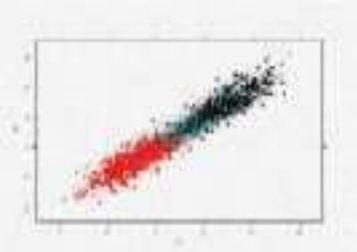
ESTADÍSTICA (SISTEMAS)

Profesores: Hilario Navarro. Jorge Martín



DEPARTAMENTO DE ESTADÍSTICA,
INVESTIGACIÓN OPERATIVA Y CÁLCULO
NUMÉRICO

Cuarta unidad didáctica.
Soluciones a los problemas propuestos de Otros Métodos Estadísticos

Curso 2004-2005

Problema 1. Con el fin de seleccionar el sistema más rápido de almacenamiento y recuperación de datos para un determinado tipo de procesos, se realizó un experimento consistente en hacer 4 pruebas con cada uno de los tres sistemas considerados: *CD, Disco* y *Cinta*. Los tiempos —en minutos— requeridos en cada ocasión se reflejan en la siguiente tabla

	CD	*Disco*	*Cinta*
	8.7	7.0	7.2
	9.3	6.4	9.1
	7.9	9.8	7.5
	8.0	8.2	7.7
Suma	33.9	31.4	31.5
Media	8.475	7.850	7.875
Varianza	0.429	2.250	0.709

Utilizando un nivel de significación $\alpha = 0.01$, contraste la hipótesis de igualdad de los tiempos medios.

Solución

Suponiendo que la respuesta —tiempo de almacenacimiento y recuperación— se distribuye según una ley *normal*, con la misma dispersión en los tres sistemas considerados y que las tres muestras aleatorias son independientes, podemos aplicar un contraste basado en el **análisis de la varianza**. Los resultados de dicho análisis se resumen en la siguiente tabla:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico
Entre grupos	1.08	2	0.54	0.48
Dentro de los grupos	10.087	9	1.12	
Total	11.167	11		

donde:

- $\sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2 = 4[(8.475-8.0)^2 + (7.850-8.0)^2 + (7.875-8.0)^2] = 1.08$
- La suma de cuadrados " dentro de los grupos " se puede obtener como diferencia entre la variabilidad total —dato auxiliar— y la suma anterior:

$$11.167 - 1.08 = 10.087$$

El resto de las entradas de la tabla se obtienen aplicando las definiciones correspondientes.

Dado que $F_{2,9;0.01} = 8.0215$, notablemente superior al valor del estadístico que aparece en la tabla del análisis de la varianza (véase la figura 1), la conclusión debe ser la aceptación de la hipótesis de igualdad de los tiempos medios.

□

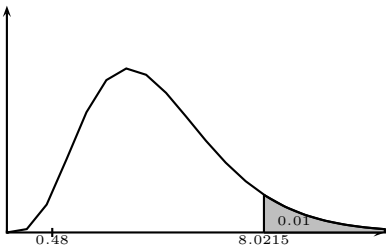


Figura 1: Punto crítico para una $F_{2,9}$ ($\alpha = 0.01$)

Problema 2. El número de trabajadores que diariamente integran una cadena de montaje varía a causa del nivel de absentismo. La tabla siguiente contiene los datos registrados en una muestra aleatoria de la producción diaria, siendo X el número de trabajadores ausentes e Y el número de productos defectuosos generados por dicha cadena.

X	1	3	5	0	2
Y	10	16	20	9	12

Si la recta de regresión estimada es

$$y = 8.26 + 2.34x \ ,$$

calcule un intervalo de confianza (99 %) para la predicción de la cantidad de productos defectuosos que se obtendrán cuando el número de operarios ausentes sea 4.

Solución

La teoría establece que, bajo un modelo de regresión lineal, el intervalo solicitado es de la forma

$$I = \left[\hat{y}_0 \pm t_{n-2;\alpha/2} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})}{n v_x}} \right] .$$

En este caso, el valor de cada unos de los elementos integrantes de la expresión anterior es:

- $\hat{y}_0 = 8.26 + 2.34 \times 4 = 17.62$
- $t_{5-2;0.005} = 5.8409$
- $S_R = \sqrt{\frac{1}{5-2} \sum_{i=1}^{5} (y_i - \hat{y}_i)^2} = 0.878$ donde $\sum_{i=1}^{5} (y_i - \hat{y}_i)^2 =$

$$= [(10-10.6)^2 + (16-15.28)^2 + (20-19.96)^2 + (9-8.26)^2 + (12-12.94)^2] = 2.3112$$

- $(x_0 - \bar{x})^2 = (4 - 2.20)^2 = 3.24$
- $v_x = \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \bar{x}^2 = 7.8 - 4.84 = 2.96$

resultando el intervalo

$$\left[17.62 \pm 5.8409 \times 0.878\sqrt{1 + \frac{1}{5} + \frac{3.24}{5 \times 2.96}}\right]$$

es decir, $[11.511, 23.729]$.

□

Problema 3. Nos dicen que un programa de ordenador genera observaciones de una distribución *normal estándar*, es decir, $N(0;1)$. En una muestra aleatoria de 100 observaciones producidas mediante dicho programa se obtienen los siguientes resultados:

6 observaciones menores que -1.5,	20 entre -1.5 y -0.5,
30 entre -0.5 y 0,	25 entre 0 y 0.5,
15 entre 0.5 y 1.5,	4 mayores que 1.5

Sabiendo que una variable aleatoria Z con distribución $N(0;1)$ asigna las probabilidades: $P(Z < -1.5) = 0.07$, $P(-1.5 \leq Z \leq -0.5) = 0.24$, $P(-0.5 < Z < 0) = 0.19$, ¿qué regla utilizaría para decidir, al nivel $\alpha = 0.01$, si el programa funciona correctamente o no?

Solución

Las diferencias entre la frecuencia observada y la frecuencia esperada bajo la hipótesis nula —la población es $N(0,1)$— se recogen en la tabla siguiente:

Clase	Fr. observada (o_i)	Fr. esperada (e_i)	$(o_i - e_i)^2$
menor que -1.5	6	7	1
entre -1.5 y -0.5	20	24	16
entre -0.5 y 0.0	30	19	121
entre 0.0 y 0.5	25	19	36
entre 0.5 y 1.5	15	24	81
mayor que 1.5	4	7	9

Como consecuencia, el estadístico χ^2 toma el valor

$$\sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = \frac{1}{7} + \frac{16}{24} + \frac{121}{19} + \frac{36}{19} + \frac{81}{24} + \frac{9}{7} = 13.733.$$

Al ser el valor obtenido de las tablas de una distribución χ^2 con 5 grados de libertad $\chi^2_{5;0.01} = 15.086 > 13.733$, los datos no detectan *mal funcionamiento* en el programa considerado.

□

Problema 4. Se ha realizado un experimento orientado a detectar diferencias en el comportamiento de la respuesta Y en tres situaciones diferentes: A, B y C. Tras algunos cálculos con los datos recogidos, se han obtenido los siguientes resultados:

- Suma de cuadrados total $\boxed{172.93}$
- Suma de cuadrados "entre grupos" $\boxed{66.93}$
- Grados de libertad "dentro de los grupos" $\boxed{12}$
- Cuadrado medio "dentro de los grupos" $\boxed{8.83}$

Se pide:

(a) Construir la tabla de análisis de la varianza.

(b) ¿Cuántas unidades componen la muestra?

(c) ¿Qué conclusiones se obtienen a partir de la información contenida en dicha tabla?

Solución

(a) La tabla completa de análisis de la varianza es

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico
Entre grupos	66.93	3-1=2	66.93/2=33.465	$\frac{33.465}{8.83} = 3.79$
Dentro de los grupos	172.93-66.93=106	12	106/12=8.83	
Total	172.93	14		

(b) Los grados de libertad de la suma de cuadrados "dentro de los grupos" son la diferencia entre el número total de observaciones y el número de grupos. Por tanto, la muestra está compuesta por 15 unidades.

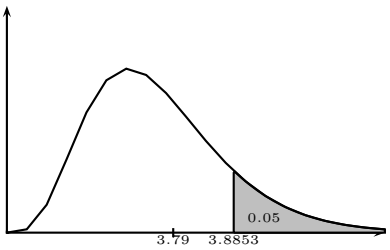


Figura 2: Punto crítico para una $F_{2,12}$ ($\alpha = 0.05$)

(c) De la comparación entre el valor del estadístico resultante del análisis de la varianza, 3.79, y el punto crítico obtenido de las tablas de la distribución F, dado por $F_{2,12;0.05} = 3.8853$ —véase figura 2—, se concluye que los datos no muestran una desviación significativa —con $\alpha = 0.05$— respecto a la hipótesis nula de igualdad de la respuesta media.

□

Problema 5. En las poblaciones **M** y **H** se han recogido dos muestras aleatorias independientes de tamaño 50. La frecuencia de aparición de cada una de las dos clases de la variable cualitativa Y se refleja en la siguiente tabla (la frecuencia que figura entre paréntesis es la esperada):

	Y_1	Y_2
M	14 (12)	36 (38)
H	10 (12)	40 (38)
Total	24	76

Se pide:

(a) Calcular el valor del estadístico χ^2.

(b) ¿Qué se puede concluir sobre la distribución de la característica Y?

Solución

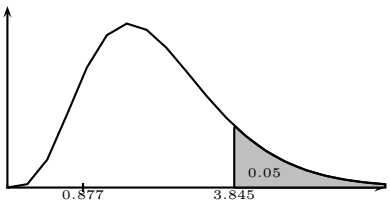


Figura 3: Punto crítico para una χ^2_1 ($\alpha = 0.05$)

(a) En la situación planteada —homogeneidad de poblaciones— el valor del estadístico χ^2 es

$$\sum_{j=1}^{2}\sum_{i=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(14-12)^2}{12}+\frac{(36-38)^2}{38}+\frac{(10-12)^2}{12}+\frac{(40-38)^2}{38} = 0.877$$

(b) Como $\chi^2_{1;0.05} = 3.845 > 0.877$ (ver figura 3), la homogeneidad de las distribuciones de Y —en la poblaciones **M** y **H**— es una hipótesis aceptable.

Problema 6. Se piensa que el tiempo de respuesta de un equipo informático, cuando se le solicita cierto tipo de información, sigue una distribución exponencial con parámetro $\lambda = 1$ seg. (por tanto, la densidad es $f(x) = e^{-x}$, para $x \geq 0$). Contraste dicha hipótesis utilizando los siguientes datos:

► Los puntos $0.22, 0.51, 0.92, 1.61$ determinan 5 clases equiprobables para la densidad citada.

► La frecuencia observada en cada una de estas clases, en un muestreo aleatorio de tamaño 40, es:

$$6, 8, 10, 7, 9$$

respectivamente.

Solución

Comparamos las frecuencias observada y esperada mediante el estadístico $\chi^2 = \sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i}$. Los resultados se muestran en la siguiente tabla

Clases	O_i	e_i	$(O_i - e_i)^2$	$\frac{(O_i - e_i)^2}{e_i}$	$\sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i}$
Menores que 0.22	6	8	4	$4/8$	1.25
$0.22 - 0.51$	8	8	0	0	
$0.51 - 0.92$	10	8	4	$4/8$	
$0.92 - 1.61$	7	8	1	$1/8$	
Mayores que 1.61	9	8	1	$1/8$	

El procedimiento que vamos a aplicar consiste en rechazar la hipótesis de interés —para el nivel de significación α— si se satisface la desigualdad

$$\sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i} > \chi^2_{5-1;\alpha}$$

En nuestro caso, el estadístico proporciona el valor 1.25 y, según los datos auxiliares, $\chi^2_{4;0.05} = 9.488$. Como consecuencia, no hay suficiente evidencia —al nivel de significación 0.05— contra la hipótesis de que "*el tiempo de respuesta sigue una distribución exponencial con parámetro* $\lambda = 1$ *seg* ", ya que el valor del estadístico es inferior al punto crítico —véase la figura 4—.

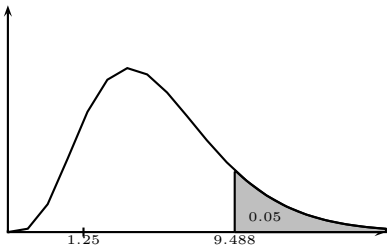


Figura 4: Punto crítico para una χ^2_4 ($\alpha = 0.05$)

□

Problema 7. Para estudiar la posible relación entre dos variables, X e Y, se registraron los valores de dichas variables en una muestra aleatoria formada por 20 individuos. De la información recogida se obtuvieron los siguientes datos estadísticos:

$$\frac{1}{20}\sum x_i = 53.69 \qquad \frac{1}{20}\sum y_i = 58.81$$

$$\frac{1}{20}\sum x_i^2 = 2948.82 \qquad \frac{1}{20}\sum y_i^2 = 3505.15$$

$$\frac{1}{20}\sum x_i y_i = 3197.39$$

Suponiendo un modelo de regresión lineal, ¿se puede concluir, con un nivel de significación $\alpha = 0.05$, que la pendiente de la recta de Y sobre X es positiva?

Solución

La pregunta alude al contraste

$$H_0 : \beta_1 \leq 0 \quad , \quad H_1 : \beta_1 > 0$$

Entonces, al nivel $\alpha = 0.05$, rechazaremos la hipótesis nula —concluyendo que β_1 es positiva— cuando

$$\frac{\hat{\beta}_1}{S_R\sqrt{\frac{1}{20 v_x}}} > t_{18;0.05}$$

Con los datos de este ejercicio resulta:

- $\hat{\beta}_1 = \frac{cov_{x,y}}{v_x} = \frac{3197.39 - 53.69 \times 58.81}{2948.82 - 53.69^2} = 0.60$
- $S_R = \sqrt{31.36} = 5.6$
- $\sqrt{\frac{1}{20(2948.82 - 53.69^2)}} = 0.03$

y, como consecuencia, $\dfrac{\hat{\beta}_1}{S_R\sqrt{\frac{1}{n v_x}}} = \dfrac{0.6}{5.6 \times 0.03} = 3.57.$

Dado que $t_{18;0.05} = 1.734$, es decir, el valor del estadístico es mayor que el punto crítico —véase la figura 5—, los datos nos permiten concluir que la pendiente β_1 es positiva.

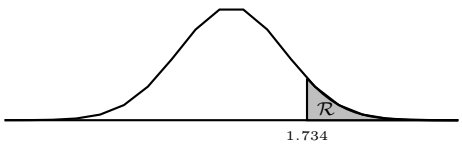


Figura 5: Punto crítico para una t_{18} ($\alpha = 0.01$)

□

Problema 8. Se está realizando un estudio sobre los fallos de un dispositivo electrónico. Este elemento se puede montar en dos posiciones diferentes y hay cuatro tipos de fallos posibles. Un muestreo aleatorio proporciona la siguiente distribución de frecuencias:

	Tipo de fallo			
Posición de Montaje	*A*	*B*	*C*	*D*
1	14	18	8	20
2	6	12	12	10

¿Concluiría que el tipo de fallo es independiente de la posición de montaje?

Solución

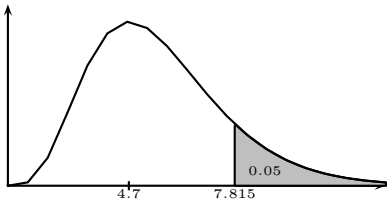


Figura 6: Punto crítico para una χ^2_3 ($\alpha = 0.05$)

Rechazaremos la hipótesis nula (*Posición de Montaje* independiente del *Tipo de Fallo*), con un nivel de sgnificación α, si

$$\sum_{j=1}^{4}\sum_{i=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} > \chi^2_{(2-1)(4-1);\alpha}$$

El enunciado nos proporciona la frecuencia observada (o_{ij}); multiplicando las correspondientes frecuencias marginales y dividiendo por el tamaño de la muestra obtenemos la frecuencia esperada bajo la hipótesis nula (e_{ij}):

	Tipo de fallo				
Posición de Montaje	*A*	*B*	*C*	*D*	
1	12	18	12	18	60
2	8	12	8	12	40
	20	30	20	30	100

Con estos datos podemos realizar la operación $(o-e)^2/e$ para cada celda de la tabla, resultando:

	Tipo de fallo			
Posición de Montaje	*A*	*B*	*C*	*D*
1	4/12	0	16/12	4/18
2	4/8	0	16/8	4/12

Entonces,

$$\sum_{j=1}^{4}\sum_{i=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{4+16+4}{12}+\frac{16+4}{8}+\frac{4}{18} = \frac{144+180+16}{72} = \frac{340}{72} \simeq 4.7$$

Como $4.7 < \chi^2_{3;0.05} = 7.815$ (véase la figura 6), concluimos que, con un nivel de significación $\alpha = 0.05$, la condición de **independencia** es **aceptable**.

□