Tema 3: Estimación estadística de modelos probabilistas. (tercera parte)

Estructura de este tema:

1. Técnicas de muestreo y estimación puntual.
2. Estimación por intervalos de confianza.
3. **Contrastes de hipótesis**.

Contraste de Hipótesis Paramétrico:

Dada una variable poblacional X, asumimos que *conocemos su función de distribución*, F_θ, pero el parámetro θ *es desconocido*: $X \sim F_\theta$.
Un Contraste de Hipótesis Paramétrico es la técnica estadística que se usa para decidir si una afirmación o **hipótesis** sobre el parámetro poblacional θ se acepta o se rechaza a partir de los datos de una muestra extraída de dicha población.

Contraste de Hipótesis no paramétrico

Dada una variable poblacional X, *queremos averiguar si sigue una determinada distribución* F_θ: ¿$X \sim F_\theta$?. En este caso tanto la función de distribución F como el parámetro poblacional θ son desconocidos.

Planteamiento del problema de contraste de hipótesis paramétrico: Hipótesis nula e hipótesis alternativa.

- **Hipótesis nula (H_0):** Es la hipótesis que mantendremos como cierta a no ser que los datos muestren de forma evidente su falsedad.
- **Hipótesis alternativa (H_1):** Es la hipótesis que refleja la situación contraria a la hipótesis nula.

Ejemplo "Principio de presunción de inocencia": Como se quiere evitar condenar a una persona inocente, sólo se hará cuando haya quedado claramente demostrada su culpabilidad. En caso de duda, se primará la inocencia frente a la culpabilidad. El contraste propuesto sería

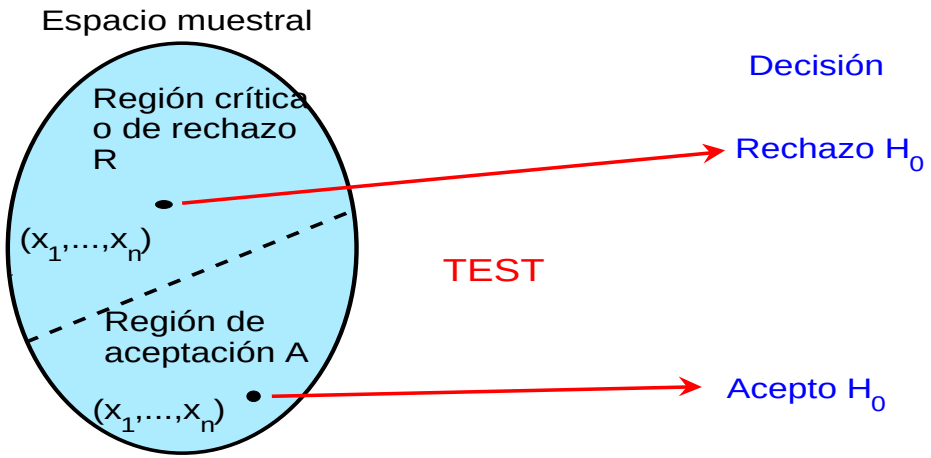
$$H_0: \text{ Inocente}$$
$$H_1: \text{ Culpable}$$

Obs: Los contrastes que estudiamos son muy *conservadores* con H_0: *sólo rechazamos H_0 cuando hay una fuerte evidencia muestral en su contra.*

Elementos de un Contraste de Hipótesis Paramétrico

Paso 1. **Plantear el contraste**: Se establece una hipótesis básica, que llamaremos *hipótesis nula* y denotaremos H_0, y contraria a ella otra hipótesis H_1, o *hipótesis alternativa*.

Paso 2. Fijar una **regla de decisión** para aceptar o rechazar H_0: un test estadístico queda perfectamente especificado, en cuanto hayamos definido la **región de rechazo de H_0**.



Paso 3. **Tomar la decisión**: Con los datos recogidos a partir de la muestra, se decide finalmente si se acepta o se rechaza H_0.

	Rechazar H_0	Aceptar H_0
H_0 cierta	**Error de Tipo I**	**Decisión correcta**
H_0 falsa	**Decisión correcta**	**Error de Tipo II**

Ejemplo "presunción de inocencia" (cont.):

H_0 : Inocente

H_1 : Culpable

	Rechazar H_0	Aceptar H_0
H_0 cierta	Se condena a un inocente	Se absuelve a un inocente
H_0 falsa	Se condena a un culpable	Se absuelve a un culpable

De los dos errores *sólo vamos a poder controlar el Error de Tipo I*, por ello, se deben definir las hipótesis de forma que éste sea el más grave. Para ello: H_1 *debe ser la hipótesis que queremos confirmar*. Lo que se hace es fijar un *nivel de significación* $\alpha \in (0, 1)$ tal que

$$\alpha = P(\text{Error de tipo I}) = P(\text{Rechazar } H_0 \,|\, H_0 \text{ cierta})$$

Elección de la hipótesis nula H_0

Los contrastes de hipótesis se clasifican en:

- Contraste bilateral:

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

En el caso bilateral no hay duda de cómo elegir H_0: la de la "$=$".

- Contraste unilateral:

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

En el caso unilateral, y teniendo en cuenta que los contrastes de hipótesis son conservadores con H_0, debemos tomar como H_1 aquella que deseemos "probar" estadísticamente, es decir, H_0 *debe de ser la contraria de la que queremos "probar".*

Construcción de la región de rechazo R

Ejemplo "ilustrativo": La resistencia de ciertos componentes eléctricos fabricados en un proceso es una v.a. que sigue una distribución $N(\mu, \sigma)$. Una empresa que suministra dichos componentes asegura, basada en estudios previos, que la resistencia media es superior a 18 ohmios. Para simplificar la cuestión asumimos σ conocido.

Se trata de un contraste unilateral con $\mu_0 = 18$, con el contraste de hipótesis propuesto por la empresa:

$$H_0 : \mu \leq 18$$
$$H_1 : \mu > 18$$

Sólo se rechazará H_0 si los datos muestrales dejan clara evidencia de lo contrario, e.d. si la media muestral, $\bar{X}$, es "mucho mayor" de $\mu_0 = 18$: $R = \{\bar{X} - \mu_0 \geq C\}$ para cierto valor de C a determinar.

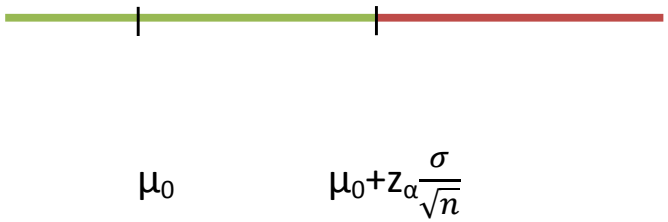
El grado de "alejamiento" permitido dependerá del tamaño muestral n, y del nivel de siginificación α, fijado:

$$\alpha = P(R) = P(\bar{X} - \mu_0 \geq C)$$

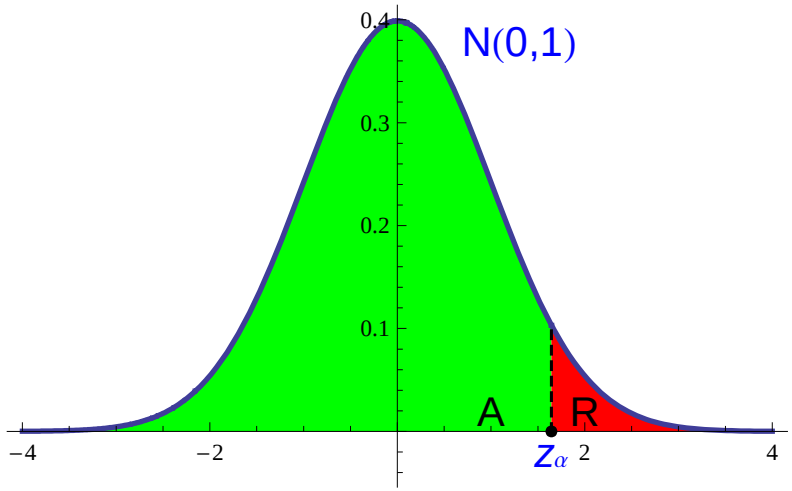
$$\underset{tipificamos}{=} P\Big(\underbrace{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}_{\sim N(0,1)} \geq \underbrace{\frac{C}{\sigma/\sqrt{n}}}_{z_\alpha}\Big) \quad \rightarrow \quad C = z_\alpha \frac{\sigma}{\sqrt{n}}$$

Por lo tanto

$$R = \left\{\bar{x} - \mu_0 \geq z_\alpha \frac{\sigma}{\sqrt{n}}\right\} \equiv \left\{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha\right\}$$



=



Contrastes en poblaciones normales

- Sea $X_1, \ldots, X_n$ una m.a. de $X \sim N(\mu, \sigma)$ con σ conocido.

$$H_0 : \mu = \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : |\bar{x} - \mu_0| \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

$$H_0 : \mu \leq \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : \bar{x} - \mu_0 \geq z_{\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

$$H_0 : \mu \geq \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : \bar{x} - \mu_0 \leq z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

Contrastes en poblaciones normales

- Sea $X_1, \ldots, X_n$ una m.a. de $X \sim N(\mu, \sigma)$ con σ desconocido.

$$H_0 : \mu = \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : |\bar{x} - \mu_0| \geq t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} \right\}$$

$$H_0 : \mu \leq \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : \bar{x} - \mu_0 \geq t_{n-1;\alpha} \frac{s}{\sqrt{n}} \right\}$$

$$H_0 : \mu \geq \mu_0 \qquad R = \left\{ (x_1, \ldots, x_n) : \bar{x} - \mu_0 \leq t_{n-1;1-\alpha} \frac{s}{\sqrt{n}} \right\}$$

$$H_0 : \sigma = \sigma_0 \qquad R = \left\{ \frac{(n-1)s^2}{\sigma_0^2} \notin (\chi^2_{n-1;1-\alpha/2}, \chi^2_{n-1;\alpha/2}) \right\}$$

$$H_0 : \sigma \leq \sigma_0 \qquad R = \left\{ \frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{n-1;\alpha} \right\}$$

$$H_0 : \sigma \geq \sigma_0 \qquad R = \left\{ \frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1;1-\alpha} \right\}$$

Ejemplo 3.10: Se certifica que un material estándar de referencia de un suelo contiene 94,6 ppm de un contaminante orgánico. Un análisis repetido arrojó los siguientes resultados:

$$98,6 \qquad 98,4 \qquad 97,2 \qquad 94,6 \qquad 96,2 \qquad \text{ppm}$$

A un nivel de significación $\alpha = 0,05$ ¿hay suficiente evidencia estadística para concluir que los resultados mantienen el valor esperado?
Si se hace una medida más y se obtiene 94,5 ¿cambiaría la respuesta?

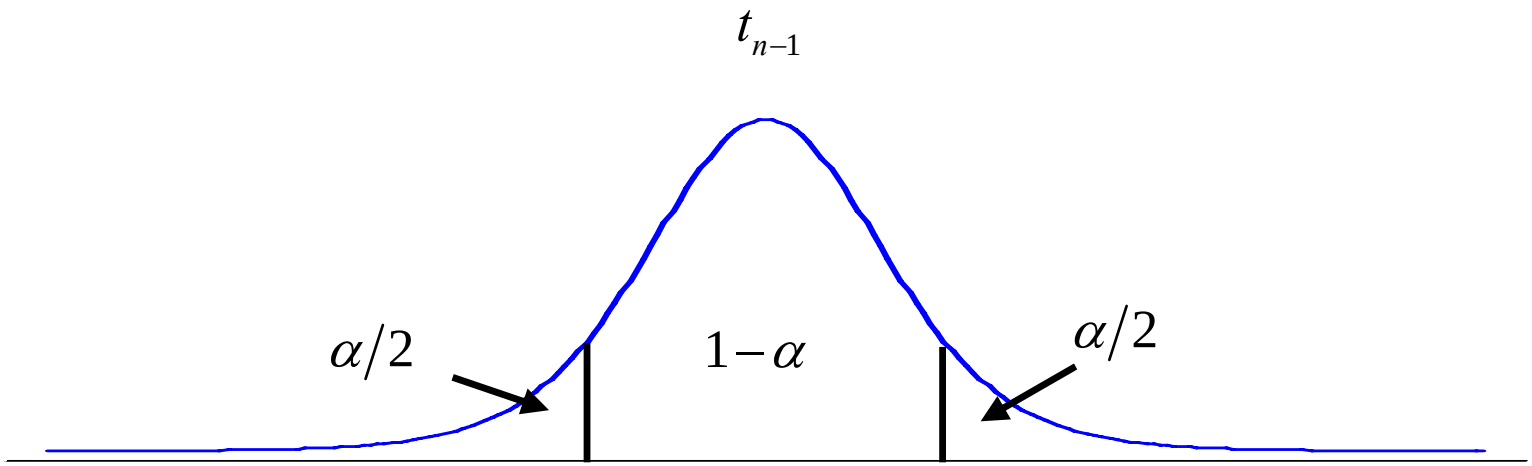
Relación entre contrastes de hipótesis bilaterales e intervalos de confianza

En un contraste bilateral con nivel de significancia α y región de rechazo R_α

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

se verifica:

- Rechazar H_0 con nivel de significación $\alpha \Leftrightarrow \theta \notin IC_{1-\alpha}(\theta)$.
- No rechazar H_0 con nivel de significación $\alpha \Leftrightarrow \theta \in IC_{1-\alpha}(\theta)$.



Ejemplo "ilustrativo" (cont.): Dada la variable $X =$"resistencia de ciertos componentes eléctricos" $\sim N(\mu, \sigma)$ (asumimos ahora la situción, mucho más real, de σ desconocida).

Paso 1. Plantear el contraste. Se trata de estudiar el contraste unilateral con $\mu_0 = 18$:

$$H_0 : \mu \leq 18$$
$$H_1 : \mu > 18$$

Para comprobar la veracidad de la empresa, la persona que supervia los componentes eléctricos obtiene una muestra de tamaño 10 y obtiene los siguientes resultados muestrales: $\bar{x} = 19.02$ y $s = 1.196$.

La pregunta que trata de responder un contraste de hipótesis es, a partir de los resultados muestrales que tenemos
¿proporcionan estos datos evidencia de que la resistencia media de los componentes es mayor que 18 ohmios?

Ejemplo "ilustrativo" (cont.):

Paso 2. Fijar una **regla de decisión**. Queremos contrastar $H_0 : \mu \leq 18$ frente a $H_1 : \mu > 18$ a nivel $\alpha = 0,05$.

Como σ no es conocida utilizamos la región de rechazo:

$$R = \left\{ \bar{x} - \mu_0 \geq t_{n-1;\alpha} \frac{s}{\sqrt{n}} \right\} = \left\{ \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \geq t_{n-1;\alpha} \right\}$$

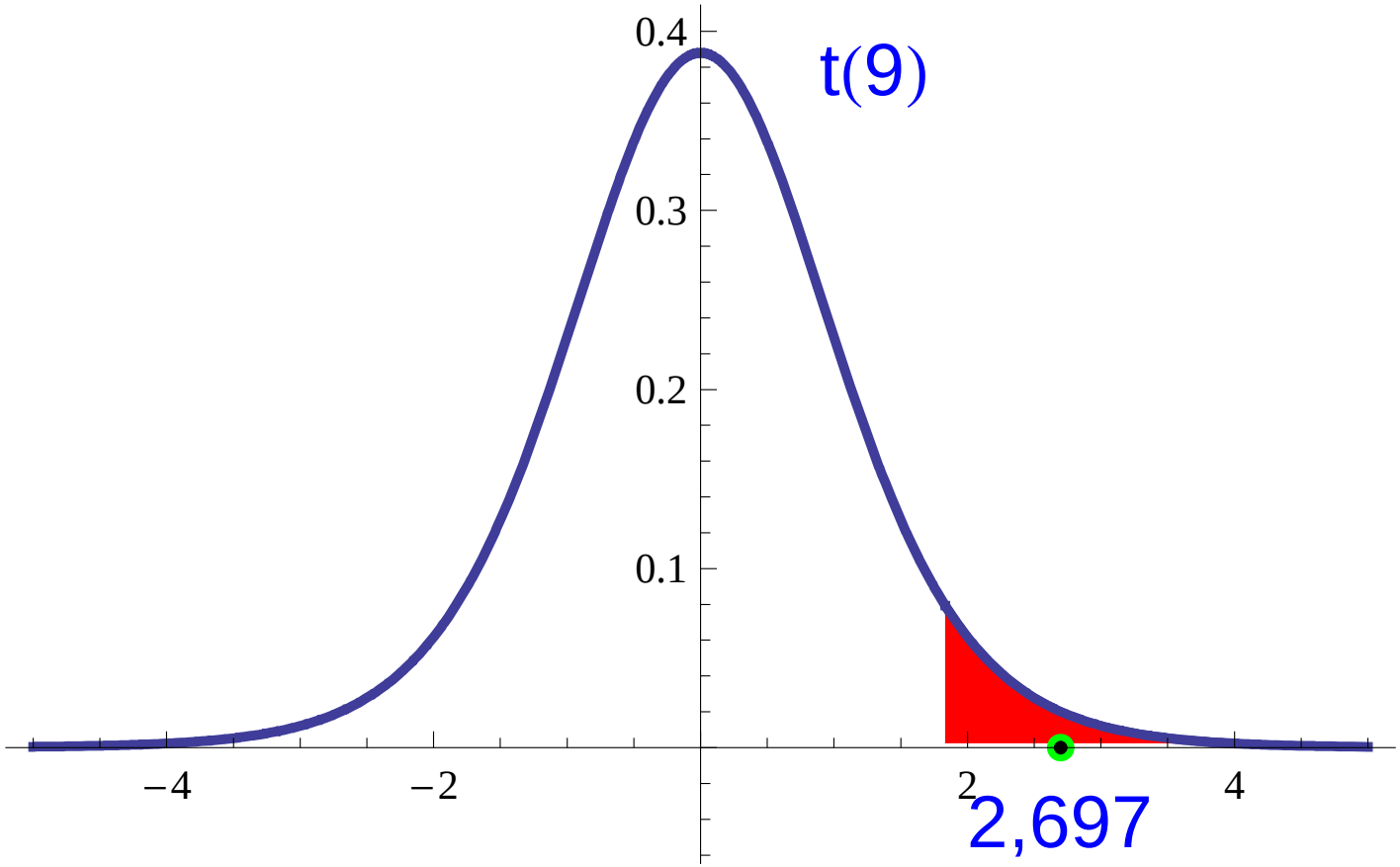
donde

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.02 \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = 1.196$$

$$t_{9;0,05} = 1,833 \qquad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.02 - 0}{1.196/\sqrt{10}} = 2.697$$

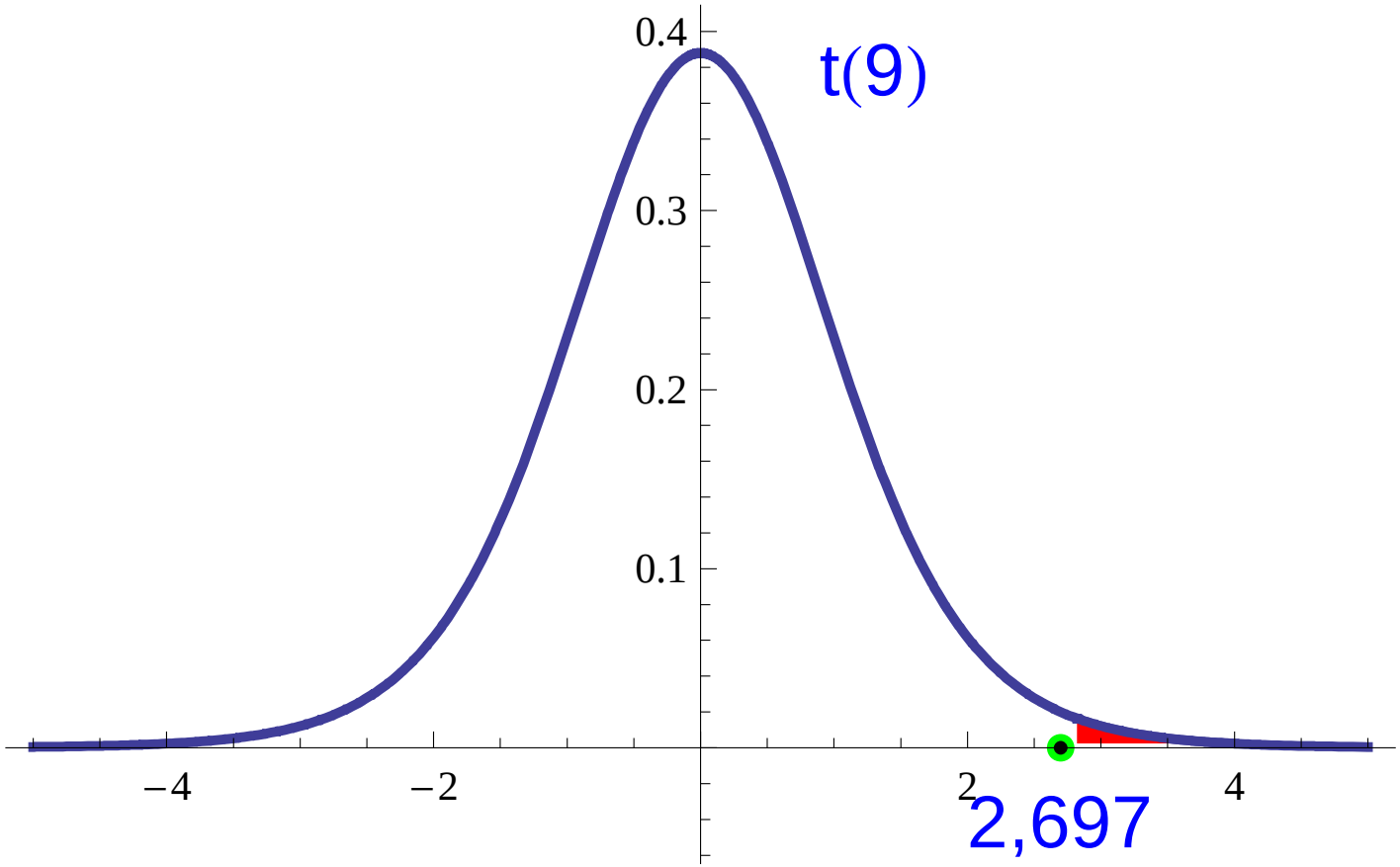
Ejemplo "ilustrativo" (cont.):

Paso 3. Tomar la decisión. Como $2,697 > t_{9;0,05} = 1,833$ estamos en la región crítica y rechazamos H_0 a nivel $\alpha = 0,05$.



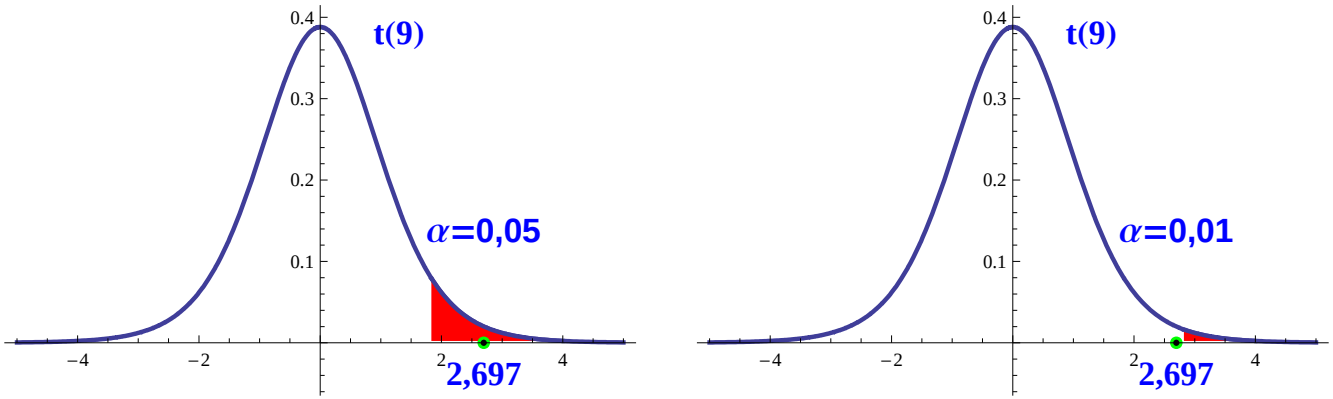
Ejemplo "ilustrativo" (cont.):

¿Cuál es la conclusión si fijamos $\alpha = 0.01$?

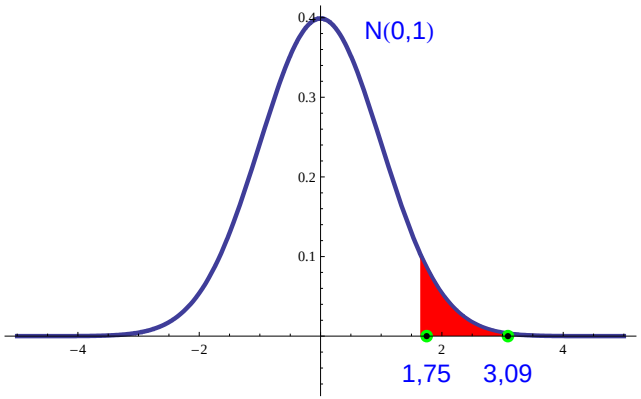


Crítica a la selección del nivel de significación α

(a) El resultado del test puede depender mucho del valor de α elegido, que es arbitrario, siendo posible rechazar H_0 con $\alpha = 0.05$ y aceptarlo con $\alpha = 0.01$.



(b) Dar sólo el resultado del test (aceptación o rechazo) no permite diferenciar el grado de evidencia que la muestra indica a favor o en contra de H_0.



El p-valor de un contraste

A medida que el nivel de significación disminuye es más difícil rechazar la hipótesis nula (manteniendo los mismos datos).

Hay un valor del nivel de significación llamado ***p*-valor**, a partir del cual ya no podemos rechazar H_0. Es decir, para cualquier nivel de significación menor que el p-valor: $\alpha < p$-valor; **no se rechaza** H_0.

Interpretación: El p-valor indica el punto de división entre el rechazo y la aceptación:

- Si $\alpha < p$-valor, no podemos rechazar H_0 a nivel α.
- Si $\alpha > p$-valor, podemos rechazar H_0 a nivel α.

Utilización: Es más informativo utilizar el p-valor que fijar un nivel de significación ya que proporciona una idea de hasta qué punto la información muestral soporta H_0.

Contrastes para dos poblaciones normales

- Sean $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ m.a. independientes de $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$ respectivamente con σ_1 y σ_2 conocidas. X e Y son v.a. independientes.

$$H_0 : \mu_1 = \mu_2 \qquad R = \left\{ |\bar{x} - \bar{y}| \geq z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

$$H_0 : \mu_1 \leq \mu_2 \qquad R = \left\{ \bar{x} - \bar{y} \geq z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

$$H_0 : \mu_1 \geq \mu_2 \qquad R = \left\{ \bar{x} - \bar{y} \leq z_{1-\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\}$$

Contrastes para dos poblaciones normales

- Sean $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ m.a. independientes de $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$ respectivamente con $\sigma_1 = \sigma_2$ desconocida. X e Y son v.a. independientes.

$$H_0 : \mu_1 = \mu_2 \qquad R = \left\{ |\bar{x} - \bar{y}| \geq t_{n_1+n_2-2;\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

$$H_0 : \mu_1 \leq \mu_2 \qquad R = \left\{ \bar{x} - \bar{y} \geq t_{n_1+n_2-2;\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

$$H_0 : \mu_1 \geq \mu_2 \qquad R = \left\{ \bar{x} - \bar{y} \leq t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

donde

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Contrastes para dos poblaciones normales

- Sean $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ m.a. independientes de $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$ respectivamente con $\sigma_1 \neq \sigma_2$ desconocidas. X e Y son v.a. independientes.

$$H_0 : \mu_1 = \mu_2 \qquad R = \left\{ |\bar{x} - \bar{y}| \geq t_{f;\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\}$$

$$H_0 : \mu_1 \leq \mu_2 \qquad R = \left\{ \bar{x} - \bar{y} \geq t_{f;\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\}$$

$$H_0 : \mu_1 \geq \mu_2 \qquad R = \left\{ \bar{x} - \bar{y} \leq t_{f;1-\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\}$$

donde f es el entero más próximo a $\dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$.

Contrastes para dos poblaciones normales

- Sean $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ m.a. independientes de $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$ respectivamente con σ_1 y σ_2 desconocidas. X e Y son v.a. independientes.

$$H_0 : \sigma_1 = \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} \notin (F_{n_1-1;n_2-1;1-\alpha/2}, F_{n_1-1;n_2-1;\alpha/2}) \right\}$$

$$H_0 : \sigma_1 \leq \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} > F_{n_1-1;n_2-1;\alpha} \right\}$$

$$H_0 : \sigma_1 \geq \sigma_2 \qquad R = \left\{ \frac{s_1^2}{s_2^2} < F_{n_1-1;n_2-1;1-\alpha} \right\}$$

Ejemplo 3.11: Con el objeto de averiguar si difieren de forma significativa los pesos entre nadadoras y corredoras olímpicas se eligieron al azar 13 nadadoras y 10 corredoras y se observó:

	Nadadoras	Corredoras
Tamaño muestral	13	10
Media muestral	63.9	67.8
Cuasi-desviación típica muestral	9.16	8.37

Asumiendo normalidad, con nivel de significación $\alpha = 0.1$, ¿qué conclusión podemos extraer de estos datos?
El p-valor ¿es mayor o menor que 0.1?

Contrastes sobre una proporción p de una Bernoulli

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una v.a. X$\sim$Bernoulli(p). Suponemos que n es grande.

$$H_0 : p = p_0 \qquad R = \left\{ |\bar{x} - p_0| > z_{\alpha/2} \sqrt{\frac{p_0(1-p_0)}{n}} \right\}$$

$$H_0 : p \leq p_0 \qquad R = \left\{ \bar{x} - p_0 > z_{\alpha} \sqrt{\frac{p_0(1-p_0)}{n}} \right\}$$

$$H_0 : p \geq p_0 \qquad R = \left\{ \bar{x} - p_0 < z_{1-\alpha} \sqrt{\frac{p_0(1-p_0)}{n}} \right\}$$

Comparación de dos proporciones Bernoulli

Sean $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ muestras de $X \sim \text{Bernoulli}(p_1)$ e $Y \sim \text{Bernoulli}(p_2)$, v.a. independientes. Entonces

$$H_0 : p_1 = p_2 \qquad R = \left\{ |\bar{x} - \bar{y}| > z_{\alpha/2} \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right\}$$

$$H_0 : p_1 \leq p_2 \qquad R = \left\{ \bar{x} - \bar{y} > z_{\alpha} \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right\}$$

$$H_0 : p_1 \geq p_2 \qquad R = \left\{ \bar{x} - \bar{y} < z_{1-\alpha} \sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right\}$$

donde $\bar{p} = \dfrac{\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j}{n_1 + n_2} = \dfrac{n_1 \bar{x} + n_2 \bar{y}}{n_1 + n_2}$.

Ejemplo 3.12: El presidente de un cierto partido político quiere averiguar si saldrá elegido en las próximas elecciones. Para ello, como sabemos, seleccionó una m.a.s. de 1000 habitantes a los que les preguntó si tenían intención de votarle o no, a lo que el 55% respondió que sí. A la vista de estos datos y con $\alpha = 0.01$ ¿podemos afirmar que saldrá elegido?

Contrastes sobre una proporción λ de una Poisson

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una v.a. X$\sim$Pois(λ). Suponemos que n es grande.

$$H_0 : \lambda = \lambda_0 \qquad R = \left\{ |\bar{x} - \lambda_0| > z_{\alpha/2} \sqrt{\frac{\lambda_0}{n}} \right\}$$

$$H_0 : \lambda \leq \lambda_0 \qquad R = \left\{ \bar{x} - \lambda_0 > z_{\alpha} \sqrt{\frac{\lambda_0}{n}} \right\}$$

$$H_0 : \lambda \geq \lambda_0 \qquad R = \left\{ \bar{x} - \lambda_0 < z_{1-\alpha} \sqrt{\frac{\lambda_0}{n}} \right\}$$

Datos emparejados:

Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra aleatoria de (X, Y) donde X e Y no son independientes, pero los pares (X_i, Y_i), para $i = 1, \ldots, n$, son independientes entre sí.

Denotemos $E(X) = \mu_1$ y $E(Y) = \mu_2$ y supongamos que $D = X - Y \sim N(\mu = \mu_1 - \mu_2, \sigma)$. Entonces $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ es una muestra aleatoria de D.

Podemos realizar los siguientes contrastes de hipótesis basándonos en los tests "correspondientes":

$$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_0 : \mu = 0$$

$$H_0 : \mu_1 \leq \mu_2 \Leftrightarrow H_0 : \mu \leq 0$$

$$H_0 : \mu_1 \geq \mu_2 \Leftrightarrow H_0 : \mu \geq 0$$

Ejemplo 3.13:

Se usan cinco dosis de una sustancia ferrosa para determinar si existen diferencias entre llevar a cabo un análisis químico de laboratorio o un análisis de fluorescencia por rayos X para determinar el contenido de hierro. Cada dosis se divide en dos partes iguales a las que se aplica cada uno de los dos procedimientos. Los resultados obtenidos son los siguientes:

Dosis	1	2	3	4	5
Rayos X	2.0	2.0	2.3	2.1	2.4
Análisis Químico	2.2	1.9	2.5	2.3	2.4

Se supone que las poblaciones son normales. ¿Aportan los datos suficiente evidencia a nivel $\alpha = 0.05$ para afirmar que el contenido medio de hierro determinado con el análisis químico es diferente del contenido medio determinado cuando se utilizan rayos X?

Ejemplo 3.13 (cont.):

Variables:

- X es el contenido de hierro detectado por rayos X
- Y es el contenido de hierro detectado por análisis químico.

Parámetros:

- μ_1 es el contenido medio detectado por rayos X
- μ_2 es el contenido medio detectado por análisis químico.

Hipótesis: Cuando las muestras **no son independientes**, lo reducimos a una sola variable

- $D = X - Y$: diferencia en el contenido de hierro detectado por ambos procedimientos, $D \sim N(\mu, \sigma)$
- $\mu = \mu_1 - \mu_2$: diferencia en el contenido medio detectado por ambos procedimientos

y en lugar de contrastar $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$, se contrasta

$$H_0 : \mu = 0 \quad \text{frente a} \quad H_1 : \mu \neq 0,$$

Ejemplo 3.13 (cont.):

Dosis	1	2	3	4	5
x_i	2.0	2.0	2.3	2.1	2.4
y_i	2.2	1.9	2.5	2.3	2.4
d_i	-0.2	0.1	-0.2	-0.2	0

Región crítica:

$$R = \left\{ |\bar{d} - \mu_0| > t_{n-1;\alpha/2} \frac{s_d}{\sqrt{n}} \right\}$$

donde $\bar{d} = -0.1$, $s_d = 0.1414$ y $t_{4;0.025} = 2.776$.

Por tanto $R = \{0.1 > 0.18\}$ **no se cumple**, es decir los datos disponibles no permiten afirmar a nivel 0.05 que los dos métodos proporcionan cantidades medias de hierro diferentes.

A diferencia de los paramétricos, ahora, el desconocimiento de la población que vamos a estudiar no se reduce al valor de un parámetro poblacional θ, sino que es mucho más amplio.

La pregunta que nos hacemos ahora es: *¿Es razonable aceptar esa modelización, a la vista de los datos disponibles?*

El estadístico que se usará para realizar el contraste tendrá una distribución χ^2.

Contraste de Bondad de Ajuste: primer caso

Sea X una variable aleatoria poblacional con distribución desconocida. Extraemos una m.a.s. de la población $(X_1, \cdots, X_n)$. A la vista de la muestra y para un θ *conocido*, **¿es razonable admitir que X sigue la distribución F_θ?**

$$\begin{aligned} H_0&: \quad X \text{ sigue la distribución } F_\theta \\ H_1&: \quad \text{X no sigue la distribución } F_\theta \end{aligned}$$

Paso 1.: Hacer una partición del espacio muestral (posibles valores de X) en k clases $A_1, \cdots, A_k$.

Paso 2.: Calcular las siguientes frecuencias absolutas:

$O_i =$ **frecuencia observada en** $A_i =$ número de elementos de la m.a.s $(x_1, \cdots, x_n)$ que se han situado en la clase A_i

$e_i =$ **frecuencia esperada en** A_i **si** H_0 **es cierta** $= nP(A_i)$

Paso 3.: Utilizar el estadístico

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$$

Observación 1: Si H_0 es cierta, las frecuencias observadas, O_i, y las frecuencias esperadas, e_i, deben ser parecidas y, por tanto, $\sum \frac{(O_i - e_i)^2}{e_i} \approx 0$.

Observación 2: Para que realmente $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$, además de que n debe ser grande, se requiere $e_i \geq 5$.

Conclusión: Rechazamos H_0 si:

$$R = \left\{ \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} > \chi^2_{k-1,\alpha} \right\}$$

Nota: Por comodidad, normalmente se usa la siguiente expresión, equivalente a la ya dada:

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{O_i^2}{e_i} - n$$

Ejemplo 3.14: Se quiere averiguar si el número de hijos por matrimonio, X, sigue una distribución binomial de parámetros $n = 3$ y $p = 0.5$. Para ello se encuestó a 100 matrimonios obteniéndose los siguientes resultados.
¿Qué podemos afirmar a la vista de estos datos?

X	0	1	2	3
O_i	22	42	28	8

Ejemplo 3.15: Se desea saber si el número de altas diarias de un hospital difiere dependiendo del día de la semana (X). La siguiente tabla muestra la frecuencia observada para cada día en un total de 589 altas por semana.
¿Qué podemos afirmar a la vista de estos datos?

X	L	M	X	J	V	S	D	Total
O_i	78	90	94	89	110	84	44	589

Contraste de Bondad de Ajuste: segundo caso

En ocasiones queremos averiguar si los datos se ajustan a un determinado tipo de distribución pero sin precisar los valores de los parámetros que la caracteriza, e.d., *no conocemos* θ donde $\theta = (\theta_1, \cdots, \theta_r)$

$$H_0: \quad X \text{ sigue la distribución } F_\theta$$
$$H_1: \quad X \text{ no sigue la distribución } F_\theta$$

Paso 1.: Hallamos $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_r)$.
Paso 2.: Repetimos exactamente los mismos pasos que para el caso anterior pero ahora:

$$\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \sim \chi^2_{k-1-r}; \qquad R = \left\{ \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} > \chi^2_{k-1-r,\alpha} \right\}$$

Nota: $\sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{O_i^2}{e_i} - n$

Ejemplo 3.16: Para averiguar si el número de descendientes de cierta especie, X, sigue una distribución de Poisson se extrajo una m.a.s. de la población obteniéndose los siguientes resultados. ¿Qué podemos afirmar a la vista de estos datos?

X	0	1	2	3	4	5	6
O_i	25	30	24	14	5	1	1

Contraste de homogeneidad de poblaciones

Sea X es una v.a que queremos estudiar en p poblaciones independientes. Extraemos m.a.s. de cada población, a la vista de las muestras, *¿es razonable admitir que las poblaciones son homogéneas, es decir, que todas ellas siguen la misma distribución?*

H_0: Las p poblaciones siguen la misma distribución
H_1: Las p poblaciones no siguen la misma distribución

Paso 1.: Hacer una partición del espacio muestral común a las p poblaciones en k clases $A_1, \cdots, A_k$.

Paso 2.: Calcular las siguientes frecuencias absolutas:

O_{ij} = **frecuencia observada en** A_i con la muestra j-ésima.

e_{ij} = **frecuencia esperada en** A_i **con la muestra** j**-ésima si** H_0 **es cierta** $= n_j P(A_i)$ (e_{ij} es la esperanza de una $Bin(n_j, P(A_i))$)

Paso 3.: Utilizar el estadístico

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(k-1)(p-1)};$$

$$R = \left\{ \sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} > \chi^2_{(k-1)(p-1),\alpha} \right\}$$

Nota: De nuevo: $\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} = \sum_{j=1}^{p}\sum_{i=1}^{k}\frac{O_{ij}^2}{e_{ij}} - n$

Ejemplo 3.17: Un estudio sobre tabaquismo en las comunidades de Galicia, Madrid y Cataluña proporcionó los siguientes datos

Comunidad	Fumadores	No Fumadores	Total
Galicia	13	87	100
Madrid	17	83	100
Cataluña	18	82	100

¿Pueden considerarse homogéneas las tres poblaciones en cuanto a sus hábitos fumadores?