Tema 9. Contrastes paramétricos en dos poblaciones

Resumen del tema

9.1. Comparación de dos varianzas

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales. μ_1, μ_2 desconocidas.		
estadístico	$F = \dfrac{S_1^2}{S_2^2}$ con $S_1^2 \geq S_2^2$		
contraste	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$
región crítica	$F < \dfrac{1}{F_{n_2-1,n_1-1,1-\alpha/2}}$ $F > F_{n_1-1,n_2-1,1-\alpha/2}$	$F < \dfrac{1}{F_{n_2-1,n_1-1,1-\alpha}}$	$F > F_{n_1-1,n_2-1,1-\alpha}$

9.2. Comparación de dos medias

9.2.1. Muestras independientes y varianzas poblacionales conocidas

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales (o cualesquiera si $n_1, n_2 \geq 30$). σ_1, σ_2 conocidas.		
estadístico	$Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$Z < -Z_{1-\alpha/2}$ $Z > Z_{1-\alpha/2}$	$Z < -Z_{1-\alpha}$	$Z > Z_{1-\alpha}$

9.2.2. Muestras independientes y varianzas poblacionales desconocidas e iguales

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales (o cualesquiera si $n_1, n_2 \geq 30$). σ_1, σ_2 desconocidas pero iguales.		
estadístico	$T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$T < -t_{n_1+n_2-2\,,\,1-\alpha/2}$ $T > t_{n_1+n_2-2\,,\,1-\alpha/2}$	$T < -t_{n_1+n_2-2\,,\,1-\alpha}$	$T > t_{n_1+n_2-2\,,\,1-\alpha}$

9.2.3. Muestras independientes y varianzas poblacionales desconocidas y distintas

condiciones	Muestras aleatorias simples independientes de tamaños n_1 y n_2. Poblaciones Normales (o cualesquiera si $n_1, n_2 \geq 30$). σ_1, σ_2 desconocidas y distintas.		
estadístico	$T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$		
grados de libertad	g=n$^{\text{o}}$ natural más próximo a $\dfrac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2-1}}$		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$T < -t_{g\,,\,1-\alpha/2}$ $T > t_{g\,,\,1-\alpha/2}$	$T < -t_{g\,,\,1-\alpha}$	$T > t_{g\,,\,1-\alpha}$

9.2.4. Muestras apareadas

condiciones	Muestras aleatorias simples apareadas de tamaño n. La variable aleatoria $D = X_1 - X_2$ es Normal (o cualquiera si $n \geq 30$).		
estadístico	$T = \dfrac{\overline{D}}{\frac{S_D}{\sqrt{n}}}$ donde $\overline{D}$ y S_D son la media y la cuasidesviación típica de D		
contraste	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$
región crítica	$T < -t_{n-1\,,\,1-\alpha/2}$ $T > t_{n-1\,,\,1-\alpha/2}$	$T < -t_{n-1\,,\,1-\alpha}$	$T > t_{n-1\,,\,1-\alpha}$

9.3. Comparación de dos proporciones

condiciones	Variable que sólo toma los valores 0 (fracaso) y 1 (éxito). p_i=proporción de éxitos en la población i, para $i = 1, 2$. Muestras aleatorias simples independientes de tamaños n_1 y n_2. X_i=n$^{\text{o}}$ de éxitos en la muestra i, $\hat{p}_i = X_i/n_i$, para $i = 1, 2$. $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$, $m = \text{máximo}\{n_1, n_2\}$. $n_1 \geq 30$, $n_2 \geq 30$, $0'1 \leq \hat{p}_1 \leq 0'9$, $0'1 \leq \hat{p}_2 \leq 0'9$.		
estadístico	$Z = \dfrac{(\hat{p}_1 - \hat{p}_2) \pm \frac{1}{2m}}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$		
contraste	$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	$H_0 : p_1 \geq p_2$ $H_1 : p_1 < p_2$	$H_0 : p_1 \leq p_2$ $H_1 : p_1 > p_2$
región crítica	$Z < -Z_{1-\alpha/2}$ $Z > Z_{1-\alpha/2}$	$Z < -Z_{1-\alpha}$	$Z > Z_{1-\alpha}$

Ejemplos que se van a resolver en clase

Ejemplo 9.1. En la tabla siguiente aparece el precio, en euros, de una muestra aleatoria de 15 libros que se prestan pocas veces (X_1) y el precio, en euros, de una muestra aleatoria de 15 libros que se prestan muchas veces (X_2).

x_{1i}	x_{2i}
75	110
32	30
30	45
34	69
42	46
57	53
51	97
36	43
82	42
45	37
58	48
66	45
40	105
35	61
51	57

$\sum_{i=1}^{15} x_{1i} = 734$	$\sum_{i=1}^{15} x_{1i}^2 = 39510$	$\sum_{i=1}^{15} x_{2i} = 888$	$\sum_{i=1}^{15} x_{2i}^2 = 61426$

a) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del precio de los libros que se prestan poco es igual a la varianza poblacional del precio de los libros que se prestan mucho?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del precio de los libros que se prestan poco es igual a la media poblacional del precio de los libros que se prestan mucho?

Ejemplo 9.2. En la siguiente tabla aparece el número de palabras por resumen de una muestra aleatoria de 30 artículos científicos escritos en francés (X_1) y el número de palabras por resumen de una muestra aleatoria de 30 artículos científicos escritos en inglés (X_2).

x_{1i}	70	65	68	74	79	67	75	80	62	69
	61	57	71	74	82	91	70	64	72	67
	74	70	81	85	70	74	75	71	69	54
x_{2i}	80	47	59	67	89	57	72	78	74	72
	104	118	89	87	79	78	101	120	107	95
	85	87	90	98	89	75	90	101	85	94

$\sum_{i=1}^{30} x_{1i} = 2141$	$\sum_{i=1}^{30} x_{1i}^2 = 154627$	$\sum_{i=1}^{30} x_{2i} = 2567$	$\sum_{i=1}^{30} x_{2i}^2 = 227713$

a) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional de la longitud de los resúmenes de artículos escritos en francés es igual a la varianza poblacional de la longitud de los resúmenes de artículos escritos en inglés?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional de la longitud de los resúmenes de artículos escritos en francés es igual a la media poblacional de la longitud de los resúmenes de artículos escritos en inglés?

Ejemplo 9.3. Se está estudiando el número de palabras por resumen de los artículos científicos de un determinado volumen de Economics Abstracts. La varianza poblacional es conocida e igual a $615'04$. Se extrae una muestra aleatoria simple de 30 resúmenes escritos en alemán y se observa que la media es $67'47$, y otra muestra aleatoria simple de 32 resúmenes escritos en inglés, obteniéndose una media de $72'5$. ¿Existe diferencia significativa entre el número medio de palabras por resumen en alemán y el número medio de palabras por resumen en inglés?

Ejemplo 9.4. Dos expertos califican una muestra aleatoria de 30 libros según su calidad (1=muy mala, 2=mala, 3=regular, 4=buena, 5=muy buena). En la tabla siguiente aparece la opinión del primer experto (X_1) y la opinión del segundo experto (X_2).

x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$	x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$
2	1	1	4	4	0
5	4	1	4	3	1
4	5	-1	5	4	1
2	3	-1	5	3	2
3	3	0	1	2	-1
1	5	-4	2	5	-3
3	3	0	2	3	-1
1	3	-2	3	2	1
4	2	2	4	1	3
2	5	-3	4	2	2
3	2	1	1	3	-2
4	3	1	2	4	-2
3	3	0	1	2	-1
1	3	-2	5	5	0
2	5	-3	5	2	3

$\displaystyle\sum_{i=1}^{30} d_i = -7$	$\displaystyle\sum_{i=1}^{30} d_i^2 = 101$

¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional de los resultados de la opinión del primer experto es igual a la media poblacional de los resultados de la opinión del segundo experto?

Ejemplo 9.5. Con objeto de comparar dos pequeñas empresas A y B de encuadernación de libros, se extrajo una muestra aleatoria de 250 libros encuadernados en A y otra muestra aleatoria de 200 libros encuadernados en B, y se encontró que 50 de los libros encuadernados en A, y 32 de los encuadernados en B tenían algún defecto en su encuadernación. ¿Son igualmente buenas las dos empresas de encuadernación?

Problemas propuestos

Problema 9.1. El precio de los libros de una biblioteca es una variable aleatoria Normal de media $63'3$ euros y desviación típica $19'4$ euros. Se sospecha que el precio medio de los libros de ciencias físicas es mayor que el precio medio de los libros de ciencias sociales. Para obtener alguna evidencia sobre la sospecha, se selecciona una muestra aleatoria simple de 20 libros de ciencias físicas y otra de 30 libros de ciencias sociales, obteniéndose una media de $57'5$ euros para los primeros, y $52'6$ euros para los segundos. ¿Podemos afirmar, con un nivel de significación de $0'05$, que es cierta nuestra sospecha?

Problema 9.2. Se nos ha señalado la posibilidad de que se paguen sueldos distintos a documentalistas según el sexo. Presumiblemente, a los hombres se les ha pagado más que a las mujeres. Un estudio de los sueldos anuales durante los cinco años anteriores al actual arroja los siguientes resultados:

	hombres	mujeres
media muestral	21.980	20.470
cuasidesviación típica muestral	1.810	2.290
tamaño muestral	25	50

A la vista de estos datos, y utilizando un nivel de significación de $0'01$, ¿podemos afirmar que el sueldo de los hombres documentalistas es mayor que el de las mujeres documentalistas?

Problema 9.3. Elegimos al azar 30 matrimonios y observamos el número de veces que los hombres han visitado alguna biblioteca en los tres últimos meses (X_1) y el número de veces que las mujeres han visitado alguna biblioteca en los tres últimos meses (X_2). Los resultados se muestran en la siguiente tabla.

x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$	x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$	x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$
12	8	4	8	10	-2	25	14	11
30	11	19	14	15	-1	12	16	-4
10	12	-2	20	12	8	8	10	-2
20	16	4	13	19	-6	23	20	3
15	10	5	11	6	5	14	17	-3
14	9	5	7	7	0	8	10	-2
11	12	-1	6	7	-1	12	23	-11
9	10	-1	8	6	2	27	10	17
7	7	0	15	20	-5	32	27	5
5	4	1	42	35	7	14	18	-4

$\sum_{i=1}^{30} d_i = 51$	$\sum_{i=1}^{30} d_i^2 = 1273$

¿Podemos afirmar que hay diferencia significativa entre los hombres y las mujeres de los matrimonios en cuanto al número de veces que van a la biblioteca?

Problema 9.4. En la siguiente tabla aparece el número de usuarios diarios de la biblioteca A (variable X_1) y el número de usuarios diarios de la biblioteca B (variable X_2) en 10 días elegidos al azar.

x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$
51	45	6
72	58	14
35	32	3
70	56	14
75	68	7
98	76	22
100	88	12
80	69	11
72	57	15
90	75	15

$\sum_{i=1}^{10} d_i = 119$	$\sum_{i=1}^{10} d_i^2 = 1685$

a) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la muestra de las diferencias d_i es aleatoria?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la variable diferencia $D = X_1 - X_2$ es Normal?

c) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del número de usuarios diarios de la biblioteca A es igual a la media poblacional del número de usuarios diarios de la biblioteca B?

Problema 9.5. Se quiere saber si la proporción de libros escritos en español es la misma en dos bibliotecas universitarias (la de la facultad de matemáticas y la de la facultad de filosofía). Se toma una muestra aleatoria simple de 100 libros de la biblioteca de la facultad de matemáticas y se encuentra que 35 de ellos están escritos en español y el resto en otros idiomas. Se extrae otra muestra aleatoria simple de 150 libros de la biblioteca de la facultad de filosofía y se observa que 60 están escritos en español. ¿Qué conclusión se puede extraer?

Soluciones de los problemas propuestos

Solución del problema 9.1. Sea X_1=*precio de los libros de ciencias físicas* y X_2=*precio de los libros de ciencias sociales*. Hacemos el contraste de comparación de dos medias en el que la hipótesis nula es $H_0 : \mu_1 \leq \mu_2$. Las muestras son independientes y las varianzas poblacionales se consideran conocidas: $\sigma_1^2 = \sigma_2^2 = 19'4^2$. El valor del estadístico de contraste es $Z = 0'8750$. Como el nivel de significación es $\alpha = 0'05$, entonces la región crítica es $Z > 1'6449$. En consecuencia, aceptamos H_0 y, por tanto, no podemos aceptar que el precio medio de los libros de ciencias físicas sea mayor que el precio medio de los libros de ciencias sociales. Finalmente, la respuesta a la pregunta es **NO**.

Solución del problema 9.2.

1) En primer lugar tenemos que hacer un contraste de comparación de dos varianzas poblacionales ya que éstas son desconocidas, y no sabemos si son iguales o distintas.

Debe ser X_1=*sueldo anual de las mujeres documentalistas* y X_2=*sueldo anual de los hombres documentalistas*, pues la cuasidesviación típica muestral en las mujeres es mayor que en los hombres.

Hacemos el contraste de comparación de dos varianzas en el que la hipótesis nula es $H_0 : \sigma_1^2 = \sigma_2^2$. Las muestras son independientes y se supone que las dos variables aleatorias son normales. El valor del estadístico de contraste es $F = 1'6007$. Como el nivel de significación es $\alpha = 0'01$, entonces la región crítica es $F < 0'4249$ ó $F > 2'6522$. En consecuencia, aceptamos H_0 y, por tanto, las varianzas poblacionales son desconocidas pero iguales.

2) En segundo lugar hacemos un contraste de comparación de dos medias en el que la hipótesis nula es $H_0 : \mu_1 \geq \mu_2$. Las muestras son independientes y las desviaciones típicas poblacionales son desconocidas pero iguales. El valor del estadístico de contraste es $T = -2'8751$. Como el nivel de significación es $\alpha = 0'01$, entonces la región crítica es $T < -2'3789$. En consecuencia, rechazamos H_0 y, por tanto, aceptamos que el sueldo medio de los hombres documentalistas es mayor que el sueldo medio de las mujeres documentalistas. Finalmente, la respuesta a la pregunta es **SÍ**.

Solución del problema 9.3. Hacemos el contraste de comparación de dos medias en el que la hipótesis nula es $H_0 : \mu_1 = \mu_2$. Las muestras son apareadas. El valor del estadístico de contraste es $T = 1'455832$. Si el nivel de significación es $\alpha = 0'05$, entonces la región crítica es $T < -2'0452$ ó $T > 2'0452$. En consecuencia, tenemos que aceptar H_0. Por tanto, no hay diferencia significativa entre los hombres y las mujeres de los matrimonios en cuanto al número de veces que van a la biblioteca. Finalmente, la respuesta a la pregunta es **NO**.

Solución del problema 9.4. Sea D la variable aleatoria diferencia entre X_1 y X_2; es decir $D = X_1 - X_2$.

a) Hacemos el contraste de las rachas sobre aleatoriedad de la muestra en el que la hipótesis nula es H_0 :*La muestra de datos de la variable D es aleatoria*. El valor del estadístico de contraste es $R = 8$. Como el nivel de significación es $\alpha = 0'05$, entonces la región de aceptación es el intervalo $(2, 10)$. Por tanto, aceptamos H_0. Finalmente, la respuesta a la pregunta es **SÍ**.

b) Hacemos el contraste de D'Agostino sobre normalidad en el que la hipótesis nula es H_0 :*La variable aleatoria D es Normal*. El valor del estadístico de contraste es $D_{exp} = 0'274802$. Como el nivel de significación es $\alpha = 0'05$, entonces la región de aceptación es el intervalo $(0'2513, 0'2849)$. Por tanto, aceptamos H_0. Finalmente, la respuesta a la pregunta es **SÍ**.

c) Hacemos el contraste de comparación de dos medias en el que la hipótesis nula es $H_0 : \mu_1 = \mu_2$. Las muestras son apareadas. El valor del estadístico de contraste es $T = 6'884506$. Como el nivel de significación es $\alpha = 0'05$, entonces la región crítica es $T < -2'2622$ ó $T > 2'2622$. En consecuencia, tenemos que rechazar H_0. Por tanto, la media poblacional del número de usuarios diarios de la biblioteca A no es igual a la media poblacional del número de usuarios diarios de la biblioteca B. Finalmente, la respuesta a la pregunta es **NO**.

Solución del problema 9.5. Sea p_1=*proporción de libros escritos en español en la biblioteca de la facultad de matemáticas* y p_2=*proporción de libros escritos en español en la biblioteca de la facultad de filosofía*. Hacemos el contraste de comparación de dos proporciones, con muestras independientes, en el que la hipótesis nula es $H_0 : p_1 = p_2$. El valor del estadístico de contraste es $Z = -0'851112$. Si tomamos un nivel de significación de $\alpha = 0'05$, entonces la región crítica es $Z < -1'96$ ó $Z > 1'96$. En consecuencia, tenemos que aceptar H_0. Por tanto, la proporción de libros escritos en español es la misma en las dos bibliotecas.