OPEN COURSEWARE
UNIVERSIDAD DE MURCIA

Tema 7. Contrastes no paramétricos en una población

Resumen del tema

7.1. Introducción a la Estadística Inferencial. Estimación de parámetros

Como ya sabemos, la Estadística estudia los métodos científicos para recoger, organizar, resumir y analizar datos de una o varias muestras, extrayendo, a través del cálculo de probabilidades, conclusiones válidas que nos permitan tomar decisiones sobre la población. En el bloque 1 de esta asignatura hemos estudiado la rama de la Estadística que se ocupa de describir y analizar los datos de las muestras, sin sacar conclusiones sobre un conjunto mayor de datos; es decir, hemos estudiado *Estadística Descriptiva*. En el bloque 2 se han resuelto problemas relativos al *Cálculo de Probabilidades* de ciertos sucesos relacionados con variables aleatorias que seguían determinadas distribuciones de parámetros conocidos. Sin embargo, siendo los parámetros algo característico de toda población, es usual que sean desconocidos. En el bloque 3, que ahora comienza, vamos a estudiar la rama de la Estadística que trata de sacar conclusiones o inferencias sobre un grupo grande de datos (población) a partir de un subgrupo de datos (muestra), incluyendo el problema de la determinación aproximada de los parámetros de la población. Esta rama se llama **Estadística Inferencial**.

La utilización de un método adecuado de muestreo garantiza que la muestra obtenida es representativa de la población. Esto significa que la información proporcionada por la muestra es un reflejo de la información contenida en la población. Podemos, por tanto, utilizar la información muestral para formarnos una idea sobre las propiedades de la población. Es decir, podemos servirnos de las muestras para hacer *inferencias* sobre la población.

Estas inferencias pueden adoptar diferentes formas pero las más habituales son dos: la *estimación de parámetros* y el *contraste de hipótesis*. Cuando la información deseada de la población es el valor de alguno de sus parámetros, la técnica a utilizar es la **estimación de parámetros**. Los **contrastes de hipótesis** permiten comprobar si ciertas hipótesis que se enuncian acerca de la población son correctas o no.

La estimación de parámetros puede ser:

- **Estimación puntual**: consiste en asignar un valor muestral concreto al parámetro poblacional que se desea estimar.
- **Estimación por intervalo de confianza**: consiste en atribuir al parámetro que se desea estimar, no un valor concreto, sino un rango de valores entre los que se espera que pueda encontrarse el verdadero valor del parámetro con una probabilidad alta y conocida.

7.2. Contrastes de hipótesis

– *Hipótesis estadística*: afirmación sobre la forma de una o más distribuciones, o sobre el valor de uno o más parámetros de esas distribuciones.

- *Hipótesis nula*: hipótesis estadística que se somete a contraste. Se denota por H_0.
- *Hipótesis alternativa*: es la negación de la hipótesis nula H_0, e incluye todo lo que H_0 excluye. Se denota por H_1.
- *Contraste de hipótesis*: procedimiento que nos capacita para determinar si las muestras observadas difieren significativamente de los resultados esperados, y por tanto nos ayuda a decidir si aceptamos o rechazamos la hipótesis nula.
    - *Contraste paramétrico*: la hipótesis nula es una afirmación sobre el valor de uno o más parámetros de la variable aleatoria observada en la población.
    - *Contraste no paramétrico*: la hipótesis nula no es una afirmación sobre el valor de uno o más parámetros de la variable aleatoria observada en la población.
- *Estadístico de contraste*: estadístico que se observa al realizar un contraste de hipótesis, y que nos sirve para aceptar o rechazar la hipótesis nula por poseer una distribución muestral conocida.
- *Región crítica*: zona de la distribución muestral del estadístico de contraste que corresponde a los valores que permiten rechazar la hipótesis nula, y por tanto aceptar la hipótesis alternativa.
- *Región de aceptación*: zona de la distribución muestral del estadístico de contraste que corresponde a los valores que permiten aceptar la hipótesis nula.
- *Contraste unilateral o de una cola*: la región crítica se encuentra en una sola zona de la distribución muestral del estadístico de contraste.
- *Contraste bilateral o de dos colas*: la región crítica se encuentra repartida entre dos zonas de la distribución muestral del estadístico de contraste.
- *Error de tipo I*: error que se comete cuando se decide rechazar una hipótesis nula que en realidad es verdadera.
- *Nivel de significación*: probabilidad de cometer un error de tipo I al contrastar una hipótesis. Se denota por α.
- *Error de tipo II*: error que se comete cuando se decide aceptar una hipótesis nula que en realidad es falsa. La probabilidad de cometer dicho error se denota por β.
- *Potencia de un contraste*: probabilidad de rechazar la hipótesis nula cuando es falsa. Por tanto, la potencia es igual a $1 - \beta$.
- *p-valor (o nivel crítico)*: es el nivel de significación más pequeño al que una hipótesis nula puede ser rechazada con el estadístico de contraste obtenido. Se rechaza H_0 si el p-valor es claramente menor que α; se acepta H_0 si el p-valor es claramente mayor que α; y se repite el contraste con una muestra diferente si el p-valor tiene un resultado próximo a α.

7.3. Contraste sobre aleatoriedad de la muestra

Contraste de las Rachas sobre aleatoriedad de la muestra	
contraste	H_0: la muestra es aleatoria H_1: la muestra no es aleatoria
condiciones	Los datos son sólo de dos tipos o pueden reducirse a dos tipos. N_1=número de datos de un tipo $\leq$ N_2=número de datos del otro tipo.
estadísticos	**(a)** Si $N_1 \leq N_2 \leq 20$ se calcula R =número de rachas (secuencias de datos del mismo tipo). **(b)** Si $N_1 > 20$ ó $N_2 > 20$ se calcula $Z = \dfrac{(R - E(R)) \pm 0'5}{\sqrt{V(R)}}$, donde $E(R) = \dfrac{2N_1N_2}{N_1 + N_2} + 1,$ $V(R) = \dfrac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)}.$
región crítica	**(a)** Si $N_1 \leq N_2 \leq 20$, rechazamos H_0 si el valor de R está fuera del intervalo de la tabla de los puntos críticos del test de las rachas. **(b)** Si $N_1 > 20$ ó $N_2 > 20$, rechazamos H_0 si $Z < -Z_{1-\alpha/2}$ ó $Z > Z_{1-\alpha/2}$.

7.4. Contraste sobre normalidad

Contraste de D'Agostino sobre Normalidad	
contraste	H_0: la variable aleatoria X observada en la población es Normal H_1: la variable aleatoria X observada en la población no es Normal
condiciones	Se extrae una muestra aleatoria simple de tamaño n. Se ordena la muestra de menor a mayor: $X_1 \leq X_2 \leq \cdots \leq X_n$.
estadístico	$D_{exp} = \dfrac{\sum_{i=1}^{n} i\,X_i - \frac{n+1}{2}\sum_{i=1}^{n} X_i}{n\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}}$, donde $\sum_{i=1}^{n} i\,X_i$ significa $1X_1 + 2X_2 + 3X_3 + \cdots + nX_n$.
región crítica	Rechazamos H_0 si el valor de D_{exp} está fuera del intervalo de la tabla de los puntos críticos del test de D'Agostino.

7.5. Contraste chi-cuadrado sobre independencia de dos variables aleatorias

Contraste χ^2 de Pearson sobre independencia de dos variables	
contraste	H_0: las variables X e Y son independientes H_1: las variables X e Y no son independientes
estadístico	$\chi^2_{exp} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$, donde r = número de clases de la variable X, k = número de clases de la variable Y, f_{ij} = frecuencia absoluta conjunta de la clase $A_i \times B_j$, f_{i*} = frecuencia absoluta marginal de la clase A_i de X, f_{*j} = frecuencia absoluta marginal de la clase B_j de Y, $e_{ij} = \frac{f_{i*} \cdot f_{*j}}{n}$ = frecuencia absoluta esperada bajo H_0.
condiciones	$e_{ij} \geq 1$ para todas las clases. $e_{ij} \geq 5$, salvo para un 20 % de las clases como máximo.
región crítica	$\chi^2_{exp} \geq \chi^2_{(r-1)(k-1)\,,\,1-\alpha}$

Ejemplos que se van a resolver en clase

Ejemplo 7.1. En la tabla siguiente aparecen los datos de 10 bibliotecas, en las cuales se ha observado las siguientes variables: número total de títulos catalogados en un año (X), número de horas totales al año que emplea la biblioteca en catalogar sus títulos (Y) y costo, en euros, de una hora de catalogación (Z).

x_i	y_i	z_i
1550	220	15'75
1640	230	14'50
1000	140	16'40
950	135	16'70
750	110	17'10
1700	255	12'50
1650	228	14'80
1860	270	15'25
1900	280	18'50
900	130	17'30

$\sum_{i=1}^{10} z_i = 158'8$	$\sum_{i=1}^{10} z_i^2 = 2547'965$

a) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra de datos de la variable Z es aleatoria?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'02$, que la variable aleatoria Z es Normal?

Ejemplo 7.2. Se ha estudiado el uso de la biblioteca pública por parte de los profesores universitarios, encontrándose que 42 de 113 psicólogos, 17 de 68 biólogos, 33 de 203 ingenieros y 20 de 78 profesores de inglés son usuarios de la biblioteca pública (y el resto no). ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que existe relación entre la especialidad de los profesores y el uso de la biblioteca pública?

Problemas propuestos

Problema 7.1. Los siguientes datos corresponden a las edades de una muestra de 10 personas que visitan una biblioteca.

19	24	83	30	17	23	33	19	68	56

a) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra es aleatoria?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable aleatoria *edad de las personas que visitan la biblioteca* es Normal?

Problema 7.2. El rector de una universidad española desea saber la opinión del profesorado en relación con un proyecto por el cual todos los libros comprados por los departamentos se llevarían a una biblioteca general universitaria ubicada en un edificio independiente de las facultades. Para ello, selecciona una muestra aleatoria de 370 profesores de distintos rangos académicos (A.E.U.= Ayudante de Escuela Universitaria, A.F.= Ayudante de Facultad, T.E.U.=Titular de Escuela Universitaria, T.U.= Titular de Universidad, C.U.= Catedrático de Universidad). Los resultados se reflejan en la siguiente tabla:

	A.E.U.	A.F.	T.E.U.	T.U.	C.U.
en contra	30	55	95	14	12
indiferente	15	20	17	8	10
a favor	10	25	38	8	13

¿Se puede aceptar que existe relación entre el rango académico y la opinión de los profesores respecto del proyecto mencionado?

Problema 7.3. Un profesor de estadística de un Grado en Información y Documentación quiere estudiar la mejor forma de obtener un buen resultado en la asignatura y para ello solicita la colaboración de los alumnos durante varios cursos académicos planteándoles el siguiente esquema: al final del primer parcial califica a todos los alumnos según los resultados del examen en A (sobresaliente y notable), B (aprobado) y C (suspenso); luego les pide que contesten cuál ha sido su método de trabajo ante la signatura (I= sólo estudia teoría, II= sólo estudia problemas, III= estudia teoría y problemas). Conocidos los resultados, el profesor construye la siguiente tabla:

		Método de trabajo		
		I	II	III
Calificación	A	15	12	65
	B	58	70	85
	C	40	102	53

¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la calificación es independiente del método de trabajo empleado?

Problema 7.4. En una determinada facultad se considera una muestra de 807 alumnos y se realiza una encuesta para saber cuántas horas diarias dedica cada alumno al estudio en la biblioteca, obteniéndose la siguiente tabla de resultados:

		Curso de la licenciatura				
		1º	2º	3º	4º	5º
Nº de horas	menos de 1 hora	18	20	32	77	96
	entre 1 y 3 horas	22	35	90	83	50
	más de 3 horas	60	70	80	60	14

¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que existe relación entre el curso al que pertenece el alumno y el tiempo que dedica al estudio en la biblioteca?

Soluciones de los problemas propuestos

Solución del problema 7.1. Sea X=*Edad de las personas que visitan la biblioteca.*

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 X es aleatoria.* El valor del estadístico de contraste es $R = 6$. 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 X es Normal.* El valor del estadístico de contraste es $D_{exp} = 0'261150$. 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**.

Solución del problema 7.2. Sean las dos variables aleatorias: X=*Opinión de los profesores respecto del proyecto (en contra, indiferente, a favor)* e Y=*Rango académico de los profesores universitarios.* Se hace el contraste χ^2 de Pearson sobre independencia de dos variables, en el que la hipótesis nula es H_0 :*Las variables X e Y son independientes.* El valor del estadístico de contraste es $\chi^2_{exp} = 17'295681$.

Si tomáramos un nivel de significación de $\alpha = 0'05$, entonces la región crítica sería $\chi^2_{exp} \geq 15'5073$. Con este nivel de significación tendríamos que rechazar H_0 y, por tanto, aceptaríamos que existe relación entre el rango académico y la opinión de los profesores respecto del proyecto. Es decir, con $\alpha = 0'05$, la respuesta a la pregunta es **SÍ**.

Sin embargo, si tomáramos un nivel de significación de $\alpha = 0'01$, entonces la región crítica sería $\chi^2_{exp} \geq 20'0902$. Con este nivel de significación tendríamos que aceptar H_0 y, por tanto, aceptaríamos que no existe relación entre el rango académico y la opinión de los profesores respecto del proyecto. Es decir, con $\alpha = 0'01$, la respuesta a la pregunta es **NO**.

Solución del problema 7.3. Sean las dos variables aleatorias: X=*Calificación* e Y=*Método de trabajo empleado.* Se hace el contraste χ^2 de Pearson sobre independencia de dos variables, en el que la hipótesis nula es H_0 :*Las variables X e Y son independientes.* El valor del estadístico de contraste es $\chi^2_{exp} = 60'900070$. Como el nivel de significación es $\alpha = 0'05$, entonces la región crítica es $\chi^2_{exp} \geq 9'48773$. Por tanto, se rechaza H_0 y, por tanto, se concluye que la calificación no es independiente del método de trabajo empleado. Finalmente, la respuesta a la pregunta es **NO**.

Solución del problema 7.4. Sean las dos variables aleatorias: X=*Tiempo que dedica cada alumno al estudio en la biblioteca* e Y=*Curso al que pertenece cada alumno.* Se hace el contraste χ^2 de Pearson sobre independencia de dos variables, en el que la hipótesis nula es H_0 :*Las variables X e Y son independientes.* El valor del estadístico de contraste es $\chi^2_{exp} = 158'754042$. Como el nivel de significación es $\alpha = 0'05$, entonces la región crítica es $\chi^2_{exp} \geq 15'5073$. Por tanto, se rechaza H_0 y, por tanto, se concluye que existe relación entre el curso al que pertenece el alumno y el tiempo que dedica al estudio en la biblioteca. Finalmente, la respuesta a la pregunta es **SÍ**.