PROB ERROR DE TIPO I : $P(\,Rechazar\ Ho\ /\ \ Ho\ es\ verdadero\,) = \alpha$ (**Nivel de significancia**)
- PROB ERROR DE TIPO II : $P(\,No\ Rechazar\ Ho\ /\ \ Ho\ es\ falso) = \beta$
- CONFIANZA : $P(\,NO\ Rechazar\ Ho\ /\ \ Ho\ es\ verdadero\,) = (1-\alpha)$
- POTENCIA DE UNA PRUEBA : $P(\,Rechazar\ Ho\ /\ \ Ho\ es\ falso) = (1-\beta)$

Introducción

Ejemplo:

En los sistemas legales en general, al acusado se le considera inocente hasta que se demuestre que es culpable. Considera una Hipótesis Nula, en la que el acusado es inocente, y una Hipótesis Alternativa en la que el acusado es culpable. El jurado tiene 2 posibilidades: Encarcelar al acusado o exonerarlo.
Explique los riesgos de cometer un error de tipo I o un error de tipo II.

Introducción

Ejemplo:

	SITUACION REAL	
	H_0 : INOCENTE	H_a : CULPABLE
Se Rechazo H_0	α Decidir que es culpable, dado que es inocente	Decisión Correcta $(1-\beta)$
NO Rechazo H_0	Decisión Correcta $(1-\alpha)$	β Decidir que es inocente, dado que es culpable

- Si se minimiza el error tipo I (α), se desea evitar encarcelar al inocente.
- No se puede aceptar H_0 (lo correcto es no rechazar H_0). No se puede asegurar la inocencia.
- Si se rechaza H_0, es porque se encontró evidencia (en los datos) a favor de H_1.

Introducción

Decisión (Data)

- No rechazo H_0
- Se rechaza H_0

Conclusión (Contexto)

- No hay evidencia a favor de H_1
- Hay evidencia a favor de H_1

Decisión (control del error tipo I) (fijado por el investigador)

Se define $\alpha = 0.05$, nivel de significancia máximo de 0.05. Máximo permito equivocación en el 5% de los casos al rechazar H_0 cuando esta es verdadera.

Niveles de significancia con frecuencia usados en investigación:

$\alpha = 0.01, 0.05, 0.1.$

Introducción

Decisión (control del error tipo I) (fijado por el investigador)

- Se define $\alpha = 0.05$, nivel de significancia máximo de 0.05. Máximo permito equivocación en el 5% de los casos al rechazar H_0

Error real (datos y una distribución de probabilidad)

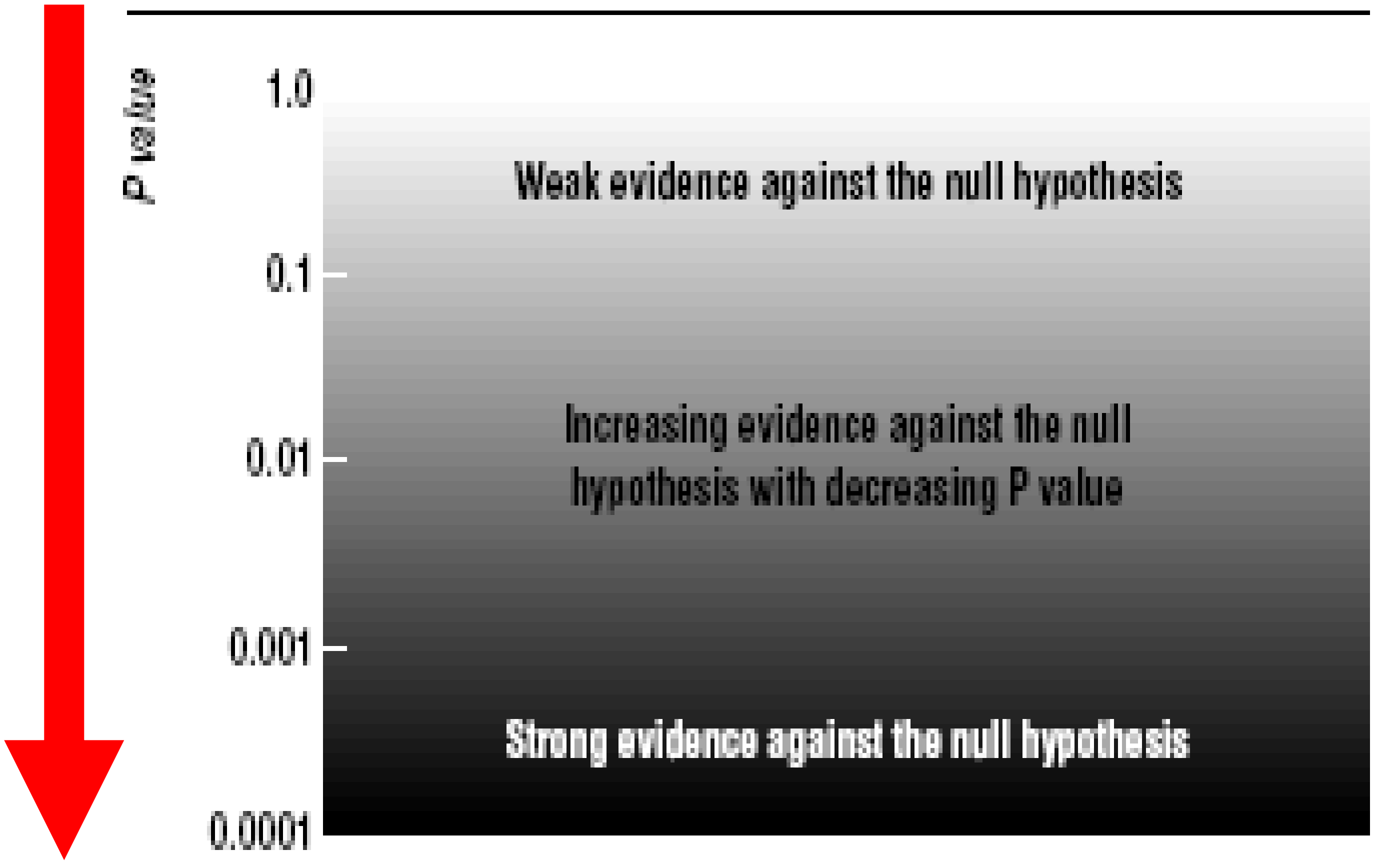
- Del 90%. Mayor al permitido. Mejor no rechazo H_0. Me equivoco en el 90% de los casos al rechazar
- (Es poco probable equivocarse)

P-value = Es el nivel de significancia (probabilidad) observado o "real" a partir de los datos

Introducción

Se calcula un valor de probabilidad, P, que representa la fortaleza de la evidencia en los datos contra H_0

Se interpreta así



Introducción

REGLAS DE DECISION PARA RECHAZAR H_0:

Si $p - value \leq \alpha$: H_0 se rechaza.

Si $p - value > \alpha$: H_0 no se rechaza.

Pasos para realizar un contraste de hipótesis:

✓ Tipos de test:

- Univariado, bivariado o multivariado,
- Paramétrico o no paramétrico
- Sobre:
  - parámetros distribucionales
  - una distribución
  - una relación
  - parámetros de un modelo
  - supuestos de un modelo

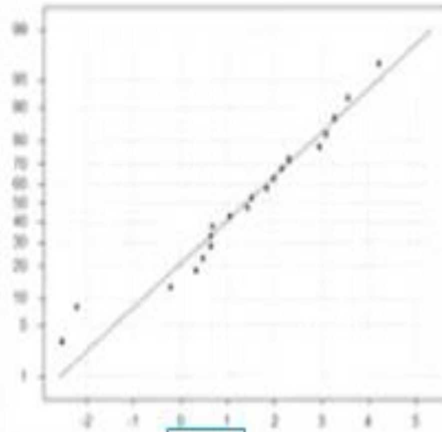
Pasos para realizar un contraste de hipótesis:

EJEMPLOS DE TEST DE HIPÓTESIS

1

¿Están correlacionadas las variables X e Y?

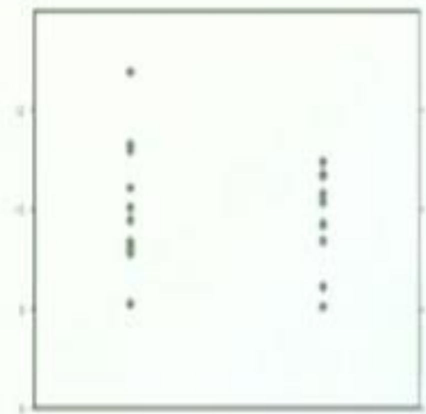
$H_o: \rho = 0$



2

¿Las observaciones siguen una distribución normal?

$H_o: X \sim N$



3

¿La varianza de A y B son iguales?

$H_o: \sigma_A^2 = \sigma_B^2$

Pasos para realizar un contraste de hipótesis:

1. Definir hipótesis nula y alternativa
2. Identificar si es una prueba unilateral o bilateral
3. Definir nivel de significancia
4. Definir estadístico de prueba (distribución de probabilidad que permite hallar el P-value o tomar la decisión).
   - Depende del **parámetro**(s) de interés, la información conocida (**supuestos distribucionales**) y los datos a través del estimador (Z_c, t_c, X_c^2, F_c).
   - Se contrasta la información bajo H_o del parámetro de interés con la información derivada del estimador del parámetro.
   - Halla el p-value o los valores críticos.
5. Decisión sobre el rechazo o no de H_o,
   - Usando p-value
   - Usando valores críticos
6. Conclusión en función de H_1

Pasos para realizar un contraste de hipótesis:

Ejemplo: Tenemos las siguientes mediciones de los mg/l de hierro en sangre de un paciente:

2,4 2,2 2,5 3 3,2 3,3 3 3,4 3,2 3,3

Queremos saber si es realista pensar que la media de mg/l de hierro en sangre es de 2,4 mg/l (nivel "aceptable") o diferente

X = «Concentración medida en una determinación en mg/l».
Supongamos que la variable aleatoria es normal y que se conoce $\sigma^2 = 0.1$

Solución

1 y 2.

$$H_0: \mu = 2.4$$
$$H_1: \mu \neq 2.4$$

Contraste bilateral

3. Fijamos $\alpha = 0.01$

4. Necesitamos una distribución de probabilidad que involucre μ, su estimador ($\bar{x}$) (**datos**) y que asuma normalidad en la variable

$$z_c = \frac{\bar{x} - \overbrace{\mu}^{?}}{\sigma / \sqrt{n}} = 5.50$$

$\mu_0 = 2.4$
$\bar{x}$=2.95
$n = 10$
$\sigma^2 = 0.1$

Pasos para realizar un contraste de hipótesis:

Solución

1 y 2.

$$H_0: \mu = 2.4$$
$$H_1: \mu \neq 2.4$$

Contraste bilateral

3. Fijamos $\alpha = 0.01$

4. Necesitamos una distribución de probabilidad que involucre μ, su estimador ($\bar{x}$) (**datos**) y que asuma normalidad en la variable

$$z_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 5.50$$

$\mu_0 = 2.4$
$\bar{x}$=2.95
$n = 0.1$
$\sigma^2 = 0.1$

5. Decisión:

$$P - value = 2 * P(Z > z_c) = 3.8 * 10^{-8}$$

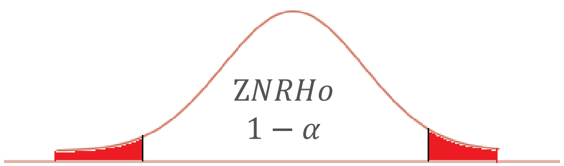
Con base en los datos de la muestra nos equivocamos menos de lo permitido al rechazar la H_0, casi en el 0% de las veces.

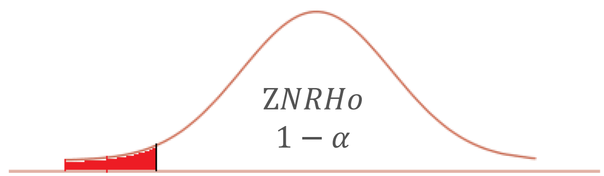
Se rechaza H_0

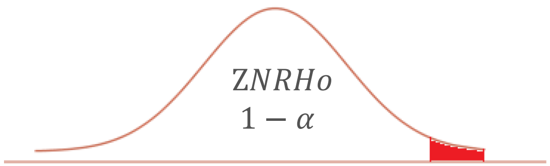
5. Conclusión:

Hay evidencia estadística suficiente para afirmar que la media en sangre del paciente es diferente de 2.4

Prueba de Hipótesis para la media (μ) con la varianza (σ^2) conocida

<table>
<tr><th>Prueba Bilateral</th><th colspan="2">Prueba Unilateral</th></tr>
<tr><td>$H_0: \mu = \mu_0$
$H_a: \mu \neq \mu_0$</td><td>$H_0: \mu \geq \mu_0$
$H_a: \mu < \mu_0$</td><td>$H_0: \mu \leq \mu_0$
$H_a: \mu > \mu_0$</td></tr>
<tr><td>

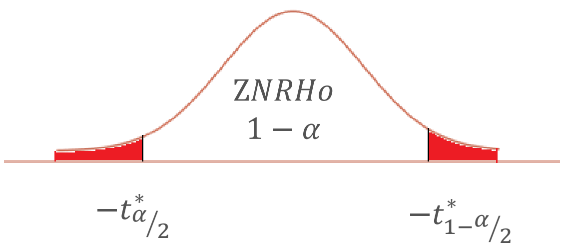
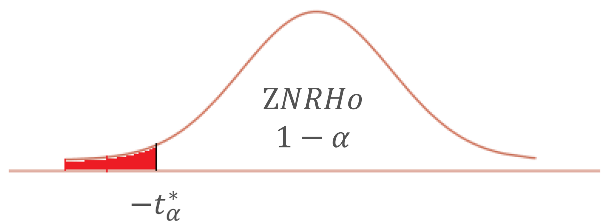
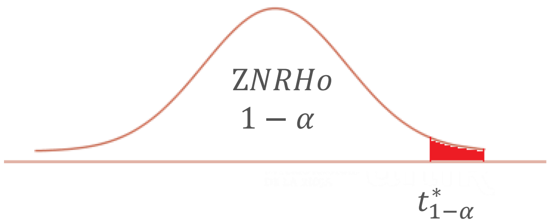
$Z^*_{\alpha/2}$ $Z^*_{1-\alpha/2}$</td><td>

$-Z^*_{\alpha}$</td><td>

$Z^*_{1-\alpha}$</td></tr>
<tr><td>$P - value = 2 * P(Z > z_c)$</td><td>$P - value = P(Z < z_c)$</td><td>$P - value = P(Z > z_c)$</td></tr>
</table>

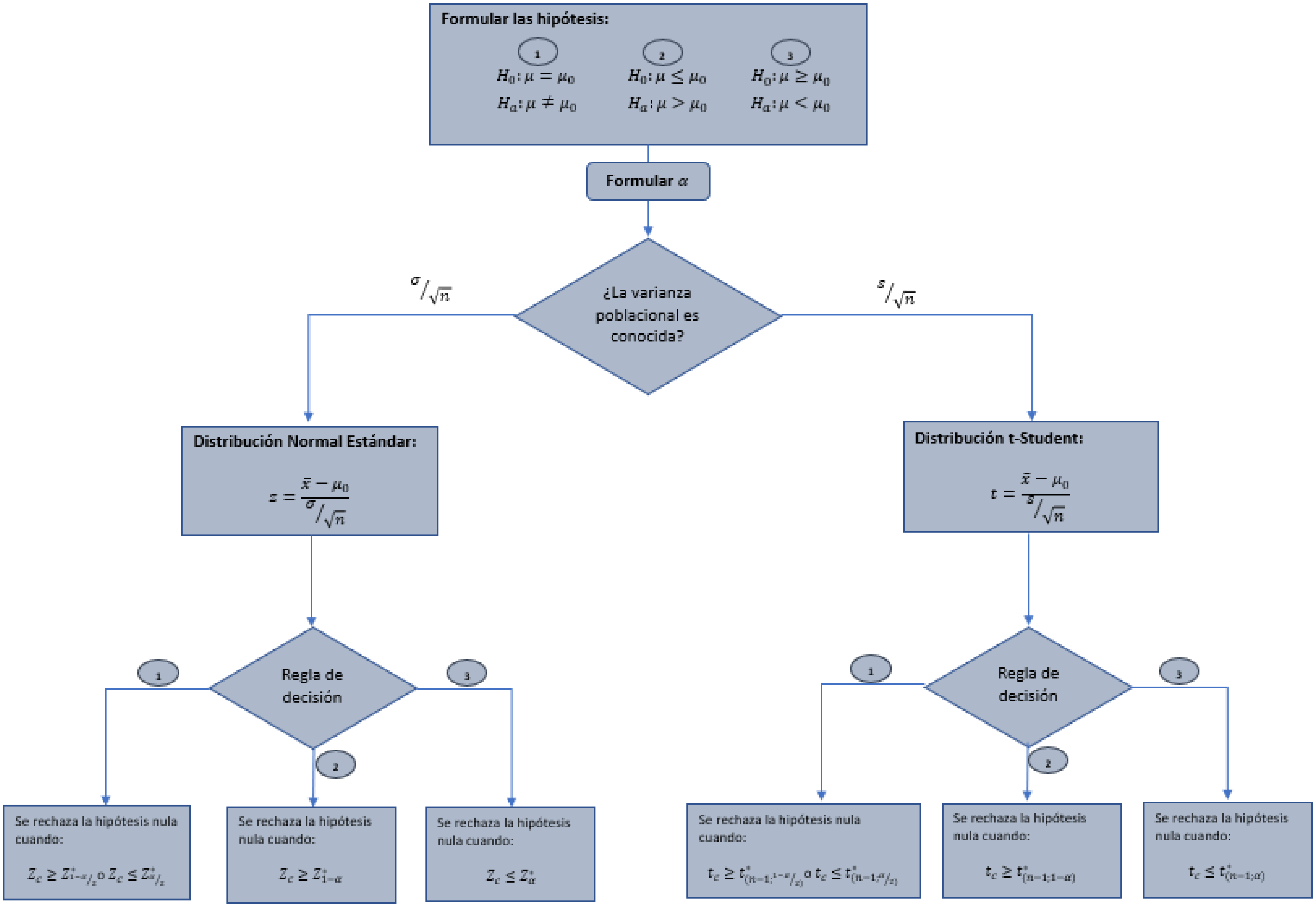
$$Valor\ calculado\ Z \rightarrow z_c = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Prueba de Hipótesis para la media (μ) con la varianza (σ^2) desconocida

Prueba Bilateral	Prueba Unilateral	
$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$
		
$P - value = 2 * P(t > t_c)$	$P - value = P(t < t_c)$	$P - value = P(t > t_c)$

$$Valor\ calculado\ t \rightarrow t_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Prueba de Hipótesis para la media (μ)



Prueba de Hipótesis para la media (μ)

Ejemplo: Tenemos las siguientes mediciones de los mg/l de hierro en sangre de un paciente:

2,4	2,2	2,5	3	3,2	3,3	3	3,4	3,2	3,3

X = «Concentración medida en una determinación en mg/l».
Supongamos que la variable aleatoria es normal y que no se conoce σ^2

Solución

1 y 2.

$$H_0: \mu = 2.4$$
$$H_1: \mu \neq 2.4$$

Contraste bilateral

3. Fijamos $\alpha = 0.01$

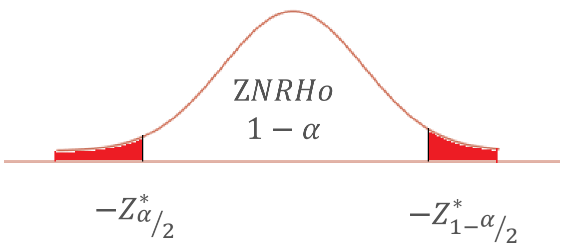
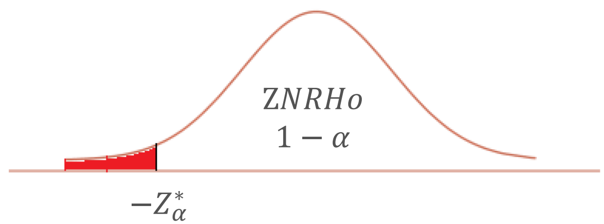
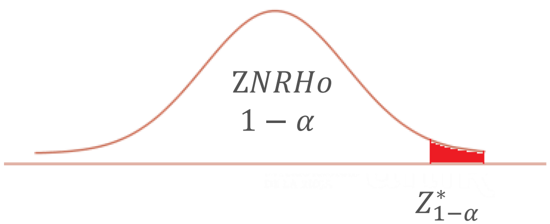
4. Necesitamos una distribución de probabilidad que involucre μ, su estimador ($\bar{x}$) (**datos**) y que asuma normalidad en la variable

$$t_c = \frac{\overbrace{\bar{x} - \mu}^{?}}{s/\sqrt{n}} = 4.06$$

$\mu_0 = 2.4$
$\bar{x}$=2.95
$n = 10$
GL =9
s =4.07

5. Decisión: $P - value = 2 * P(t > t_c) = 0.0028$ **Se rechaza H_0**

Prueba de Hipótesis para la proporción (P)

Prueba Bilateral	Prueba Unilateral	
$H_0: P = P_0$ $H_1: P \neq P_0$	$H_0: P \geq P_0$ $H_1: P < P_0$	$H_0: P \leq P_0$ $H_1: P > P_0$
		
$P - value = 2 * P(Z > Z_c)$	$P - value = P(Z < Z_c)$	$P - value = P(Z > Z_c)$

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

Donde $\hat{p}$ la proporción muestral, n es el tamaño de la muestra y P_0, es la proporción poblacional (hipotética)

Próxima sesión

Tema 8_II: Contraste de hipótesis

- Contraste paramétrico para la varianza
- Contrastes paramétricos para dos muestras.
- Contrastes de hipótesis robustos.

Sesión doble

unir

LA UNIVERSIDAD
EN INTERNET

www.unir.net