Tema 5
Contrastes de hipótesis estadísticas

5.1 Contrastes sobre los parámetros de poblaciones normales

Las técnicas que veremos en este tema obedecen a una necesidad ya planteada en el tema anterior: estudiar una característica desconocida, en una población de individuos, a partir de una muestra de la misma mucho menos numerosa que la población completa. La diferencia en este caso es que no nos interesa estimar el valor de la característica, sino extraer conclusiones sobre algún aspecto desconocido de la misma. Por ejemplo: ¿el pH medio del agua de una piscina es superior a 7.6?¿La varianza de las longitudes de las piezas manufacturadas por una fábrica es distinta de 0.1?La variable que modela los ingresos de la población, ¿sigue una distribución normal?

Definición 5.1.1.– *Una hipótesis estadística es una conjetura sobre alguna característica desconocida de la población bajo estudio.*

Las hipótesis estadísticas que plantearemos en este tema serán hipótesis sobre los parámetros de una o dos poblaciones normales y sobre la normalidad de la distribución que sigue la magnitud bajo estudio [1].

Definición 5.1.2.– *Se llama* **hipótesis nula**, *y se denota por H_0, a la hipótesis bajo estudio. La realización de un contraste sobre H_0 supone la contraposición de esta hipótesis frente a otra* **hipótesis alternativa** *que se denota H_1.*

Definición 5.1.3.– *Una hipótesis se dice que es simple cuando asigna un único valor al parámetro bajo estudio, no dejando ningún grado de libertad. En caso contrario, se dice que la hipótesis es compuesta.*

[1] Al hablar de normalidad, nos estamos refiriendo a distribución de probabilidad normal.

Ejemplo. dada $X \sim N(\mu, \sigma^2)$ los siguientes son ejemplos de hipótesis simples y compuestas:

Hipótesis simples	Hipótesis compuestas
$\mu = 10$	$\mu \leq 10$
$\sigma^2 = 9$	$\sigma^2 \neq 9$
$\sigma = 1$	$\sigma > 1$

Definición 5.1.4.– *En un contraste paramétrico una hipótesis compuesta unilateral es de la forma $\theta < \theta_0$ o bien de la forma $\theta > \theta_0$, mientras que una hipótesis bilateral es de la forma $\theta \neq \theta_0$, siendo θ el parámetro sobre el cual se realiza el contraste y θ_0 el valor de prueba.*

Plantear un contraste de hipótesis consiste en el establecimiento de una hipótesis nula y una hipótesis alternativa. Generalmente se toma como H_0 una hipótesis simple y como H_1 una hipótesis compuesta (unilateral o bilateral). Por ejemplo, dada una variable aleatoria X que sigue una distribución $N(\mu, \sigma^2)$, podemos plantear el contraste

$$\begin{cases} H_0 : \mu = 10 \\ H_1 : \mu < 10 \end{cases}$$

La resolución de un contraste, que posteriormente veremos cómo se realiza, da como resultado tomar una de las dos decisiones siguientes:

- **Rechazar H_0:** cuando, a partir de los datos experimentales, se tiene evidencia significativa de que H_1 es cierta.
- **No rechazar H_0:** cuando, a partir de los datos experimentales, no se tiene evidencia significativa de que H_1 es cierta.

Ahora bien, como la decisión se toma a partir de una muestra podemos equivocarnos, siendo posibles cada una de las siguientes situaciones:

	H_0 cierta	H_0 falsa
No rechazar H_0	Correcto	**Error de Tipo II**
Rechazar H_0	**Error de Tipo I**	Correcto

Definición 5.1.5.– *Se denota por α a la probabilidad de cometer un Error de Tipo I, también llamada nivel de significación del contraste, y se denota por β a la probabilidad de cometer un Error de Tipo II; esto es:*

$$P\left[\textit{Rechazar } H_0|_{H_0 \textit{ es cierta}}\right] = \alpha$$

$$P\left[\textit{No rechazar } H_0|_{H_0 \textit{ es falsa}}\right] = \beta$$

¿Se pueden controlar los errores? Lo ideal sería que ambas probabilidades fueran cero o estuvieran muy próximas a cero, pero esto no es posible. Más aún, aunque una probabilidad no es complementaria de la otra y, en general, $\alpha + \beta \neq 1$ cuando una disminuye la otra aumenta, por lo que no podemos minimizar las dos a la vez.

En la práctica lo que se suele hacer es fijar α como un valor pequeño. De este modo, la probabilidad de cometer un error de tipo I será pequeña y estará controlada. Por otra parte sería deseable que, una vez fijado α, el valor de β fuera lo menor posible si bien la consecución de este objetivo excede las pretensiones de este tema.

De este modo cuando la hipótesis nula es cierta la probabilidad de rechazarla, adoptando así una decisión incorrecta, es igual a α que es conocido y sabemos que toma un valor pequeño. Por lo tanto cuando rechazamos H_0 podemos estar 'seguros' de nuestra decisión. Por el contrario si H_0 es falsa la probabilidad de no rechazarla, tomando así una decisión incorrecta, es igual a β que no sabemos cuánto vale y que en la práctica toma un valor lo suficientemente grande para que nuestra decisión venga acompañada de una incertidumbre considerable. Por todo esto, al resolver un contraste de hipótesis, la conclusión debe venir dada en términos de H_1 y nos dirá o bien que *existe evidencia significativa para afirmar que* H_1 *es cierta* o bien que *no existe evidencia significactiva para afirmar que* H_1 *es cierta*, pero no podremos extraer conclusiones sobre la hipótesis nula propiamente dicha.

Si H_1 es compuesta, la probabilidad de cometer un Error de Tipo II, β, depende del parámetro θ, esto es, $\beta = \beta(\theta)$ y representa la probabilidad de cometer un Error de Tipo II para cada valor θ de la hipótesis alternativa. Esto nos lleva a la siguiente definición.

Definición 5.1.6.– *Se define la función potencia del test como* $\rho(\theta) = 1 - \beta(\theta)$. *Representa la probabilidad de rechazar* H_0 *cuando es falsa y conviene que sea lo mayor posible.*

Para resolver un contraste se utiliza un estadístico T unido a una regla de decisión, que dependerá del α fijado. La regla de decisión será de la forma: rechazo $H_0 \Leftrightarrow$ "T verifica una condición". De este modo, el recorrido de T queda dividido en dos partes:

Región de aceptación: valores de T que hacen que no se rechace H_0.

Región crítica o de rechazo: valores de T que hacen que se rechace H_0.

En lo que sigue se muestran las regiones críticas de los posibles contrastes sobre los parámetros de una población normal y sobre los parámetros de dos poblaciones normales independientes.

Contrastes de hipótesis, sobre poblaciones normales, a nivel de significación α

Contrastes sobre una población

Contrastes sobre μ con σ conocida	
Hipótesis	Región crítica
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$	$\overline{X} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$ ó $\overline{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$	$\overline{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$	$\overline{X} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$

Contrastes sobre μ con σ desconocida	
Hipótesis	Región Crítica
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$	$\overline{X} \leq \mu_0 - \frac{S_c}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}}$ ó $\overline{X} \geq \mu_0 + \frac{S_c}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}}$
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$	$\overline{X} \geq \mu_0 + \frac{S_c}{\sqrt{n}} \cdot t_{n-1,1-\alpha}$
$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$	$\overline{X} \leq \mu_0 - \frac{S_c}{\sqrt{n}} \cdot t_{n-1,1-\alpha}$

Contrastes sobre σ^2	
Hipótesis	Región Crítica
$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{cases}$	$S_c^2 \geq \frac{\sigma_0^2 \chi^2_{n-1,1-\frac{\alpha}{2}}}{n-1}$ ó $S_c^2 \leq \frac{\sigma_0^2 \chi^2_{n-1,\frac{\alpha}{2}}}{n-1}$
$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases}$	$S_c^2 \geq \frac{\sigma_0^2 \chi^2_{n-1,1-\alpha}}{n-1}$
$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 < \sigma_0^2 \end{cases}$	$S_c^2 \leq \frac{\sigma_0^2 \chi^2_{n-1,\alpha}}{n-1}$

Contrastes de hipótesis, sobre poblaciones normales, a nivel de significación α

Contrastes sobre dos poblaciones independientes

Contrastes sobre diferencias de medias con varianzas conocidas	
Hipótesis	Región Crítica
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y \neq \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \leq \delta_0 - \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \cdot z_{1-\frac{\alpha}{2}}$ o bien $\overline{X} - \overline{Y} \geq \delta_0 + \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \cdot z_{1-\frac{\alpha}{2}}$
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y > \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \geq \delta_0 + \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \cdot z_{1-\alpha}$
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y < \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \leq \delta_0 - \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \cdot z_{1-\alpha}$

Contrastes sobre diferencias de medias con varianzas iguales y desconocidas	
Hipótesis	Región Crítica*
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y \neq \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \leq \delta_0 - S_p \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \cdot t_{n_x+n_y-2,1-\frac{\alpha}{2}}$ o bien $\overline{X} - \overline{Y} \geq \delta_0 + S_p \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \cdot t_{n_x+n_y-2,1-\frac{\alpha}{2}}$
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y > \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \geq \delta_0 + S_p \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \cdot t_{n_x+n_y-2,1-\alpha}$
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y < \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \leq \delta_0 - S_p \cdot \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \cdot t_{n_x+n_y-2,1-\alpha}$

$^{*}S_p^2 = \frac{(n_x-1)S_{cx}^2 + (n_y-1)S_{cy}^2}{n_x+n_y-2}$

Contrastes de hipótesis, sobre poblaciones normales, a nivel de significación α

Contrastes sobre dos poblaciones independientes

Contrastes sobre diferencias de medias con varianzas distintas y desconocidas	
Hipótesis	Región Crítica*
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y \neq \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \leq \delta_0 - \sqrt{\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}} \cdot t_{g,1-\frac{\alpha}{2}}$ o bien $\overline{X} - \overline{Y} \geq \delta_0 + \sqrt{\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}} \cdot t_{g,1-\frac{\alpha}{2}}$
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y > \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \geq \delta_0 + \sqrt{\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}} \cdot t_{g,1-\alpha}$
$\begin{cases} H_0: & \mu_x - \mu_y = \delta_0 \\ H_1: & \mu_x - \mu_y < \delta_0 \end{cases}$	$\overline{X} - \overline{Y} \leq \delta_0 - \sqrt{\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}} \cdot t_{g,1-\alpha}$

Contrastes sobre varianzas de dos poblaciones	
Hipótesis	Región Crítica
$\begin{cases} H_0: & \sigma_x^2 = \sigma_y^2 \\ H_1: & \sigma_x^2 \neq \sigma_y^2 \end{cases}$	$\frac{S_{cx}^2}{S_{cy}^2} \geq f_{n_x-1,n_y-1,1-\frac{\alpha}{2}}$ o bien $\frac{S_{cx}^2}{S_{cy}^2} \leq \frac{1}{f_{n_y-1,n_x-1,1-\frac{\alpha}{2}}}$
$\begin{cases} H_0: & \sigma_x^2 = \sigma_y^2 \\ H_1: & \sigma_x^2 > \sigma_y^2 \end{cases}$	$\frac{S_{cx}^2}{S_{cy}^2} \geq f_{n_x-1,n_y-1,1-\alpha}$
$\begin{cases} H_0: & \sigma_x^2 = \sigma_y^2 \\ H_1: & \sigma_x^2 < \sigma_y^2 \end{cases}$	$\frac{S_{cx}^2}{S_{cy}^2} \leq \frac{1}{f_{n_y-1,n_x-1,1-\alpha}}$

$$^* g \approx \frac{\left(\frac{S_{cx}^2}{n_x} + \frac{S_{cy}^2}{n_y}\right)^2}{\frac{\left(S_{cx}^2/n_x\right)^2}{n_x+1} + \frac{\left(S_{cy}^2/n_y\right)^2}{n_y+1}} - 2$$

5.2 Estudio de la normalidad de los datos: el test de Shapiro-Wilk

Tanto a la hora de calcular intervalos de confianza, en el tema anterior, como de realizar contrastes de hipótesis, en este tema, hemos partido de la hipótesis de que la muestra procede de una variable aleatoria que sigue una distribución normal. En la práctica, la hipótesis de normalidad debe comprobarse antes de aplicar dichas técnicas porque, en caso contrario, podemos obtener conclusiones erróneas.

Existen distintos test para estudiar la normalidad de un conjunto de datos. En esta sección describiremos el test de Shapiro-Wilk, válido para muestras de tamaño no superior a 50. Partimos de $X_1, \ldots, X_n$ una muestra aleatoria simple de una variable aleatoria X. El objetivo es resolver el contraste

$$\begin{cases} \mathrm{H}_0: & X \sim \text{Normal} \\ \mathrm{H}_1: & X \nsim \text{Normal} \end{cases}$$

La prueba de Shapiro-Wilk utiliza el estadístico W, definido como

$$W = \frac{b^2}{(n-1)S_c^2} \qquad \text{con} \qquad b = \sum_{i=1}^{[n/2]} (X_{(n-i+1)} - X_{(i)}) a_{n-i+1}$$

donde $X_{(i)}$ denota la i-ésima variable de la muestra ordenada en orden no decreciente de valores, esto es, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, y $[\cdot]$ denota parte entera. El estadístico W verifica que $0 \leq W \leq 1$. Obsérvese que no es necesario conocer el valor de los parámetros de la distribución.

Los a_{n-i+1} son unos coeficientes denominados de Shapiro Wilk, que encontraremos en la tabla correspondiente. Se rechazará la hipótesis nula si $W \leq W_{n,\alpha}$, donde $W_{n,\alpha}$ está tabulado en la tabla de valores críticos del test de Shapiro-Wilk.

Si el tamaño muestral es impar el valor muestral que ocupa la posición $[n/2]+1$, en la muestra ordenada, no influye en el valor del sumatorio utilizado para calcular W. A pesar de esto, no podemos eliminar dicho valor de la muestra y trabajar con una muestra de tamaño par pues, en ese caso, los coeficientes a_{n-i+1} y por tanto el valor del estadístico W podrían cambiar significativamente.