ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2010-2011.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Como todos los años, el 10 % de la población enfermará de gripe en otoño. Si en primavera se eligen 7 personas al azar de la población, calcular la probabilidad de que no haya padecido la gripe ninguna persona. Y si se eligen 1000 personas, ¿cuál es la probabilidad de que haya más de 90 que la hayan padecido?

Problema 2

Se ha comprobado que el número de intervenciones en el foro virtual de una asignatura de la UNED sigue una distribución de Poisson y se desea determinar un intervalo de confianza, de coeficiente de confianza 0′95, para la desviación típica de dicho número de intervenciones. Para ello se eligieron al azar las intervenciones de 30 alumnos en el foro, anotándose el número de veces que habían participado en dicho foro estos alumnos. Determine el intervalo de confianza buscado si los resultados fueron:

2 , 4 , 2 , 1 , 3 , 3 , 5 , 1 , 2 , 5 , 4 , 1 , 4 , 3 , 3

5 , 3 , 1 , 3 , 2 , 6 , 4 , 3 , 2 , 3 , 4 , 3 , 3 , 2 , 3

Problema 3

Se desea saber si el porcentaje de aprobados en una determinada asignatura es significativamente superior al 75 %. Para ello se eligieron al azar 150 individuos que se había examinado de ella, obteniéndose de dicha muestra un porcentaje de aprobados del 77 %. ¿Se puede concluir que es significativamente superior al 75 %?

Problema 1

El problema se puede formalizar mediante un modelo binomial (EBR-sección 4.4.1) en donde cada prueba de Bernoulli sea el preguntar a la persona elegida si padeció la gripe el otoño pasado, y el suceso *éxito*, el que responda que sí. De esta forma, la variable *número de personas, de entre los siete, que padecieron gripe el otoño*, se puede modelizar mediante una variable X con distribución binomial $B(7, 0'1)$, al ser $p = 0'1$ la probabilidad de que se dé el suceso *éxito*.

La probabilidad pedida será ahora, utilizando la Tabla 1 de la distribución binomial (ADD, página 30)

$$P\{X = 0\} = 0'4783.$$

También se podía calcular esta probabilidad sin utilizar la Addenda, mediante la función de masa de la binomial:

$$P\{X = 0\} = \binom{7}{0} 0'1^0\ 0'9^7 = 0'9^7 = 0'4782969.$$

Si se aumenta el número de *pruebas de Bernoulli*, modelizándose el problema con una variable $X \rightsquigarrow B(1000, 0'1)$, el cálculo de probabilidades de distribuciones binomiales para un gran número de ensayos, como aquí ocurre, se realiza aproximando dicha distribución mediante el *teorema central del límite* (EBR-sección 4.7).

En el caso de una distribución binomial $X \rightsquigarrow B(n, p)$, su aproximación mediante una normal

$$N(np\,,\ \sqrt{np(1-p)}\,)$$

es válida (EBR-sección 4.7) cuando, supuesto sea $p \leq 0'5$ (como aquí ocurre) entonces sea también $np > 5$ (como aquí ocurre).

Por tanto, aproximaremos la $X \rightsquigarrow B(1000, 0'1)$, por una

$$N\left(1000 \cdot 0'1\,,\ \sqrt{1000 \cdot 0'1 \cdot 0'9}\right) = N(100\,,\ 9'4868)$$

quedando la probabilidad pedida igual a

$$P\{X > 90\} = P\left\{\frac{X - 100}{9'4868} > \frac{90 - 100}{9'4868}\right\} \simeq P\{Z > -1'054\} = 1 - P\{Z \leq -1'054\} = 1 - P\{Z > 1'054\}$$

siendo Z una variable aleatoria $N(0, 1)$.

Se podría aproximar ahora la probabilidad $P\{Z > 1'054\}$ por $P\{Z > 1'055\}$ y como este valor es equidistante de $P\{Z > 1'05\}$ y $P\{Z > 1'06\}$, de la Tabla 3 de la Addenda, página 33, tenemos que es $P\{Z > 1'05\} = 0'1469$

y $P\{Z > 1'06\} = 0'1446$, por lo que la semisuma de estas dos probabilidades nos daría el valor buscado, $(0'1469 + 0'1446)/2 = 0'14575$.

Por mostrar cómo se calcularía con más precisión esta probabilidad mediante interpolación lineal, haríamos la regla de tres que dice que, si para un aumento de abscisas de $1'06-1'05 = 0'01$ hay una disminución de probabilidad de $0'1469-0'1446 = 0'0023$, para un aumento de abscisas de $1'054-1'05 = 0'004$ habrá una disminución de probabilidad de $0'004 \cdot 0'0023/0'01 = 0'00092$, por lo que será $P\{Z > 1'054\} = 0'1469 - 0'00092 = 0'14598$, y por tanto, la probabilidad buscada,

$$P\{X > 90\} \simeq 1 - P\{Z > 1'054\} = 1 - 0'14598 = 0'85402.$$

Problema 2

Según el enunciado, la variable en observación sigue una distribución de Poisson de parámetro, digamos, λ. Como la media y la varianza de esta distribución son iguales al parámetro (EBR-sección 4.4.2), primero determinaremos el intervalo de confianza para λ, es decir, para la media, y luego extraeremos la raíz cuadrada de los extremos del intervalo así calculado.

El intervalo de confianza para λ, de coeficiente de confianza $1-\alpha$ es (EBR-sección 6.3 y ADD, página 15))

$$\left[\overline{x} - z_{\alpha/2}\sqrt{\overline{x}/n}\ ,\ \overline{x} + z_{\alpha/2}\sqrt{\overline{x}/n}\right] = [3-1'96\sqrt{3/30}\ ,\ 3-1'96\sqrt{3/30}\,] = [2'3802\ ,\ 3'6198]$$

con lo que el correspondiente para $\sqrt{\lambda}$ será

$$\left[\sqrt{\overline{x} - z_{\alpha/2}\sqrt{\overline{x}/n}}\ ,\ \sqrt{\overline{x} + z_{\alpha/2}\sqrt{\overline{x}/n}}\right] = [\sqrt{2'3802}\ ,\ \sqrt{3'6198}\,] = [1'5428\,,\ 1'9026].$$

Problema 3

El problema se puede modelizar mediante una binomial $B(1,p)$ en donde *éxito* sea aprobar la asignatura en cuestión y fracaso, no aprobarla. En este modelo p es el porcentaje de aprobados (parámetro poblacional) sobre el que queremos saber si puede admitirse que es $p > 0'75$ utilizando una muestra aleatoria de $n = 150$ individuos, muestra que suministró un porcentaje de aprobados (de *éxitos*) de $\widehat{p} = 0'77$.

Con este esquema estamos ante el caso de un contraste de hipótesis para la media de una población no necesariamente normal, muestras grandes, EBR-sección 7.3 y ADD página 15; más en concreto de Poblaciones binomiales.

Como habitualmente ocurre, lo que queremos contrastar se establece como hipótesis alternativa de manera que contrastaremos $H_0 : p \leq 0'75$ frente a

$H_1 : p > 0'75$ (además no tenemos fórmulas para contrastar *mayor* frente a *menor o igual*), rechazándose H_0 cuando (EBR-sección 7.3 y ADD, página 15) sea

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} > z_\alpha.$$

Se podría fijar un nivel de significación α, calcular el valor de la abscisa de una normal N(0,1) que deje a la derecha un área de probabilidad α, es decir, calcular z_α y comparar el valor del estadístico de contraste

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

con z_α, para tomar la decisión de aceptar o rechazar H_0. No obstante, lo más habitual y conveniente es calcular el p-valor del test cosa que, en todo caso, siempre hay que hacer para valorar la decisión así tomada.

Es decir, en lugar de fijar un nivel de significación, ejecutar el test y luego valorar la decisión tomada, lo más razonable es calcular el p-valor directamente y, en base a éste, tomar la decisión de aceptar o rechazar H_0.

Dado que el estadístico de contraste toma para los datos del problema el valor

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} = \frac{0'77 - 0'75}{\sqrt{\frac{0'75 \cdot 0'25}{150}}} = 0'5657$$

el p-valor del test es $P\{Z > 0'5657\}$, valor que no viene exactamente en las tablas de la normal N(0,1). Se podría calcular por interpolación, pero en realidad sólo necesitamos acotarlo porque (EBR-página 194), un p-valor $0'2$ o mayor indica aceptar la hipótesis nula y un p-valor $0'01$ o menor, indica rechazar H_0, indicando valores intermedios del p-valor la toma de otra muestra.

Como es

$$P\{Z > 0'57\} < P\{Z > 0'5657\} < P\{Z > 0'56\}$$

de la Tabla 3 (ADD, página 33), vemos que es

$$0'2843 < P\{Z > 0'5657\} < 0'2877$$

es decir, suficientemente grande como para no poder rechazar la hipótesis nula, aceptándola en consecuencia. En definitiva, con sólo el 77 % de aprobados en nuestra muestra, no podemos concluir que el porcentaje de aprobados de la asignatura sea superior al 75 %.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2010-2011.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Si X es una variable aleatoria con distribución normal de media 3 y desviación típica 2, calcular de forma razonada las probabilidades, $P\{|3 - X| > 2\}$ y $P\{|2\,X| \leq 3\}$.

Problema 2

Los siguientes datos proceden de un ensayo clínico llevado a cabo por Ezdinli y otros en 1976, para comparar dos tratamientos en el linfoma de linfocitos. Un tratamiento, denominado CP, administrado a 135 pacientes, consistía en cytoxan + prednisone, y otro tratamiento, denominado BP y administrado a 138 pacientes, estaba compuesto por carmustina (BCNU) + prednisone. La variable observada fue la respuesta del tumor en cada paciente, medida en una escala cualitativa desde "Respuesta Completa" (lo mejor que puede ocurrir) a "Progresa" (lo peor que puede ocurrir). Los datos obtenidos son los que aparecen en la siguiente tabla:

	Respuesta Completa	Respuesta Parcial	Sin Cambios	Progresa
BP	26	51	21	40
CP	31	59	11	34

¿Difieren significativamente los tratamientos en su eficacia?

Problema 3

La evolución del Producto Interior Bruto en España, medido por la tasa de variación interanual en cuatro momentos de 2008 y 2009, viene dada por la siguiente tabla

Tiempo X	1	2	3	4	5	6	7	8
Tasa Y	$2'5$	$1'7$	$0'5$	$-1'2$	$-3'2$	$-4'2$	$-4'0$	$-3'1$

Determine la recta de mínimos cuadrados (también llamada de regresión de Y sobre X) de los datos anteriores y analice si es o no significativa mediante un test de hipótesis.

Problema 1

El objetivo que perseguimos es transformar las probabilidades buscadas en términos de la normal N(0,1) considerando sucesos equivalentes, es decir con la misma probabilidad, porque de esta distribución tenemos tablas.

Aunque hay varias formas de llegar al resultado buscado, la más rápida es, seguramente, la siguiente,

$$P\{|3-X|>2\} = P\{|X-3|>2\} = P\{X-3>2\} + P\{X-3<-2\}.$$

Como $X \rightsquigarrow N(3,2)$, para conseguir la $Z \rightsquigarrow N(0,1)$, ya sólo tenemos que dividir por la desviación típica,

$$= P\{Z>1\} + P\{Z<-1\} = 2\cdot P\{Z>1\}$$

en donde la última igualdad se obtiene por ser simétrica la distribución N(0,1). En general, suele ayudar mucho un dibujo de la densidad N(0,1) en el cálculo de las probabilidades de áreas bajo esta curva.

Ahora, sólo tenemos que buscar en la Tabla 3 de N(0,1), ADD página 33, para completar el ejercicio,

$$P\{|3-X|>2\} = 2\cdot P\{Z>1\} = 2\cdot 0'1587 = 0'3174.$$

Respecto a la otra probabilidad, con los mismos argumentos y notación,

$$P\{|2\,X|\leq 3\} = P\{-3\leq 2\,X\leq 3\} = P\{-1'5\leq X\leq 1'5\} = P\{-2'25\leq Z\leq -0'75\} =$$

$$= P\{0'75\leq Z\leq 2'25\} = P\{Z>0'75\} - P\{Z>2'25\} = 0'2266 - 0'0122 = 0'2144.$$

Problema 2

Como los datos aportados son recuentos de observaciones clasificados por clases, comparar ambas poblaciones debe hacerse mediante un test de la χ^2 de *homogeneidad de varias muestras* (EBR-sección 8.2.3), en donde la hipótesis nula que se establece es que ambos tratamientos pueden considerase homogéneos. Esta hipótesis nula se rechazará cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i\, m_j/n)^2}{n_i\, m_j/n}$$

el estadístico de Pearson que, observar, siempre es la suma de las *frecuencias observadas menos las esperadas al cuadrado, dividido por las esperadas*. En nuestro caso, toma el valor $\lambda = 4'5995$.

De la Tabla 4 de la χ^2 de Pearson (ADD, página 34) vemos que el p-valor es

$$P\{\chi^2_{(r-1)(s-1)} > 4'5995\} = P\{\chi^2_3 > 4'5995\} > P\{\chi^2_3 > 6'251\} = 0'1$$

suficientemente grande como para aceptar la hipótesis nula de homogeneidad con bastante seguridad.

Problema 3

La recta de regresión (EBR-sección 10.2) o mínimos cuadrados (EBR-sección 2.4.2)

$$y = \widehat{\beta}_0 + \widehat{\beta}_1\, x$$

es la que tiene por coeficientes los valores

$$\widehat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{8\cdot(-91'4) - 36\cdot(-11)}{8\cdot 204 - 1296} = -0'997$$

y

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \frac{-11 - (-0'997)\cdot 36}{8} = 3'11.$$

Es decir,

$$y = 3'11 - 0'997\, x$$

Analizar si esta recta es significativa o no, es ejecutar el contraste de la regresión lineal simple (EBR-sección 10.3), para lo que debemos construir la tabla de Análisis de la Varianza (EBR página 301 y ADD página 26)

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal simple	$SSEX = \widehat{\beta_1}^2 \left(\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \right)$	1	$SSEX$	$\frac{SSEX}{\frac{SSNEX}{n-2}}$
Residual	$SSNEX = SST - SSEX$	$n-2$	$\frac{SSNEX}{n-2}$	
Total	$SST = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$	$n-1$		

que en nuestro problema toma el valor

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal	$SSEX = (-0'997)^2 \cdot (204 - 1296/8) = 41'75$	1	$41'75$	$F = 33'64$
Residual	$SSNEX = 49'195 - 41'75 = 7'445$	6	$1'241$	
Total	$SST = 64'32 - 121/8 = 49'195$	7		

Ahora el test que se plantea es H_0 : *La regresión no es significativa*, o lo que es lo mismo, H_0 : *La recta de regresión no me sirve para explicar la variable Y en función de la X*, frente a la hipótesis alternativa, H_1 : *La regresión es significativa*. Para ello, lo mejor es determinar el p-valor o una acotación suya. Puesto que el estadístico del contraste sigue una F de Snedecor con $(1, 6)$ grados de libertad, el p-valor del test es, utilizando la Tabla 6 de la distribución F de Snedecor (ADD, página 36)

$$\text{p-valor} = P\{F_{(1,6)} > 33'64\} < P\{F_{(1,6)} > 18'635\} = 0'005.$$

Es decir, suficientemente pequeño como para rechazar la hipótesis nula y concluir con que la recta de regresión o mínimos cuadrados determinada, es válida para predecir la evolución del Producto Interior Bruto.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2010-2011.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

El flujo de tráfico es una variable aleatoria X que se suele modelizar mediante una distribución de Poisson. Se observó durante cierto tiempo un cruce determinado de una avenida en una gran ciudad, cuyo número de coches se modelizó mediante una distribución de Poisson, sacándose además la conclusión de que el número medio de coches que pasan por el cruce era de 8 coches por minuto. Con objeto de establecer la duración del semáforo en dicho cruce conteste a las siguientes preguntas:

a) Calcular la probabilidad de que no haya coches en el cruce, en un periodo elegido al azar, de 30 segundos.

b) Calcular la probabilidad de que haya 2 coches o más en el cruce, en un periodo elegido al azar, de 30 segundos.

c) Calcular el número mínimo de coches en el cruce de manera que la probabilidad de observar en el cruce este número o un número menor de coches en un periodo de 30 segundos, sea al menos del 95 %.

d) Si la varianza del número de coches por minuto en la intersección es 20, ¿cree que el modelo Poisson es apropiado? Explicar la respuesta.

Problema 2

Luigi Fortuna y otros (1995, pág. 76) estudiaron la aceleración de redes neuronales celulares obteniendo los siguientes datos:

2′51 , 3′35 , 2′85 , 3′45 , 3′51 , 3′18 , 3′53 , 3′25 , 3′54 , 3′29 , 3′54 , 3′33 , 3′55

Suponiendo que puede admitirse que estos datos proceden de una distribución normal, determinar un intervalo de confianza de coeficiente de confianza de 0′95 para la aceleración media.

Problema 3

Williams, Lockley y Hurst (1981) tomaron muestras de braquiópodos fósiles en los niveles CD8 y CD9 de una excavación en el centro de Gales. El número de fósiles según las especies encontradas aparece en la siguiente tabla:

	Nivel CD8	Nivel CD9
Macrocoelia	4	2
Sowerbyella	25	36
Dalmanella	15	8
Hesperorthis	3	2
Crinoideos	8	3
Briozoos ramificados	4	5

A la vista de estos datos, ¿existen diferencias significativas entre ambos niveles?

Problema 1

Si consideramos la variable aleatoria X =*número de coches en un periodo de 30 segundos*, en el cruce en cuestión, es razonable, según el enunciado modelizar X con una distribución de Poisson de media 4, es decir, de parámetro 4 por ser la media de la distribución de Poisson, su parámetro.

a) La probabilidad pedida será

$$P\{X = 0\} = 0'0183$$

utilizando unas tablas de la distribución de Poisson.

b) La probabilidad pedida será

$$P\{X \geq 2\} = 1 - P\{X < 2\} = 1 - (P\{X = 0\} + P\{X = 1\}) = 1 - (0'0183 + 0'0733) = 0'9084.$$

c) Si denominamos n al número buscado, la ecuación planteada por el enunciado es

$$P\{X \leq n\} \geq 0'95$$

o bien, haciendo algunas operaciones, n debe de ser tal que

$$P\{X > n\} \leq 0'05.$$

Si miramos la línea de $\lambda = 4$ en la tabla de la distribución de Poisson, sumando los últimos 4 valores obtenemos

$$0'0298 + 0'0132 + 0'0053 + 0'0019 = 0'0502 > 0'05$$

con lo que nos pasamos de $0'05$. Por tanto debe ser $n = 8$ para sumar sólo los tres últimos y que sea

$$P\{X > 8\} = P\{X = 9\} + P\{X = 10\} + P\{X = 11\} = 0'0132 + 0'0053 + 0'0019 = 0'0204 < 0'05.$$

d) Una de las características de la distribución de Poisson es que su parámetro es su media y su varianza. Si los datos del cruce suministran una media de 4 coches cada 30 segundos y una varianza de 10 coches cada 30 segundos, que son valores muy distintos, el modelizar X con una distribución de Poisson no parece muy razonable.

Problema 2

Nos piden determinar un intervalo de confianza para la media de una distribución normal de varianza desconocida y tamaños muestrales pequeños (EBR-sección 6.2 ó CB-sección 6.2) cuya expresión es

$$\left[\overline{x} - t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \, , \, \overline{x} + t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \, \right].$$

De los datos se deduce que es $\overline{x} = 3'3$, $S = 0'31$ y $t_{n-1;\alpha/2} = t_{12;0'025} = 2'179$, por lo que el intervalo de confianza buscado será

$$\left[\overline{x} - t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \, , \, \overline{x} + t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \, \right] = \left[3'3 - 2'179 \frac{0'31}{\sqrt{13}} \, , \, 3'3 + 2'179 \, \frac{0'31}{\sqrt{13}} \, \right]$$

$$= [3'11 \, , \, 3'49].$$

Problema 3

Estamos ante un caso de comparación de dos poblaciones, *fósiles braquiópodos en el nivel CD8* y *fósiles braquiópodos en el nivel CD9*, en donde los datos son recuentos de observaciones. Por tanto, la comparación debe realizarse con un test de la χ^2 de *homogeneidad de varias muestras* (EBR-sección 8.2.3), en donde la hipótesis nula que se establece es que ambas poblaciones pueden considerase homogéneas. Ésta se rechaza cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_i \, m_j/n)^2}{n_i \, m_j/n}$$

el estadístico de Pearson. En nuestro caso, toma el valor $\lambda = 7'29$.

De la tabla de la χ^2 de Pearson vemos que el p-valor es

$$0'1 < P\{\chi^2_{(r-1)(s-1)} > 7'29\} = P\{\chi^2_5 > 7'29\} < 0'3$$

que, en principio es suficientemente grande como para aceptar la hipótesis nula de homogeneidad.

Como algunas frecuencias esperadas son menores que 5, se podría agrupar algunas clases contiguas, argumento a añadir a que el p-valor no es contundente. Así, la matriz de datos que utilizaremos en el test es

	Nivel CD8	Nivel CD9
Macrocoelia y Sowerbyella	29	38
Dalmanella y Hesperorthis	18	10
Crinoideos y Briozoos ramificados	12	8

para la que se obtendría un valor del estadístico de contraste igual a $\lambda = 4'2193$ y un p-valor de $0'1213$. Aunque ha bajado un poco, sigue siendo suficientemente grande como para admitir que las muestras de ambos niveles de excavación pueden considerarse homogéneas.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 41206AD01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2011-2012.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Sea X una variable aleatoria discreta que toma los valores -2 , 0 , 2 y 3 con probabilidades respectivas 1/4 , 1/3 , 1/3 y 1/12 . Calcular la función de distribución de X así como su media y su varianza.

Problema 2

Se quiere analizar si existen o no diferencias significativas en el consumo habitual de dieta mediterránea con aceite de oliva virgen (Tratamiento 1) frente a una dieta de adelgazamiento habitual (Tratamiento 2) en las enfermedades del corazón. Para ello se seleccionaron al azar 181 individuos para el Tratamiento 1 y a 177 para el Tratamiento 2 y se analizó en ambos grupos, al cabo de tres meses, la diferencia de las medias muestrales en apolipoproteína B del grupo de Tratamiento 1 menos el del Tratamiento 2. Si ésta fue $-2'9$ mg/dL y las cuasivarianzas muestrales en los dos grupos fueron $S_1^2 = 160$ y $S_2^2 = 170$, ¿existen diferencias significativas?

Problema 3

Un psicólogo ha dividido los estados de ánimo de sus pacientes en una escala que va de 0 a 10, correspondiendo una mayor puntuación a un mejor estado de ánimo, siendo el valor 0 un estado de ánimo de "depresión", 5 un estado de ánimo "normal" y 10 un estado de ánimo "eufórico". Con objeto de analizar el estado de ánimo de sus pacientes durante los meses veraniegos, eligió al azar 10 de ellos y, después de una serie de tests psicológicos obtuvo como estado de ánimo los siguientes valores:

$$4 \ , \ 4'2 \ , \ 3 \ , \ 6 \ , \ 5'3 \ , \ 7 \ , \ 3'4 \ , \ 8 \ , \ 2'1 \ , \ 3'5$$

A nivel de significación $\alpha = 0'05$ y utilizando el test de los signos, ¿puede admitirse un estado de ánimo en promedio menor del "normal" en los meses de verano analizados?

Problema 1

Del enunciado se desprende que X es una variable aleatoria discreta, pues toma valores aislados, con función de masa (EBR-sección 4.2)

$$p_X(x) = \begin{cases} 1/4 & \text{si} \quad x = -2 \\ 1/3 & \text{si} \quad x = 0 \\ 1/3 & \text{si} \quad x = 2 \\ 1/12 & \text{si} \quad x = 3 \end{cases}$$

Como la función de distribución de una variable aleatoria discreta, como X, es una función en escalera que da saltos en los valores de X, siendo el tamaño de los saltos igual a los valores de la función de masa allí, la función de distribución de X será la función de x (que da el valor de la probabilidad acumulada hasta el punto x),

$$F(x) = \begin{cases} 0 & \text{si} \quad x < -2 \\ 1/4 & \text{si} \quad -2 \leq x < 0 \\ 1/4 + 1/3 = 7/12 & \text{si} \quad 0 \leq x < 2 \\ 7/12 + 1/3 = 11/12 & \text{si} \quad 2 \leq x < 3 \\ 11/12 + 1/12 = 1 & \text{si} \quad x \geq 3 \end{cases}$$

La media de X, al ser esta variable una variable aleatoria discreta, es la suma de los valores que toma X por las probabilidades con que los toma,

$$E[X] = -2 \cdot \frac{1}{4} + 0 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{12} = \frac{5}{12}.$$

Para calcular la varianza es más simple calcular primero la media de los cuadrados $E[X^2]$ y después aplicar la expresión,

$$V(X) = E[X^2] - (E[X])^2.$$

Como es

$$E[X^2] = 4 \cdot \frac{1}{4} + 0 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 9 \cdot \frac{1}{12} = \frac{37}{12}$$

será

$$V(X) = E[X^2] - (E[X])^2 = \frac{37}{12} - \frac{25}{144} = \frac{419}{144} = 2'91.$$

Problema 2

Si denominamos μ_1 al nivel medio poblacional de apolipoproteína B del grupo de Tratamiento 1 y μ_2 al nivel medio poblacional de apolipoproteína B

del grupo de Tratamiento 2, se desea saber si existen o no diferencias significativas entre μ_1 y μ_2, es decir, contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$.

Del enunciado se desprende que no sabemos la distribución de los niveles de apolipoproteína en las poblaciones pero que tenemos tamaños muestrales suficientemente grandes, $n_1 = 181$ y $n_2 = 177$, con lo que estamos ante un caso de comparación de las medias de dos poblaciones no necesariamente normales muestras grandes (EBR-sección 7.7), siendo las varianzas poblacionales desconocidas. En este caso, se rechaza la hipótesis nula anterior cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} > z_{\alpha/2}$$

De nuestros datos se obtuvo que $\overline{x}_1 - \overline{x}_2 = -2'9$, que $S_1^2 = 160$ y que $S_2^2 = 170$, con lo que será

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} = \frac{|-2'9|}{\sqrt{\dfrac{160}{181} + \dfrac{170}{177}}} = 2'14.$$

Si se toma como nivel de significación $\alpha = 0'05$ será $z_{\alpha/2} = z_{0'025} = 1'96 < 2'14$ rechazándose, por tanto, la igualdad de las medias de ambas poblaciones, es decir, concluyendo que sí existen diferencias significativas entre ambos tratamientos a ese nivel de significación.

Si, en lugar de fijar el nivel de significación calculamos el p-valor, camino más recomendable puesto que al final habrá que calcularlo, éste será,

$$\text{p-valor} = P\{|Z| > 2'14\} = 2 \cdot P\{Z > 2'14\} = 2 \cdot 0'0162 = 0'0324$$

valor no concluyente, puesto que no es menor que $0'01$, aunque sí se podría afirmar que existen diferencias significativas.

Problema 3

Este ejercicio es similar al Problema 11.1 del texto PREB. Si representamos por X a la variable aleatoria *estado de ánimo del paciente* elegido al azar, podemos resumir la distribución de X, la cual es completamente desconocida, por una medida de posición como su mediana M.

Como la situación de "normalidad" se ha establecido en el valor 5 de la variable, la hipótesis que estamos interesados en validar es $M < 5$, por lo que contrastaremos la hipótesis nula $H_0 : M \geq 5$ frente a la alternativa $H_1 : M < 5$.

Además, como los estados de ánimo asignados por la psicóloga no representan, en realidad, una puntuación numérica sino más bien una *ordenación* de los pacientes, por eso el test a utilizar será el *test de los signos* (EBR-sección 8.3.1).

Si T representa el *número de diferencias* $X_i - 5$ *positivas*, el test de los signos indica rechazar H_0 cuando sea

$$T \leq n - t_\alpha$$

siendo t_α el menor entero tal que $P\{W \geq t_\alpha\} \leq \alpha$, en donde W es una variable aleatoria con distribución binomial $B(n, 0'5)$.

Las diez diferencias $X_i - 5$ son

$$-1\ ,\ -0'8\ ,\ -2\ ,\ 1\ ,\ 0'3\ ,\ 2\ ,\ -1'6\ ,\ 3\ ,\ -2'9\ ,\ -1'5$$

con lo que el número T de diferencias positivas es $T = 4$.

El nivel de significación indicado es $\alpha = 0'05$. Buscando en la tabla 1 de la distribución binomial $B(10, 0'5)$, obtenemos que es

$$P\{W \geq 8\} = P\{W = 8\} + P\{W = 9\} + P\{W = 10\} = 0'0439 + 0'0098 + 0'0010 = 0'0547 > 0'05$$

por lo que tenemos que quedarnos con el valor 9 ya que

$$P\{W \geq 9\} = P\{W = 9\} + P\{W = 10\} = 0'0098 + 0'0010 = 0'0108 \leq 0'05.$$

(Recordemos que si un número es menor que otro, entonces es menor o igual.)

Por tanto, el menor número entero que la verifica, el cual es por definición t_α, será $t_\alpha = 9$. Al ser $n - t_\alpha = 10 - 9 = 1$ y $T = 4$, es $T > n - t_\alpha$, aceptándose en consecuencia H_0 y concluyéndose, en definitiva que, en base a ese estudio, no se produce una disminución significativa del estado mediano "normal" de ánimo en los pacientes.

El p-valor del test es, a partir de la tabla 1,

$$P\{W \leq 4\} = 0'0010 + 0'0098 + 0'0439 + 0'1172 + 0'2051 = 0'377$$

suficientemente grande como para confirmar la aceptación de la hipótesis nula.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2011-2012.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Sea X una variable aleatoria con función de distribución

$$F(x) = \begin{cases} 0 & \text{si} \quad x < -2 \\ 1/3 & \text{si} \quad -2 \leq x < -1 \\ 1/2 & \text{si} \quad -1 \leq x < 1 \\ 1 & \text{si} \quad x \geq 1 \end{cases}$$

Calcular la función de masa de X así como su media y su varianza.

Problema 2

Se quiere averiguar si habitantes actuales de un país mediterráneo pueden considerarse descendientes de una antigua civilización mediterránea que vivió en esas tierras y que desapareció. Para ello se midió, en milímetros, la máxima anchura, X, de 6 cráneos de restos de la civilización desaparecida y la máxima anchura, Y, de la cabeza de 6 varones de los habitantes actuales, todos ellos elegidos al azar. Los resultados obtenidos fueron los siguientes:

Civilización Desaparecida	140	132	154	142	141	150
Habitantes Actuales	133	138	136	128	143	130

En base a los datos obtenidos y utilizando un contraste de Wilcoxon-Mann-Whitney, ¿se puede concluir con la existencia de diferencias significativas entre las dos poblaciones a nivel $\alpha = 0'05$?

Problema 3

En un reciente estudio se dividió a pacientes con problemas graves de arteriosclerosis en tres grupos: Grupo I, que consumieron

dieta mediterránea con un suplemento de aceite de oliva virgen (15 litros cada trimestre); Grupo II, de individuos que consumieron dieta mediterránea y un suplemento de nueces (30 gramos diarios); y el Grupo III de Control. En los tres grupos se midió, mediante imágenes de ultrasonido, el grosor en milímetros de una arteria carótida al concluir el tratamiento al cabo de un año, menos el grosor que tenían al inicio del estudio. Los resultados obtenidos fueron los siguientes:

Grupo I	**−0′079**	**−0′029**	**−0′085**	**−0′120**	**−0′082**
Grupo II	**−0′072**	**−0′062**	**−0′096**	**−0′123**	**−0′007**
Grupo III	**−0′001**	**−0′002**	**0′001**	**−0′006**	**0′002**

¿Existen diferencias significativas entre los tres grupos?

Problema 1

Al ser la función de distribución F una función en escalera, la variable X es de tipo discreto, por eso nos solicitan su función de masa (EBR-sección 4.2) $p_X(x)$ que nos va dando las probabilidades con las que toma X sus valores.

De la función de distribución se deduce que los valores que toma X son -2, -1 y 1 pues estos son los valores en donde F salta. La función de masa buscada será el valor de esos saltos. Así,

$$p_X(-2) = \frac{1}{3} - 0 = \frac{1}{3}$$

$$p_X(-1) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$

$$p_X(1) = 1 - \frac{1}{2} = \frac{1}{2}$$

Por tanto,

$$p_X(x) = \begin{cases} 1/3 & \text{si} \quad x = -2 \\ 1/6 & \text{si} \quad x = -1 \\ 1/2 & \text{si} \quad x = 1 \end{cases}$$

La media de X, al ser esta variable una variable aleatoria discreta, será la suma de los valores que toma X por las probabilidades con que los toma,

$$E[X] = -2 \cdot \frac{1}{3} - 1 \cdot \frac{1}{6} + 1 \cdot \frac{1}{2} = -\frac{1}{3}.$$

Para calcular la varianza es más simple calcular primero la media de los cuadrados $E[X^2]$ y después aplicar la expresión,

$$V(X) = E[X^2] - (E[X])^2.$$

Como es

$$E[X^2] = 4 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6} + 1 \cdot \frac{1}{2} = 2$$

será

$$V(X) = E[X^2] - (E[X])^2 = 2 - \frac{1}{9} = \frac{17}{9} = 1'89.$$

Problema 2

Este ejercicio es muy parecido al Ejercicio 8.5 de EBR y al Problema 8.2 de EEA.

Estamos ante un caso de comparación de dos poblaciones independientes mediante la comparación de sus medianas, al realizar el contraste de la hipótesis nula $H_0 : M_X = M_Y$ frente a la alternativa $H_1 : M_X \neq M_Y$, utilizando el *test de Wilcoxon-Mann-Whitney* (EBR-sección 8.4.1).

El estadístico del test de Wilcoxon-Mann-Whitney es

$$U = \sum_{i=1}^{6} \sum_{j=1}^{6} D_{ij}$$

con

$$D_{ij} = \begin{cases} 1 & \text{si es } Y_j < X_i \\ 0 & \text{si es } Y_j \geq X_i \end{cases}$$

es decir, el número de observaciones Y_j que preceden a cada X_i fijo. Si subrayamos los valores Y_j en la muestra combinada de las 12 observaciones, obtenemos

$$\underline{128}\ ,\ \underline{130}\ ,\ 132\ ,\ \underline{133}\ ,\ \underline{136}\ ,\ \underline{138}\ ,\ 140\ ,\ 141\ ,\ 142\ ,\ \underline{143}\ ,\ 150\ ,\ 154$$

que proporcionan un valor para U de

$$U = 2 + 5 + 5 + 5 + 6 + 6 = 29.$$

La región crítica de este contraste es

$$C = \{U \leq n \cdot m - u_{m,n;\alpha/2}\} \cup \{U \geq u_{m,n;\alpha/2}\}$$

siendo $u_{m,n;\alpha/2}$ el punto crítico del test. Como los tamaños muestrales son suficientemente grandes (mayores que 5) y similares, se puede utilizar la aproximación normal, fijado el nivel de significación $\alpha = 0'05$,

$$u_{m,n;\alpha/2} = \frac{m\,n}{2} + z_{\alpha/2}\sqrt{\frac{m\,n\,(n+m+1)}{12}} = \frac{6 \cdot 6}{2} + 1'96\sqrt{\frac{6 \cdot 6 \cdot 13}{12}} = 30'24$$

es decir, $C = \{U \leq 5'76\} \cup \{U > 30'24\}$. Como $U = 29$ no cae en la región crítica, se acepta la hipótesis nula y se concluye con que ambas poblaciones no presentan diferencias significativas en los tamaños de sus cráneos y que, utilizando esta característica como definitoria de la población, que los habitantes actuales tiene como origen la civilización desaparecida.

Problema 3

La comparación de más de dos poblaciones; en este caso, tres poblaciones se ejecuta con un Análisis de la varianza para un factor y un diseño completamente aleatoriazado (EBR-sección 9.2) .

Para contrastar $\left\{ \begin{array}{ll} H_0: & \mu_I = \mu_{II} = \mu_{III} \\ H_1: & \textit{alguna distinta} \end{array} \right.$ la tabla de Análisis de la Varianza que se obtiene es

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Tratamientos	$SST_i = 0'0185$	2	$0'00925$	$F = 9'44$
Residual	$SSE = 0'0118$	12	$0'00098$	
Total	$SST = 0'0303$	14		

Si fijamos un nivel de significación $\alpha = 0'01$, al ser $F = 9'44 > 8'5096 = F_{(2,12);0'005}$, se rechaza la igualdad de los tres tratamientos, H_0, además con un p-valor menor que $0'005$.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2011-2012.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Sea X una variable aleatoria continua con función de densidad

$$f(x) = \frac{2\,x}{\theta^2} \qquad , \qquad 0 < x < \theta$$

siendo $\theta > 0$ un parámetro fijo desconocido. Se pide: Comprobar que f es una función de densidad verificando que su integral vale 1, calcular la función de distribución de X, así como su media y su varianza. Calcular por último la probabilidad $P\{\theta/2 < X < \theta\}$.

Problema 2

El tiempo en minutos que tardan en ser atendidos los clientes de un banco, sigue una distribución normal de media desconocida y desviación típica igual a 2'5. Los tiempos de espera de 10 clientes fueron los siguientes:

4'5 , 5 , 6'5 , 4 , 10 , 4'5 , 7 , 8'5 , 5'5 , 12

Determinar el intervalo de confianza, de coeficiente de confianza 0'95, para el tiempo medio de espera.

Problema 3

Se preguntó a 200 individuos si practicaban o no al menos 5 horas de ejercicio moderado a la semana y si habían padecido o no infarto de miocardio. Los resultados aparecen recogidos en la siguiente tabla:

	Ejercicio SÍ	Ejercicio NO
INFARTO SÍ	**22**	**43**
INFARTO NO	**87**	**48**

¿Puede decirse que el hacer ejercicio moderado está relacionado significativamente con padecer un infarto de miocardio?

Problema 1

Este ejercicio es parecido al Problema 2.12 del texto PREB.

Una función será de densidad de alguna variable aleatoria si es positiva y si su integral vale 1. En este caso será,

$$\int_0^\theta f(x)\,dx = \int_0^\theta \frac{2\,x}{\theta^2}\,dx = \left|\frac{1}{\theta^2}x^2\right|_0^\theta = 1.$$

La función de distribución de X será (EBR-sección 4.2), si es $0 < x < \theta$,

$$F(x) = \int_{-\infty}^x f(y)\,dy = \int_0^x \frac{2\,y}{\theta^2}\,dy = \left|\frac{1}{\theta^2}y^2\right|_0^x = \frac{x^2}{\theta^2}.$$

En resumen, la función de distribución será

$$F(x) = \left\{\begin{array}{ccc} 0 & \text{si} & x \leq 0 \\ x^2/\theta^2 & \text{si} & 0 < x < \theta \\ 1 & \text{si} & x \geq \theta \end{array}\right.$$

La media de X será

$$E[X] = \int_{-\infty}^\infty x\,f(x)\,dx = \int_0^\theta x\,f(x)\,dx = \int_0^\theta \frac{2\,x^2}{\theta^2}\,dx = \left|\frac{2}{3\theta^2}x^3\right|_0^\theta = \frac{2\,\theta}{3}.$$

Para calcular la varianza es más simple calcular primero la media de los cuadrados $E[X^2]$ y después aplicar la expresión,

$$V(X) = E[X^2] - (E[X])^2.$$

Como es

$$E[X^2] = \int_{-\infty}^\infty x^2\,f(x)\,dx = \int_0^\theta x^2\,f(x)\,dx = \int_0^\theta \frac{2\,x^3}{\theta^2}\,dx = \left|\frac{2}{4\theta^2}x^4\right|_0^\theta = \frac{\theta^2}{2}$$

será

$$V(X) = E[X^2] - (E[X])^2 = \frac{\theta^2}{2} - \frac{4\,\theta^2}{9} = \frac{\theta^2}{18}.$$

Por último,

$$P\{\theta/2 < X < \theta\} = F(\theta) - F(\theta/2) = \frac{\theta^2}{\theta^2} - \frac{\theta^2}{4\,\theta^2} = \frac{3}{4}.$$

Problema 2

Este ejercicio es muy similar al Problema 4.2 del texto PREB y al Problema 2.10 del texto EEA.

Se trata de determinar el intervalo de confianza para la media de una población normal de varianza conocida (EBR-sección 6.2) dado por

$$\left[\overline{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]$$

es decir,

$$\left[6'75 - 1'96\,\frac{2'5}{\sqrt{10}}\ ,\ 6'75 + 1'96\,\frac{2'5}{\sqrt{10}}\right] = [5'2\ ,\ 8'3].$$

Problema 3

Este ejercicio es similar al Problema 10.2 del texto PREB y al Problema 7.11 del texto EEA.

Se trata de un contraste de independencia de caracteres (EBR-sección 8.2.4), *Ejercicio* e *Infarto*, en donde la hipótesis nula a contrastar es la independencia de ambos.

El estadístico de Pearson

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

siendo a y b el número de clases que presentan ambos caracteres, toma el valor en este caso,

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = 16'566.$$

Su p-valor será (a partir de la Tabla 4 de la Adenda)

$$\text{p-valor} = P\{\chi^2_1 > 16'566\} < 0'005.$$

Un p-valor tan bajo sugiere rechazar claramente la hipótesis nula de independencia de ambas poblaciones y concluir que sí hay relación entre hacer ejercicio y no padecer infarto.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 41206AD01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2012-2013.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los siguientes datos corresponden al número de horas reales trabajadas en un año por 20 enfermeras de un determinado hospital, es decir, descontadas vacaciones, días de baja, etc. y añadidas las horas extras.

1235 , 1925 , 1850 , 1500 , 2015 , 1925 , 1750 , 1967 , 925 , 1500

1714 , 955 , 1800 , 1645 , 1992 , 1985 , 1555 , 1956 , 1962 , 2015

Calcular, la Media aritmética, la Mediana, el Sexto Decil, el Recorrido, la Desviación Típica, el Coeficiente de Variación de Pearson y el Coeficiente de Asimetría de Fisher. Agrupar los datos en los 7 intervalos de amplitud 200: 800-1000, 1000-1200, ..., 2000-2200, obtener la distribución de frecuencias absolutas y el Histograma de estos datos agrupados.

Problema 2

El New York Post publicó el 30 de Agosto de 1955 una noticia en la que investigaba la hipótesis de que la posición en la línea de salida en una carrera de caballos influía en sus posibilidades de victoria, a pesar de que todos los caballos corrían la misma distancia y en línea recta. La posición de los ganadores en 144 carreras en el mismo hipódromo y en el mismo tipo de carrera que publicó el periódico fueron las siguientes,

Posición de salida	1	2	3	4	5	6	7	8
Número de victorias	29	19	18	25	17	10	15	11

¿Cree usted que la posición de salida influye en el número de victorias a un nivel de significación nivel $\alpha = 0'05$?

Problema 3

Los siguientes datos corresponden a niveles de salinidad (en partes por mil) de tres masas de agua separadas en la laguna de Bimini (Bahamas).

Zona I	**37'54**	**36'71**	**36'75**	**38'85**	**37'32**
Zona II	**40'17**	**40'81**	**39'38**	**39'79**	**39'70**
Zona III	**39'04**	**39'05**	**38'53**	**38'89**	**38'51**

Analizar si puede aceptarse la hipótesis nula de ser iguales los niveles de salinidad en las tres zonas.

Problema 1

Aunque el alumno habrá resuelto el ejercicio con la ayuda de una calculadora, los resultados deben de coincidir con los que obtenemos a continuación con R.

Primero incorporamos los datos y, a partir de `(1)` calculamos las medias de posición, dispersión etc.

```
> horas<-c(1235,1925,1850,1500,2015,1925,1750,1967,925,1500,1714,955,1800,
+ 1645,1992,1985,1555,1956,1962,2015)


> mean(horas)
[1] 1708.55                                                                  (1)
> median(horas)
[1] 1825                                                                     (1)
> quantile(horas,probs=6/10)
 60%
1925                                                                         (1)
> sort(horas)
 [1]  925  955 1235 1500 1500 1555 1645 1714 1750 1800 1850 1925 1925 1956
 [15] 1962 1967 1985 1992 2015 2015
> 2015-925
[1] 1090                                                                     (1)
> sqrt(19/20*var(horas))
[1] 330.2512                                                                 (1)
> 100*330.2512/mean(horas)
[1] 19.32933                                                                 (1)
> sum( ( horas - mean(horas) ) ^3 ) / (length(horas) *(sd(horas)^3))                 # asimetría a
[1] -1.10934                                                                 (1)     # la izquierda
```

La distribución de frecuencias de los datos agrupados en intervalos solicitada es

800-1000	2
1000-1200	0
1200-1400	1
1400-1600	3
1600-1800	3
1800-2000	9
2000-2200	2

Obsérvese que el dato 1800 debe de ir en el intervalo que comienza con este valor.

El histograma es el dado por la Figura 0.1. Se podría obtener directamente ejecutando

```
> hist(horas,main="Histograma de Horas",col=c(2,3,4,5),right=F)
```

en donde el último argumento es para formar a que los valores que coincidan con el extremo del intervalo, vayan en el siguiente.

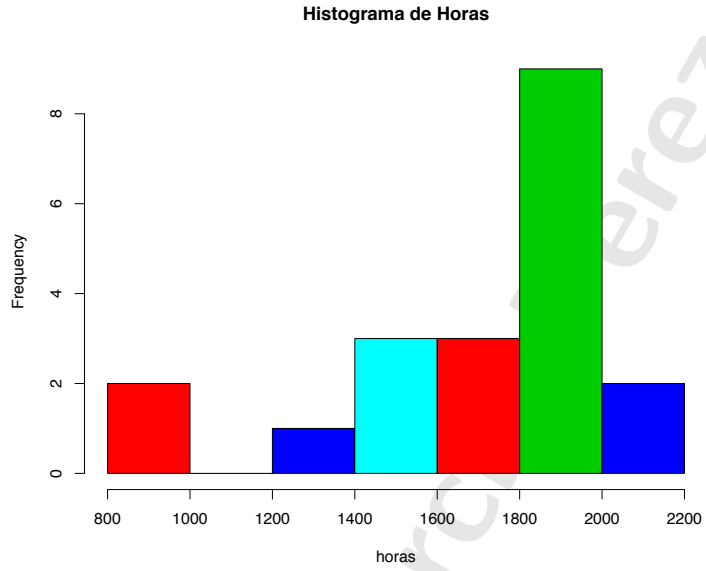


Figura 0.1 : Histograma

Problema 2

Se trataría de un test de bondad del ajuste de los 144 datos a una distribución uniforme sobre los 8 puntos, es decir, que asigna probabilidad 1/8 a cada una de las 8 posiciones y que, por tanto, asigna una frecuencia esperada de $144/8 = 18$ victorias en cada una de las 8 posiciones. Si esta discrepancia es significativa o no se analiza con el estadístico de Pearson ejecutando el test sobre la hipótesis nula de que el ajuste es válido, es decir, que no existen diferencias significativas entre los 8 lugares, frente a la hipótesis alternativa de que sí influye el lugar en las probabilidades de victoria. (Véase el Ejemplo 8.1 de EBR.)

El estadístico de contraste es

$$\lambda = \sum_{i=1}^{8} \frac{(n_i - n\,p_i)^2}{n\,p_i} = 16'33.$$

El p-valor del test es

$$P\{\chi^2_7 > 16'33\} < P\{\chi^2_7 > 14'07\} = 0'05$$

Como se ven en el p-valor dado en `(1)`, la decisión no está nada clara, pero a nivel $\alpha = 0'05$ se rechazaría la hipótesis nula de homogeneidad (por ser el p-valor menor que el nivel de significación) y concluiríamos que el lugar sí que influye en las oportunidades de victoria.

Aunque se podría platear la resolución de este ejercicio como un problema de Regresión Lineal, las suposiciones necesarias para que esta técnica se pudiera aplicar, no se verificarían, por lo que es más aconsejable un planteamiento como problema no paramétrico.

Problema 3

La comparación de más de dos poblaciones, en este caso tres, se ejecuta con un Análisis de la varianza para un factor y un diseño completamente aleatoriazado (EBR-sección 9.2) .

Para contrastar $\left\{ \begin{array}{ll} H_0: & \mu_I = \mu_{II} = \mu_{III} \\ H_1: & \textit{alguna distinta} \end{array} \right.$ la tabla de Análisis de la Varianza que se obtiene es

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Tratamientos	$SST_i = 16'113$	2	$8'0565$	$F = 21'458$
Residual	$SSE = 4'5$	12	$0'3755$	
Total	$SST = 20'613$	14		

Si fijamos un nivel de significación $\alpha = 0'01$, al ser $F = 21'458 > 8'5096 = F_{(2,12);0'005}$, se rechaza la igualdad de los tres tratamientos, H_0, además con un p-valor menor que $0'005$.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2012-2013.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los problemas que presentan los ordenadores se clasifican en problemas de hardware H, de software S, y otras razones O (como por ejemplo malas conexiones), habiéndose establecido que sus probabilidades de aparición son 0′1, 0′6 y 0′3 respectivamente.

Una empresa que fabrica ordenadores cree que, si hay un problema de hardware, sus ordenadores fallan con probabilidad 0′9; que si hay un problema de software, fallan con probabilidad 0′85 y, que si hay otras razones, fallan con probabilidad 0′4.

Si uno de esos ordenadores falla, ¿por cuál de las tres razones cree que sería?

Problema 2

En el artículo de Higham et al. (1980) aparecen las longitudes de las mandíbulas en milímetros, de 10 chacales (*Canis aureus*) macho y 10 hembra que reproducimos a continuación y que forman parte del Museo de Historia Natural del British Museum de Londres.

Macho	**120**	**107**	**110**	**116**	**114**	**111**	**113**	**117**	**114**	**112**
Hembra	**110**	**111**	**107**	**108**	**110**	**105**	**107**	**106**	**111**	**111**

¿Existen diferencias significativas entre las medias de las poblaciones, supuestamente normales, de donde proceden los datos?

Problema 3

Los siguientes datos son parte de los datos obtenidos por Wallach (1987) en donde midió el peso x en kilogramos y el consumo energético diario y en Megacalorías por día, de 8 ovejas Merinas de Australia

Peso x	**22′1**	**26′2**	**33′2**	**34′3**	**49**	**52′6**	**27′6**	**31**
Consumo y	**1′31**	**1′27**	**1′25**	**1′14**	**1′78**	**1′7**	**1′39**	**1′47**

Determine la recta de mínimos cuadrados (también llamada de regresión de y sobre x) de los datos anteriores y analice si es o no significativa a nivel $\alpha = 0'05$ mediante un test de hipótesis.

Problema 1

Del enunciado se desprende que es $P(H) = 0'1$, $P(S) = 0'6$ y $P(O) = 0'3$ y que es $P(F|H) = 0'9$, $P(F|S) = 0'85$ y $P(F|O) = 0'4$. Por el teorema de Bayes será

$$P(H|F) = \frac{P(F|H) \cdot P(H)}{P(F)} = \frac{P(F|H) \cdot P(H)}{P(F|H) \cdot P(H) + P(F|S) \cdot P(S) + P(F|O) \cdot P(O)} =$$

$$= \frac{0'9 \cdot 0'1}{0'9 \cdot 0'1 + 0'85 \cdot 0'6 + 0'4 \cdot 0'3} = \frac{0'09}{0'09 + 0'51 + 0'12} = \frac{0'09}{0'72} = 0'125.$$

Análogamente,

$$P(S|F) = \frac{P(F|S) \cdot P(S)}{P(F)} = \frac{0'85 \cdot 0'6}{0'72} = \frac{0'51}{0'72} = 0'7083.$$

$$P(O|F) = \frac{P(F|O) \cdot P(O)}{P(F)} = \frac{0'4 \cdot 0'3}{0'72} = \frac{0'12}{0'72} = 0'1667.$$

Con lo que, lo más probable es que sea por causa del software.

Problema 2

La igualdad de las medias de ambas poblaciones de donde proceden los datos (es decir, la hipótesis de que no hay dimorfismo sexual) se puede contrastar mediante un test para la igualdad de las medias de dos poblaciones normales independientes de varianzas desconocidas y muestras pequeñas (EBR-sección 7.6) para lo que primero debemos averiguar si pueden admitirse como iguales o no las varianzas poblacionales. Es decir, contrastar $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$. Al ser

$$I = \left[F_{n_1-1,n_2-1;1-\alpha/2} \, , \, F_{n_1-1,n_2-1;\alpha/2} \right] = \left[F_{9,9;0'975} \, , \, F_{9,9;0'025} \right] =$$

$$= \left[1/F_{9,9;0'025} \, , \, F_{9,9;0'025} \right] = [0'248 \, , \, 4'026]$$

y ser $S_1^2/S_2^2 = 13'8/5'16 = 2'67 \in I$ aceptaremos la hipótesis nula de igualdad de las varianzas. Hemos elegido un nivel de significación $\alpha = 0'05$ pero con otros se hubiera obtenido igual resultado.

Por tanto, el estadístico del test óptimo a utilizar, para contrastar la igualdad de medias, será

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|113'4 - 108'6|}{\sqrt{\dfrac{9\cdot 13'8 + 9\cdot 5'16}{18}}\sqrt{\dfrac{1}{10}+\dfrac{1}{10}}} = 3'486.$$

El p-valor es

$$\text{p-valor} = 2\cdot P\{t_{18} > 3'486\} < 2\cdot 0'0025 = 0'005$$

con lo que se rechaza la igualdad de medias y se concluye con que sí hay dimorfismo sexual en esta especie.

Problema 3

La recta de regresión (EBR-sección 10.2) o mínimos cuadrados (EBR-sección 2.4.2) para los datos del enunciado resulta ser igual a

$$y = 0'85955 + 0'01606\, x$$

Analizar si esta recta es significativa o no, es ejecutar el contraste de la regresión lineal simple (EBR-sección 10.3), para lo que debemos construir la tabla de Análisis de la Varianza (EBR página 301 y ADD página 26) que en nuestro problema toma el valor

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal	$SSEX = 0'21214$	1	$0'21214$	$F = 9'0368$
Residual	$SSNEX = 0'14085$	6	$0'023475$	
Total	$SST = 0'35299$	7		

Ahora el test que se plantea es H_0 : *La regresión no es significativa*, o lo que es lo mismo, H_0 : *La recta de regresión no me sirve para explicar la variable y en función de la x*, frente a la hipótesis alternativa, H_1 : *La regresión es significativa*. Puesto que el estadístico del contraste sigue una F de Snedecor con $(1, 6)$ grados de libertad, el punto crítico es, utilizando la Tabla 6 de la distribución F de Snedecor (ADD, página 36) $F_{(1,6);0'05} = 5'9874 < 9'0368 = F$, por lo que debemos rechazar la hipótesis nula y concluir que el ajuste es válido.

Alternativamente se podía haber contrastado la hipótesis nula de que el coeficiente de regresión de la variable x es cero frente a la alternativa de que no es cero. En este caso, el estadístico de contraste hubiera tomado el valor $3'006$ y también se hubiera rechazado la hipótesis nula.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2012-2013.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Un alumno quiere modelizar el número de veces X que ha suspendido una asignatura antes de aprobarla, suponiendo que todas las veces que se presenta tiene la misma probabilidad de aprobarla, $0'5$ y además esta probabilidad es independiente cada vez que se presenta al examen, el cual es tipo test.

Determinar la función de masa de X, su media, su desviación típica y calcular la probabilidad de que suspenda más de 5 veces.

Problema 2

Los siguientes datos corresponden a un estudio sobre trombosis (van Oost et al., 1983) en donde se midió la cantidad de tromboglobulina urinaria eliminada por 12 pacientes normales y 12 pacientes con diabetes.

Normales:
$4'1$, $6'3$, $7'8$, $8'5$, $8'9$, $10'4$, $11'5$, 12 , $13'8$, $17'6$, $24'3$, $37'2$

Diabéticos:
$11'5$, $12'1$, $16'1$, $17'8$, 24 , $28'8$, $33'9$, $40'7$, $51'3$, $56'2$, $61'7$, $69'2$

Supuesto que ambos grupos de datos proceden de distribuciones normales, ¿puede aceptarse la igualdad de las medias de ambas poblaciones a nivel $0'05$?

Problema 3

En un artículo de Agresti (1989) aparece una tabla con datos sobre un cierto deterioro mental en niños de edad escolar y el estatus socioeconómico de su familia, siendo 1 el nivel más bajo de este

estatus y 6 el más alto. Los datos de dicho trabajo son los de la siguiente tabla:

	Estatus socioeconómico padres					
	1	**2**	**3**	**4**	**5**	**6**
Situación Mental						
Buena	**64**	**57**	**57**	**72**	**36**	**21**
Deterioro mental leve	**94**	**94**	**105**	**141**	**97**	**71**
Deterioro mental moderado	**58**	**54**	**65**	**77**	**54**	**54**
Deterioro mental grave	**46**	**40**	**60**	**94**	**78**	**71**

Analizar la posible independencia entre ambas variables.

Problema 1

Si denominamos *éxito* a aprobar la asignatura, la distribución que modeliza esta variable aleatoria, *número de fallos antes del primer éxito* es la geométrica, en este caso, de parámetro $p = 0'5$ (EBR-sección 4.4.3). Su función de masa será, por tanto,

$$p_X(x) = 0'5^x\, 0'5 = 0'5^{x+1}$$

$x = 0, 1, 2, ...$

Su media y desviación típica son, respectivamente,

$$E[X] = \frac{1-p}{p} = \frac{0'5}{0'5} = 1$$

$$D[X] = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{0'5}{0'5^2}} = \frac{1}{\sqrt{0'5}} = 1'414.$$

La probabilidad pedida de que suspenda más de cinco veces será

$$P\{X > 5\} = 1 - P\{X \leq 5\} = 1 - (P\{X = 0\} + P\{X = 1\} + P\{X = 2\} +$$

$$+P\{X = 3\} + P\{X = 4\} + P\{X = 5\}) =$$

$$1 - (p_X(0) + p_X(1) + p_X(2) + p_X(3) + p_X(4) + p_X(5)) =$$

$$= 1 - (0'5 + 0'5^2 + 0'5^3 + 0'5^4 + 0'5^5 + 0'5^6) = 1 - \frac{0'5^7 - 0'5}{0'5 - 1} = 0'5^6 = 0'015625$$

por ser la suma anterior, la suma de los 6 primeros términos de una progresión geométrica de razón $0'5$ y ser la suma de los n primeros términos de una progresión geométrica de razón r igual a

$$\frac{a_n\, r - a_1}{r - 1}.$$

Problema 2

Se trataría de la comparación de medias de dos poblaciones normales independientes y muestras pequeñas, siendo las varianzas poblacionales desconocidas (EBR-sección 7.6), para lo que necesitamos primero analizar si éstas pueden considerase iguales. Para ello contrastamos la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ frente a la $H_0 : \sigma_1^2 \neq \sigma_2^2$ (EBR-sección 7.5), aceptándose la hipótesis nula si

$$\frac{S_1^2}{S_2^2} \in \left[F_{n_1-1,n_2-1;1-\alpha/2}\ ,\ F_{n_1-1,n_2-1;\alpha/2}\right].$$

A partir de nuestros datos obtenemos que es

$$\begin{array}{lll} \overline{x}_1 = 13'533 & S_1^2 = 84'54 & n_1 = 12 \\ \overline{x}_2 = 35'275 & S_2^2 = 410'87 & n_2 = 12 \end{array}$$

Como no vienen los grados de libertad $(11, 11)$ en las tablas de la F de Snedecor, se pueden promediar, quedando

$$F_{n_1-1,n_2-1;1-\alpha/2} = F_{11,11;0'975} = \frac{1}{F_{11,11;0'025}} = \frac{1}{3'47765} = 0'28755$$

utilizando las propiedades de la distribución F de Snedecor y la Tabla 6 de esta distribución. Además,

$$F_{n_1-1,n_2-1;\alpha/2} = F_{11,11;0'025} = 3'47765$$

con lo que la región de aceptación será el intervalo $[0'28755\,,\,3'47765]$.

Al ser el estadístico de contraste igual a

$$\frac{S_1^2}{S_2^2} = \frac{84'54}{410'87} = 0'21 \notin [0'28755\,,\,3'47765]$$

no aceptaremos la hipótesis nula, concluyendo con que es razonable admitir como distintas las varianzas de las poblaciones normales.

Supuestas distintas las varianzas poblaciones, la hipótesis nula de igualdad de las medias de ambos grupos, $H_0\colon \mu_1 = \mu_2$ se aceptará cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \leq t_{f;\alpha/2}$$

en donde los grados de libertad f de la t de Student se determinan mediante la aproximación de Welch, siendo éste el entero más próximo a

$$\frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1+1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2 = \frac{\left(\dfrac{84'54}{12} + \dfrac{410'87}{12}\right)^2}{\dfrac{\left(\dfrac{84'54}{12}\right)^2}{13} + \dfrac{\left(\dfrac{410'87}{12}\right)^2}{13}} - 2 = 16'13$$

con lo que tomaremos $f = 16$, siendo, por la Tabla 5 de la t de Student, el punto crítico igual a $t_{16;\alpha/2} = t_{16;0'025} = 2'12$. Como el estadístico es igual a

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|13'533 - 35'275|}{\sqrt{\frac{84'54}{12} + \frac{410'87}{12}}} = 3'384 > 2'12 = t_{f;\alpha/2}$$

se rechazará la hipótesis nula de igualdad (en promedio) de la media de los dos grupos.

El p-valor de este último test es

$$2 \cdot P\{t_{16} > 3'384\} < 2 \cdot 0'0025 = 0'005$$

lo suficientemente pequeño como para confirmar el rechazo de la hipótesis nula.

Problema 3

Se trata de un contraste de independencia de caracteres de las dos variables que forman la tabla (EBR-sección 8.2.4).

El estadístico de Pearson es

$$\lambda = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n}$$

siendo r el número de filas de la tabla anterior, s el número de columnas de la tabla, n_{ij} las frecuencias observadas de cada celdilla de la tabla, n_i el total marginal de la fila i-ésima y m_j el total marginal de la columna j-ésima.

El valor del estadístico de Pearson es

$$\lambda = 45'9853$$

y el p-valor, a partir de la Tabla 4,

$$P\{\chi^2_{15} > 45'98\} < 0'005.$$

Un p-valor tan bajo sugiere rechazar claramente la hipótesis nula de independencia de ambas poblaciones y concluir que sí hay relación entre ambas variables en la población estudiada.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 41206AD01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2013-2014

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

En el mes de Abril se suelen solapar las alergias al polen de las Cupresáceas y las Gramíneas aunque no en las mismas personas, de manera que, en ese mes, un 15 % de la población española es alérgica a las Cupresáceas, un 15 % a las Gramíneas, un 5 % a otras plantas y un 65 % no presenta síntomas de alergia al polen, perteneciendo las personas a sólo una de estas clases.

Si en ese mes se eligen 8 personas al azar en España, ¿cuál es la probabilidad de que haya menos de 2 que presenten síntomas de alergia a Cupresáceas? ¿Cuál de que haya más de 5 con alergia al polen?

Y si se eligen 1000 personas, ¿cuál es la probabilidad de que menos de 120 sean alérgicas a las Gramíneas?

Problema 2

Se quiere analizar si el parásito Plasmodium falciparum, una de las especies que causa malaria en humanos y transmitido por el mosquito Anopheles, ataca los glóbulos rojos preferentemente a personas ya infectadas con otras enfermedades infecciosas o lo hace al azar. Si la distribución del número de parásitos por eritrocito sigue una distribución de Poisson, podemos concluir que las infecciones múltiples son debidas al azar. Para analizar este problema, Wang (1970) obtuvo los siguientes datos:

Número de parásitos por eritrocito	0	1	2	3	4	≥ 5
Frecuencia observada	40000	8621	1259	99	21	0

¿Qué conclusiones obtendría?

Problema 3

Se cree que el porcentaje de pureza de oxígeno Y en un proceso de destilación químico, depende del porcentaje de hidrocarbonos X presentes en el condensador principal de la unidad de destilación. En el análisis de esta idea se obtuvieron los siguientes datos:

X	1′15	1′30	1′48	1′35	1′23	1′54	1′40
Y	91′40	93′81	97′01	94′40	91′70	99′40	93′60

Analizar si puede aceptarse esta idea mediante un Análisis de Regresión Lineal.

En la resolución de los problemas de esta semana y de la siguiente, aparece en ocasiones su resolución con R. Lógicamente, en el examen sólo puede utilizar una calculadora. Se incluye esta posibilidad de cara a cursos futuros en donde el alumno puede utilizar estas soluciones en su formación.

Problema 1

Los tres apartados del problema se pueden formalizar mediante un modelo binomial (EBR-sección 4.4.1) en donde cada prueba de Bernoulli sea analizar si el individuo elegido posee o no un tipo concreto de alergia.

a) En este apartado el suceso *éxito* es ser alérgico a Cupresáceas, de manera que la variable *número de personas, de entre los ocho, que son alérgicos a Cupresáceas*, se puede modelizar mediante una variable X con distribución binomial $B(8, 0'15)$, al ser $p = 0'15$ la probabilidad de que se dé el suceso *éxito*.

La probabilidad pedida será ahora, utilizando la Tabla 1 de la distribución binomial (ADD, página 30)

$$P\{X < 2\} = P\{X = 0\} + P\{X = 1\} = 0'2725 + 0'3847 = 0'6572.$$

b) Ahora el suceso *éxito* es ser alérgico al polen. Como, según el enunciado, una persona es alérgica al polen con probabilidad $0'35$, si X es una variable con distribución binomial $B(8, 0'35)$, la probabilidad pedida será ahora,

$$P\{X > 5\} = P\{X = 6\} + P\{X = 7\} + P\{X = 8\}$$

pero esta probabilidad de éxito $0'35$ no viene en las tablas por lo que las calcularemos mediante la función de masa de la binomial:

$$P\{X = 6\} = \binom{8}{6} 0'35^6\, 0'65^2 = 28\, 0'35^6\, 0'65^2 = 0'02174668.$$

Análogamente,

$$P\{X = 7\} = 0'003345643$$

$$P\{X = 8\} = 0'0002251875$$

con lo que la probabilidad pedida será

$$P\{X > 5\} = P\{X = 6\} + P\{X = 7\} + P\{X = 8\} = 0'02531751.$$

c) Ahora el suceso *éxito* es ser alérgico a las Gramíneas. Como, según el enunciado, una persona es alérgica a las Gramíneas con probabilidad $0'15$, si X es

una variable con distribución binomial $B(1000, 0'15)$, la probabilidad pedida será ahora,

$$P\{X < 120\}$$

Si se aumenta el número de *pruebas de Bernoulli*, modelizándose el problema con una variable $X \rightsquigarrow B(1000, 0'15)$, el cálculo de probabilidades de distribuciones binomiales para un gran número de ensayos, como aquí ocurre, se realiza aproximando dicha distribución mediante el *teorema central del límite* (EBR-sección 4.7).

En el caso de una distribución binomial $X \rightsquigarrow B(n,p)$, su aproximación mediante una normal

$$N(np\,,\,\sqrt{np(1-p)}\,)$$

es válida (EBR-sección 4.7) cuando, supuesto sea $p \leq 0'5$ (como aquí ocurre) entonces sea también $np > 5$ (como aquí ocurre).

Por tanto, aproximaremos la $X \rightsquigarrow B(1000, 0'15)$, por una

$$N\left(1000 \cdot 0'15\,,\,\sqrt{1000 \cdot 0'15 \cdot 0'85}\right) = N(150\,,\,11'2916)$$

quedando la probabilidad pedida igual a

$$P\{X < 120\} = P\left\{\frac{X - 150}{11'2916} < \frac{120 - 150}{11'2916}\right\} \simeq P\{Z < -2'66\} = P\{Z > 2'66\}$$

siendo Z una variable aleatoria $N(0,1)$.Por la Tabla 3 de la Addenda, página 33, tenemos que es

$$P\{X < 120\} = P\{Z > 2'66\} = 0'0039$$

es decir, muy baja.

Problema 2

Se trataría de hacer un test de bondad del ajuste (EBR-sección 8.2.2) de los datos a una distribución de Poisson. De los datos observados se tiene que la media muestral es

$$\overline{x} = \frac{0 \cdot 40000 + 1 \cdot 8621 + 2 \cdot 1259 + 3 \cdot 99 + 4 \cdot 21 + 5 \cdot 0}{50000} = \frac{11520}{50000} = 0'23$$

por lo que el análisis de bondad del ajuste debe de hacerse a la Poisson $\mathcal{P}(0'23)$.

El estadístico de contraste es

$$\lambda = \sum_{i=1}^{5} \frac{(n_i - n\,p_i)^2}{n\,p_i}.$$

Como el valor del parámetro $\lambda = 0'23$ no aparece en las tablas, las probabilidades teóricas p_i debemos calcularlas a través de su función de masa (EBR-sección 4.4.2)

$$p_i = \frac{e^{-0'23} \cdot 0'23^i}{i!}$$

con la que podemos construir la siguiente tabla en donde se han juntado las últimas filas (como la de ≥ 5) para que las frecuencias esperadas fueran mayores que 5.

X_i	n_i	p_i	$n \cdot p_i$
0	40000	0'7945	39725
1	8621	0'1827	9135
2	1259	0'0210	1050
3	99	0'0016	80
≥ 4	21	0'0002	10
	50000	1	50000

El estadístico de contraste toma el valor

$$\lambda = \sum_{i=1}^{5} \frac{(n_i - n\,p_i)^2}{n\,p_i} = 89'04$$

y el p-valor del test es

$$P\{\chi^2_3 > 89'04\} < 0'005$$

suficientemente pequeño como para rechazar la hipótesis nula y concluir que el ajuste no es adecuado, o bien, que las infecciones múltiples no son debidas al azar.

Con R hubiéramos ejecutado

```
x<-c(40000,8621,1259,99,21)
p1<-c(dpois(0,0.23),dpois(1,0.23),dpois(2,0.23),dpois(3,0.23),1-ppois(3,0.23))
chisq.test(x,p=p1)
```

No saliendo exactamente lo mismo por los redondeos.

Problema 3

Para hacer el Análisis de Regresión Lineal de estos datos, deberemos ajustar una recta a los datos del enunciado y contrastar si puede admitirse o no

la hipótesis nula de ser cero el coeficiente de regresión de la variable independiente.

La recta de regresión (EBR-sección 10.2) o mínimos cuadrados (EBR-sección 2.4.2) para los datos del enunciado resulta ser igual a

$$y = 68'017 + 19'598\, x$$

Analizar si esta recta es significativa o no, es ejecutar el contraste de la regresión lineal simple (EBR-sección 10.3), para lo que debemos construir la tabla de Análisis de la Varianza (EBR página 301 y ADD página 26) que en nuestro problema toma el valor

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal	$SSEX = 43'17018$	1	$43'17018$	$F = 36'7$
Residual	$SSNEX = 5'88139$	5	$1'176278$	
Total	$SST = 49'05157$	6		

Ahora el test que se plantea es H_0 : *La regresión no es significativa*, o lo que es lo mismo, H_0 : *La recta de regresión no me sirve para explicar la variable y en función de la x*, frente a la hipótesis alternativa, H_1 : *La regresión es significativa*. Puesto que el estadístico del contraste sigue una F de Snedecor con $(1, 5)$ grados de libertad, el punto crítico es, utilizando la Tabla 6 de la distribución F de Snedecor (ADD, página 36) $F_{(1,5);0'05} = 6'6079 < 36'7 = F$, por lo que debemos rechazar la hipótesis nula y concluir que el ajuste es válido.

Alternativamente se podía haber contrastado la hipótesis nula de que el coeficiente de regresión de la variable x es cero frente a la alternativa de que no es cero. En este caso, el estadístico de contraste hubiera tomado el valor $6'058$ y también se hubiera rechazado la hipótesis nula.

Tanto en uno u otro caso, el p-valor del test es muy pequeño. En concreto

$$P\{F_{(1,5)} > 36'7\} = 2 \cdot P\{t_5 > 6'058\} = 0'00177$$

suficientemente pequeño como para confirmar el rechazo de la hipótesis nula y concluir que sí es cierto que la pureza del oxígeno Y depende linealmente del porcentaje de hidrocarbonos.

Con R hubiéramos ejecutado

```
> x<-c(1.15,1.30,1.48,1.35,1.23,1.54,1.40)
> y<-c(91.40,93.81,97.01,94.40,91.70,99.40,93.60)
> recta<-lm(y~x)
recta

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      68.02        19.60

> summary(recta)
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5        6        7
 0.84529  0.31561 -0.01201 -0.07429 -0.42254  1.20212 -1.85418

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   68.017      4.386  15.506 2.03e-05 ***
x             19.598      3.235   6.058  0.00177 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.085 on 5 degrees of freedom
Multiple R-squared: 0.8801,     Adjusted R-squared: 0.8561
F-statistic:  36.7 on 1 and 5 DF,  p-value: 0.001768

> aov(recta)
Call:
   aov(formula = recta)

Terms:
                       x Residuals
Sum of Squares  43.17018   5.88139
Deg. of Freedom        1         5

Residual standard error: 1.084564
Estimated effects may be unbalanced
```

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2013-2014

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

En una empresa de 800 trabajadores, un 30 % tiene un marcador genético en el cromosoma, que indica un riesgo elevado de hipertensión arterial.

Si se eligen al azar a 10 empleados distintos de la empresa y se les hace un análisis de este marcador genético, ¿cuál es la probabilidad de que exactamente 1 lo tenga? ¿Y de que de esos 10 haya más de 1 con este marcador?

Problema 2

En un artículo de la revista Materials Engineering (1989) aparecen los resultados de unas pruebas de adherencia a la tracción de 10 muestras de la aleación U-700. La Carga (en megapascales) a la que se produjo la rotura de la muestra fueron las siguientes:

19′8 15′4 11′4 19′5 10′1 18′5 14′1 8′8 14′9 7′9

Suponiendo que la variable Carga de ruptura sigue una distribución normal, determinar un intervalo de confianza, de coeficiente de confianza 0′95 para la media de esta variable.

Problema 3

En un artículo del Journal of the American Ceramic Society (1991) se considera que el porcentaje Y de porosidad de la circonita parcialmente estabilizada, es función de la temperatura en grados centígrados X. Para ello, la publicación aporta los siguientes datos:

X	1100	1200	1300	1100	1500	1200	1300
Y	30′8	19′2	6′0	13′5	11′4	7′7	3′6

Analizar si puede aceptarse esta idea mediante un Análisis de Regresión Lineal.

Problema 1

El problema se puede modelizar mediante una distribución hipergeométrica (EBR-sección 4.4.4) de función de masa

$$p_X(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad \text{máx}\{0, n-N+D\} \leq x \leq \text{mín}\{n, D\}.$$

en donde es $N = 800$, $D = 0'3 \cdot 800 = 240$, $n = 10$ y el $x = 1$ el punto en donde calcular la probabilidad pedida, la cual será

$$p_X(1) = \frac{\binom{240}{1}\binom{560}{9}}{\binom{800}{10}} = \frac{240 \cdot 560 \cdots 552 \cdot 10}{800 \cdots 791} = 0'12007944.$$

Con R hubiera sido más fácil, ejecutando

```
> dhyper(1,240,800-240,10)
[1] 0.1200794
```

En la segunda parte del problema debemos de calcular

$$P\{X > 1\} = 1 - P\{X \leq 1\} = 1 - P\{X = 0\} - P\{X = 1\} = 1 - 0'02756824 - 0'12007944 = 0'85235232$$

y con R

```
> 1-dhyper(0,240,800-240,10)-dhyper(1,240,800-240,10)
[1] 0.8523523
```

Problema 2

Se trata de determinar un intervalo de confianza para la media de una población supuestamente normal de varianza desconocida (EBR-sección 6.2)

$$\left[\overline{x} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \, , \, \overline{x} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \right]$$

De los datos del enunciado se obtiene que es $\overline{x} = 14'04$, $S = 4'385886$ y de la Tabla 5 de t de Student, $t_{n-1;\alpha/2} = t_{9;0'05/2} = 2'262$ con lo que el intervalo solicitado será

$$\left[\ 14'04 - 2'262\ \frac{4'385886}{\sqrt{10}}\ ,\ 14'04 + 2'262\ \frac{4'385886}{\sqrt{10}}\ \right] = [\,10'903\ ,\ 17'177\,]$$

y con R

```
> t.test(x)

        One Sample t-test

data:  x
t = 10.123, df = 9, p-value = 3.232e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 10.90253 17.17747
sample estimates:
mean of x
    14.04
```

Problema 3

Para hacer el Análisis de Regresión Lineal de estos datos, deberemos ajustar una recta a los datos del enunciado y contrastar si puede admitirse o no la hipótesis nula de ser cero el coeficiente de regresión de la variable independiente.

La recta de regresión (EBR-sección 10.2) o mínimos cuadrados (EBR-sección 2.4.2) para los datos del enunciado resulta ser igual a

$$y = 55'62561 - 0'03416\,x$$

Analizar si esta recta es significativa o no, es ejecutar el contraste de la regresión lineal simple (EBR-sección 10.3), para lo que debemos construir la tabla de Análisis de la Varianza (EBR página 301 y ADD página 26) que en nuestro problema toma el valor

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal	$SSEX = 136'6829$	1	$136'6829$	$F = 1'767522$
Residual	$SSNEX = 386'6513$	5	$77'33026$	
Total	$SST = 523'3342$	6		

Ahora el test que se plantea es H_0 : *La regresión no es significativa*, o lo que es lo mismo, H_0 : *La recta de regresión no me sirve para explicar la variable y en función de la x*, frente a la hipótesis alternativa, H_1 : *La regresión es significativa*. Puesto que el estadístico del contraste sigue una F de Snedecor con $(1, 5)$ grados de libertad, el punto crítico es, utilizando la Tabla 6 de la distribución F de Snedecor (ADD, página 36) $F_{(1,5);0'05} = 6'6079 > 1'768 = F$, por lo que debemos aceptar la hipótesis nula y concluir que el ajuste no es significativo.

Alternativamente se podía haber contrastado la hipótesis nula de que el coeficiente de regresión de la variable x es cero frente a la alternativa de que no es cero. En este caso, el estadístico de contraste hubiera tomado el valor $1'329$ y también se hubiera aceptado la hipótesis nula.

Tanto en uno u otro caso, el p-valor del test es suficientemente grande. En concreto

$$P\{F_{(1,5)} > 1'768\} = 2 \cdot P\{t_5 > 1'329\} = 0'241$$

suficientemente grande como para confirmar la aceptación de la hipótesis nula y concluir que no es cierto que la porosidad de la circonita Y dependa linealmente de la temperatura.

Con R hubiéramos ejecutado

```
> x<-c(1100,1200,1300,1100,1500,1200,1300)
> y<-c(30.8,19.2,6,13.5,11.4,7.7,3.6)
> recta<-lm(y~x)
recta

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   55.62561     -0.03416

> summary(recta)
Call:
lm(formula = y ~ x)

Residuals:
     1      2      3      4      5      6      7
12.749  4.565 -5.220 -4.551  7.012 -6.935 -7.620

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.62561   32.10539   1.733    0.144
```

```
x           -0.03416    0.02569  -1.329    0.241

Residual standard error: 8.794 on 5 degrees of freedom
Multiple R-squared: 0.2612,     Adjusted R-squared: 0.1134
F-statistic: 1.768 on 1 and 5 DF,  p-value: 0.2411

> aov(recta)
Call:
   aov(formula = recta)

Terms:
                       x Residuals
Sum of Squares  136.6829  386.6513
Deg. of Freedom        1         5

Residual standard error: 8.793763
Estimated effects may be unbalanced
```

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2013-2014

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Si X es una variable aleatoria con distribución normal de media 1 y desviación típica 3, determinar de forma razonada las probabilidades $P\{|X| < 4\}$ y $P\{|3 - X| < 4\}$.

Problema 2

Youden y Beale (1934) hicieron un estudio para analizar si dos tipos de virus producen los mismos efectos en las hojas de tabaco. Para ello eligieron al azar 8 hojas de tabaco que dividieron en dos mitades impregnando cada mitad con uno y sólo uno de los dos preparados víricos. De esta forma, cada hoja es tratada con ambos tipos de virus.

Aparecieron de esta forma al cabo de un tiempo, en las 8 mitades de hoja, el siguiente número de lesiones, consistentes en pequeños anillos oscuros:

Hoja número	1	2	3	4	5	6	7	8
Preparado 1	31	20	18	17	9	8	10	7
Preparado 2	18	17	14	11	10	7	5	6

A nivel de significación $\alpha = 0'05$, ¿producen ambos preparados los mismos efectos en las hojas? Suponer, primero, que los datos proceden de poblaciones normales y, después, utilizar el test de los rangos signados de Wilcoxon.

Problema 3

Los siguientes datos (Pearson y Lee, 1902-3) corresponden a las estaturas (en pulgadas) de hermanos y hermanas de una misma familia.

Familia	1	2	3	4	5	6	7	8	9	10	11
Hermano	71	68	66	67	70	71	70	73	72	65	66
Hermana	69	64	65	63	65	62	65	64	66	59	62

Calcule el coeficiente de correlación lineal de Pearson muestral y, en base a su valor, contraste a nivel de significación $\alpha = 0'05$, si existe correlación entre las estaturas de las hermanas y hermanos de una misma familia; es decir, si el coeficiente de correlación poblacional es significativamente distinto de cero.

Problema 1

Este problema es muy parecido al Problema 2.11 de PREB. En el cálculo de las probabilidades pedidas vamos a hacer uso de las tablas de la normal $Z \rightsquigarrow N(0,1)$, para lo que deberemos tipificar la variable dada X, restándole su media 1 y dividiendo por su desviación típica 3.

a)

$$\begin{aligned} P\{|X|<4\} &= P\{-4<X<4\} \\ &= P\left\{\frac{-4-1}{3}<Z<\frac{4-1}{3}\right\} \\ &= P\left\{-1'67<Z<1\right\} \\ &= P\left\{Z<1\right\}-P\left\{Z<-1'67\right\} \\ &= 1-P\left\{Z>1\right\}-P\left\{Z>1'67\right\}. \end{aligned}$$

Utilizando ahora la Tabla 3 de la normal $N(0,1)$, será en definitiva,

$$P\{|X|<4\}=1-P\left\{Z>1\right\}-P\left\{Z>1'67\right\}=1-0'1587-0'0475=0'7938.$$

b)

$$\begin{aligned} P\{|3-X|<4\} &= P\{|X-3|<4\} \\ &= P\left\{-4<X-3<4\right\} \\ &= P\left\{3-4<X<3+4\right\} \\ &= P\left\{-1<X<7\right\} \\ &= P\left\{\frac{-1-1}{3}<Z<\frac{7-1}{3}\right\} \\ &= P\left\{-0'67<Z<2\right\} \end{aligned}$$

$$
\begin{aligned}
&= P\{Z < 2\} - P\left\{Z < -0'67\right\} \\
&= 1 - P\{Z > 2\} - P\left\{Z > 0'67\right\} \\
&= 1 - 0'0228 - 0'2514 \\
&= 0'7258.
\end{aligned}
$$

Problema 2

Está claro que los datos son dependientes al obtenerse los pares de observaciones en la misma hoja. Por tanto, lo primero es calcular la variable diferencia D que tomará los valores

$$13\ ,\ 3\ ,\ 4\ ,\ 6\ ,\ -1\ ,\ 1\ ,\ 5\ ,\ 1$$

El propósito es analizar si puede admitirse la hipótesis nula de ser cero su media o mediana.

a) Si los datos procedieran de distribuciones normales, entonces D serían datos procedentes de una distribución normal y el test que tendríamos que hacer es el de la hipótesis nula $H_0 : \mu_D = 0$ frente a la alternativa $H_0 : \mu_D \neq 0$ para este caso de varianza desconocida (EBR-sección 7.2) rechazando la hipótesis nula cuando y sólo cuando sea

$$\frac{|\overline{x} - \mu_0|}{S/\sqrt{n}} > t_{n-1;\alpha/2}$$

En este caso será

$$\frac{|\overline{x} - \mu_0|}{S/\sqrt{n}} = \frac{4}{4'31/\sqrt{8}} = 2'625$$

El p-valor de este test será $2 \cdot P\{t_7 > 2'625\}$ que estará entre los valores

$$2 \cdot 0'01 < 2 \cdot P\{t_7 > 2'62\} < 2 \cdot 0'025$$

Es decir, entre $0'02$ y $0'05$. Para un nivel de significación $\alpha = 0'05$ se rechazaría la hipótesis nula, concluyendo que ambos tipos de virus no producen el mismo efecto.

Si queremos utilizar R ejecutaríamos

```
> Preparado1<-c(31,20,18,17,9,8,10,7)
> Preparado2<-c(18,17,14,11,10,7,5,6)
> D<-Preparado1-Preparado2
> t.test(D)

        One Sample t-test

data:  D
t = 2.6253, df = 7, p-value = 0.03414
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3972029 7.6027971
sample estimates:
mean of x
        4
```

obteniendo, lógicamente, los mismos resultados.

Si queremos hacer un test de rangos signados de Wilcoxon (EBR-sección 8.3.2) las hipótesis a contrastar serían $H_0 : M_D = 0$ frente a la alternativa $H_0 : M_D \neq 0$. Como la hipótesis nula es 0, las diferencias con respecto a esa hipótesis nula serán simplemente los valores D_i anteriores

$$13\ ,\ 3\ ,\ 4\ ,\ 6\ ,\ -1\ ,\ 1\ ,\ 5\ ,\ 1$$

Sus valores absolutos serán

$$13\ ,\ 3\ ,\ 4\ ,\ 6\ ,\ 1\ ,\ 1\ ,\ 5\ ,\ 1$$

y sus rangos

$$8\ ,\ 4\ ,\ 5\ ,\ 7\ ,\ 2\ ,\ 2\ ,\ 6\ ,\ 2$$

La suma de los rangos de las diferencias positivas será

$$T^+ = 8 + 4 + 5 + 7 + 2 + 6 + 2 = 34$$

Si $\alpha = 0'05$, a partir de la Tabla 14, será $P\{T_8^+ > 30\} = 0'05$, y el valor 34 caerá en la región crítica rechazándose la hipótesis nula.

Con R ejecutaríamos

```
> wilcox.test(D)

        Wilcoxon signed rank test with continuity correction

data:  D
V = 34, p-value = 0.02917
alternative hypothesis: true location is not equal to 0

Mensajes de aviso perdidos
In wilcox.test.default(D) : cannot compute exact p-value with ties
```

obteniendo, lógicamente, los mismos resultados.

Problema 3

El coeficiente de correlación lineal muestral (EBR-sección 2.4.3 y EBR-sección 10.5.1) es

$$r = \frac{n\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} = 0'558.$$

cuyo valor se calcularía más fácilmente con R ejecutando

```
> hermanos<-c(71,68,66,67,70,71,70,73,72,65,66)
> hermanas<-c(69,64,65,63,65,62,65,64,66,59,62)
> cor(hermanos,hermanas)
[1] 0.5580547
```

Ahora el propósito es contrastar la hipótesis nula (EBR-sección 10.5.2) $H_0 : \rho = 0$ frente a la alternativa $H_1 : \rho \neq 0$, utilizando el estadístico de contraste,

$$t = r\,\sqrt{\frac{n-2}{1-r^2}}$$

que, si la hipótesis nula es cierta, sigue una distribución t de Student con $n-2$ grados de libertad. Para los datos del problema, toma el valor

$$t = 0'558\,\sqrt{\frac{9}{0'6886}} = 2'017.$$

Como el test es bilateral, el punto crítico será, a partir de a Tabla 5 de la distribución t de Student, $t_{n-2;\alpha/2} = t_{9;0'05/2} = 2'262 > 2'017$, con lo que se aceptará la hipótesis nula concluyendo que se puede aceptar la hipótesis de no haber correlación entre las estaturas de los hermanos y hermanas de la misma familia.

Obsérvese que el p-valor del test $2 \cdot P\{t_9 > 2'017\}$ estará entre los valores

$$2 \cdot P\{t_9 > 2'262\} < 2 \cdot P\{t_9 > 2'017\} < 2 \cdot P\{t_9 > 1'833\}$$

es decir, entre $0'1$ y $0'05$ no siendo, por tanto, la decisión anterior muy clara.

Con R hubiéramos ejecutado

```
> cor.test(hermanos,hermanas)

        Pearson's product-moment correlation

data:  hermanos and hermanas
t = 2.0175, df = 9, p-value = 0.07442
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06286527  0.86751705
sample estimates:
      cor
0.5580547
```

obteniendo, lógicamente, los mismos resultados.

— —

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 41206AD01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2014-2015

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Si X es una variable aleatoria con función de distribución,

$$F(x) = \begin{cases} 0 & \text{si} \quad x < 3 \\ 0'7 & \text{si} \quad 3 \leq x < 5 \\ 1 & \text{si} \quad x \geq 5 \end{cases}$$

calcular su media, su varianza y las siguientes probabilidades: $P\{X \leq 5\}$, $P\{3 \leq X \leq 4\}$, $P\{X \leq 4\}$, $P\{X > 4\}$.

Problema 2

En un artículo (Pardoen, 1989) se estudia el efecto natural de delaminación en vigas hechas a base de composición de láminas. Se anotó una medida de esa delaminación en cinco de estas vigas sometidas a carga, obteniéndose los siguientes datos:

$$230'66\,,\ 233'05\,,\ 232'58\,,\ 229'48\,,\ 232'58$$

Determinar un intervalo de confianza con un coeficiente de confianza de $0'95$ para la media de esa medida de delaminación, suponiendo que puede admitirse que sigue una distribución normal.

Problema 3

En un artículo (Yin y Jillie, 1987) se analiza el efecto de la tasa de flujo C_2F_6 sobre la uniformidad del grabado en una oblea de silicio utilizada en la fabricación de circuitos integrados.

La uniformidad observada (en porcentaje) en seis obleas escogidas al azar y tratadas con tres tasas de flujo de forma independiente fueron las siguientes:

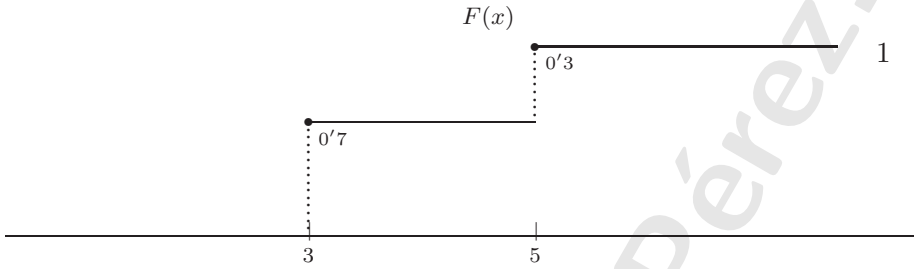
Tasas de Flujo						
125	**2′7**	**4′6**	**2′6**	**3′0**	**3′2**	**3′8**
160	**4′9**	**4′6**	**5′0**	**4′2**	**3′6**	**4′2**
200	**4′6**	**3′4**	**2′9**	**3′5**	**4′1**	**5′1**

¿Puede admitirse, a nivel de significación 0′05, la igualdad de las medias de las tres tasas de flujo mediante un Análisis de la Varianza?

Aunque no calcule exactamente el p-valor, ¿cree que la decisión tomada es fiable?

Problema 1

Por ser la función de distribución una función *en escalera*, la variable aleatoria X será de tipo discreto (EBR-sección 4.2). Una representación gráfica de la función de distribución será de utilidad,



Vemos que esta distribución sólo tiene masa en $x = 3$ (una masa de $0'7$) y en $x = 5$, en donde su masa es $0'3$.

Su media será la suma de los valores que toma por la probabilidad con que los toma (por lo que sólo son de interés los valores de X con masa positiva). La media será, por tanto,

$$E[X] = 3 \cdot 0'7 + 5 \cdot 0'3 = 3'6$$

y su varianza calculada como la *media de los cuadrados*

$$E[X^2] = 3^2 \cdot 0'7 + 5^2 \cdot 0'3 = 13'8$$

menos el cuadrado de la media, será

$$V(X) = E[X^2] - (E[X])^2 = 13'8 - 3'6^2 = 0'84.$$

Las probabilidades pedidas serán la suma de toda la masa existente en el intervalo para el cual se calcula la probabilidad. Así, una simple mirada al gráfico anterior permite concluir que

$$P\{X \leq 5\} = P\{X = 3\} + P\{X = 5\} = 0'7 + 0'3 = 1$$

$$P\{3 \leq X \leq 4\} = 0'7$$

$$P\{X \leq 4\} = 0'7$$

$$P\{X > 4\} = 0'3$$

Problema 2

Estamos en un caso de determinación del intervalo de confianza para la media de una población normal de varianza desconocida y muestras pequeñas (EBR-sección 6.2), intervalo que es

$$\left[\overline{x} - t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \, , \, \overline{x} + t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \right]$$

en donde S^2 es la cuasivarianza muestral.

De los datos del enunciado se deduce que la media y la cuasivarianza muestrales son,respectivamente, $\overline{x} = 231'67$ y $S^2 = 2'3442$. De las tablas de la distribución t de Student se obtiene el punto crítico, $t_{n-1;\alpha/2} = t_{4;0'025} = 2'776$ con lo que el intervalo de confianza buscado será

$$\left[\overline{x} - t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \, , \, \overline{x} + t_{n-1;\alpha/2} \, \frac{S}{\sqrt{n}} \right] =$$

$$= \left[231'67 - 2'776 \sqrt{\frac{2'3442}{5}} \, , \, 231'67 + 2'776 \sqrt{\frac{2'3442}{5}} \right] = [229'77 \, , \, 233'57]$$

ya que $2'776 \cdot \sqrt{2'3442/5} = 1'9$.

Problema 3

Se trata de contrastar la hipótesis nula de que no existen diferencias significativas entre las medias de tres tasas de flujo, $H_0 : \mu_{125} = \mu_{160} = \mu_{200}$, frente a la alternativa de no ser iguales las tres medias mediante un ANOVA (EBR-sección 9.2).

Para ello debemos formar la tabla ANOVA, que para los datos de este ejercicio queda igual a

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadísticos
Tasas de Flujo	$SST_i = 3'648$	2	$1'8239$	$F = 3'586$
Residual	$SSE = 7'63$	15	$0'5087$	
Total	$SST = 11'278$	17		

Ahora utilizaremos el estadístico de contraste F. Como, a partir de la Tabla 6 de la Adenda vemos que es $F_{(2,15);0'05} = 3'6823 > F = 3'586$, aceptamos

H_0, es decir, concluimos que no existen diferencias significativas entre las tres tasas de flujo a nivel $\alpha = 0'05$. No obstante, observamos que el p-valor debe de estar muy próximo a $0'05$ por lo que la decisión no es muy fiable al no ser suficientemente grande.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2014-2015

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

En un articulo de virus informáticos aparecen los 8 virus que produjeron más problemas en 2002. El nombre del virus y el porcentaje de incidentes registrados fue el siguiente:

Nombre	Porcentaje de incidentes
Gusano Klez	61′22 %
Gusano Lentin	20′52 %
Gusano Tanatos	2′09 %
Gusano BadtransII	1′31 %
Macro Word97 Thus db	1′19 %
Gusano Hybris	0′60 %
Gusano Bridex	0′32 %
Gusano Magistr	0′30 %
Otros	12′45 %

Si se comunican 3 incidentes por virus (los cuales suponemos actúan de forma independiente),

a) ¿Cuál es la probabilidad de que al menos uno sea un Gusano Klez?

b) ¿Cuál es la probabilidad de que más de dos sean un Gusano Klez?

c) ¿Cuál es la media y la desviación típica del número de incidentes por Gusano Klez de entre los 3 comunicados?

Problema 2

El Rectángulo de Oro es un rectángulo en donde el cociente Q entre el lado menor y el lado mayor es igual a $1/(0'5(\sqrt{5}+1)) = 0'618$ (razón áurea), utilizada por ejemplo por los griegos en el Partenón y hoy en día utilizada en las tarjetas de visita.

Se cree que las empresas con tarjetas de visita de valor Q mayor que la razón áurea, son conservadoras en sus negocios, mientras que empresas con tarjetas de visita de valores Q menor que la razón áurea son agresivas.

Con esta idea se eligieron al azar cinco compañías del Ibex 35 consideradas como conservadoras obteniéndose los valores Q

$$0'627\ ,\ 0'690\ ,\ 0'620\ ,\ 0'622\ ,\ 0'632$$

y cinco compañías consideradas como agresivas en las que se observaron los valores Q siguientes:

$$0'555\ ,\ 0'601\ ,\ 0'590\ ,\ 0'570\ ,\ 0'617$$

Suponiendo que los valores Q siguen una distribución normal, ¿es significativamente mayor la media de valores Q de las empresas del primer grupo (conservadoras) que las del segundo (agresivas)?

Problema 3

Se quiere analizar si existe un regresión lineal significativa entre la Edad del marido en años, X, y la de su esposa, Y, en 5 matrimonios, obteniéndose los resultados:

X	49	42	54	64	37
Y	43	40	53	61	38

Determinar la recta de regresión lineal de Y sobre X y analizar si es significativa.

Problema 1

El problema se puede modelizar con una variable aleatoria X con distribución binomial (EBR-sección 4.4.1) $B(n,p)$ en donde $n=3$ es el número de *pruebas de Bernoulli* y *éxito* es el que actúe el Gusano Klez, siendo por tanto la probabilidad de éxito $p=0'6122$. Las probabilidades buscadas serán,

a)

$$P\{X\geq 1\}=1-P\{X=0\}=1-\left(\binom{3}{0}0'6122^0(1-0'6122)^3\right)=$$

$$=1-(1-0'6122)^3=1-0'05832079=0'9417.$$

b)

$$P\{X>2\}=P\{X=3\}=\binom{3}{3}0'6122^3(1-0'6122)^0=0'6122^3=0'229.$$

c)

$$E[X]=n\cdot p=3\cdot 0'6122=1'8366$$

$$D[X]=\sqrt{n\cdot p\cdot(1-p)}=\sqrt{3\cdot 0'6122\cdot 0'3878}=\sqrt{0'7122}=0'8439.$$

Problema 2

Se trata de un problema de comparación de dos poblaciones normales con muestras pequeñas de varianzas desconocidas (EBR-sección 7.6) por lo que primero tenemos que analizar si pueden considerarse las varianzas como iguales o distintas. Para ello contrastaremos la hipótesis nula $H_0:\sigma_1^2=\sigma_2^2$ frente a la $H_0:\sigma_1^2\neq\sigma_2^2$ (EBR-sección 7.5), aceptándose la hipótesis nula si

$$\frac{S_1^2}{S_2^2}\in\left[F_{n_1-1,n_2-1;1-\alpha/2}\ ,\ F_{n_1-1,n_2-1;\alpha/2}\right].$$

A partir de nuestros datos obtenemos que es

$$\begin{array}{lll}\overline{x}_1=0'6382 & S_2^1=0'0008602 & n_2=5\\ \overline{x}_2=0'5866 & S_2^2=0'0006043 & n_1=5\end{array}$$

Como es $S_1^2/S_2^2=0'0008602/0'0006043=1'423465$, si consideramos un nivel de significación $\alpha=0'2$, será, a partir de la Tabla 6 de la F de Snedecor, $F_{4,4;1-0'1}=1/F_{4,4;0'1}=1/4'1073=0'2434689$, con lo que la región

de aceptación, a nivel $\alpha = 0'2$, es $[0'2435\,,\,4'1073]$, la cual contendrá al valor del estadístico y se aceptará la hipótesis nula de ser iguales ambas varianzas poblacionales, a ese nivel suficientemente alto. De hecho, el p-valor será mayor que $0'2$, suficientemente alto como para confirmar la aceptación de la igualdad de las varianzas poblacionales.

Por tanto, se trata ahora de un test para contrastar la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a la alternativa $H_1 : \mu_1 > \mu_2$ que aceptará H_0 cuando y sólo cuando sea

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\,\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha}$$

Como es

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\,\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{0'0516}{0'01711432} = 3'015$$

a partir de la Tabla 5 de la t de Student, vemos que el p-valor está acotado por

$$0'01 = P\{t_8 > 2'896\} > P\{t_8 > 3'015\} > P\{t_8 > 3'355\} = 0'005$$

con lo que es, en todo caso, menor que $0'01$, suficientemente pequeño como para rechazar la hipótesis nula y quedarnos con que el nivel medio de las razones Q de las tarjetas de las compañías conservadoras es significativamente mayor que el de las agresivas.

Problema 3

La recta de regresión

$$y_t = \widehat{\beta}_0 + \widehat{\beta}_1\, x$$

tiene por coeficientes de regresión $\widehat{\beta}_0$ y $\widehat{\beta}_1$ dados por (EBR-sección 10.2)

$$\widehat{\beta}_1 = \frac{n\displaystyle\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\displaystyle\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0'8966$$

y

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = 2'8889.$$

Es decir, la recta de regresión es

$$y_t = 2'8889 + 0'8966\, x.$$

Para analizar si esta recta es significativa para explicar la variable Y en función de X, debemos ejecutar el contraste de la regresión lineal simple (EBR-sección 10.3), para lo que debemos construir la tabla de Análisis de la Varianza (EBR página 301 y ADD página 26) que en nuestro problema toma el valor

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal	$SSEX = 355'9372$	1	$355'9372$	$F = 48'39873$
Residual	$SSNEX = 22'0628$	3	$7'354267$	
Total	$SST = 378$	4		

Ahora el test que se plantea es H_0 : *La regresión no es significativa*, o lo que es lo mismo, H_0 : *La recta de regresión no me sirve para explicar la variable Y en función de la X*, frente a la hipótesis alternativa, H_1 : *La regresión es significativa*. Puesto que el estadístico del contraste sigue una F de Snedecor con $(1, 3)$ grados de libertad, el punto crítico es, utilizando la Tabla 6 de la distribución F de Snedecor (ADD, página 36) $F_{(1,3);0'01} = 34'116 < 48'4 = F$, por lo que debemos rechazar la hipótesis nula y concluir que el ajuste es válido incluso a ese nivel de significación. Es decir, que el p-valor es menor que $0'01$, suficientemente pequeño como para concluir con un rechaza claro de la hipótesis nula y con que la recta es significativa.

Alternativamente se podía haber contrastado la hipótesis nula de que el coeficiente de regresión de la variable X es cero frente a la alternativa de que no es cero. En este caso, el estadístico de contraste hubiera tomado el valor $6'957$ y también se hubiera rechazado la hipótesis nula con un p-valor muy pequeño, en concreto, en este segundo caso, menor que

$$\text{p-valor} = 2 \cdot P\{t_3 > 6'957\} < 2 \cdot P\{t_3 > 5'841\} = 2 \cdot 0'005 = 0'01$$

es decir, de nuevo menor que $0'01$ y suficientemente pequeño como para confirmar el rechazo de la hipótesis nula y concluir que sí es cierto que la indica que la edad del marido X está linealmente relacionada con la de su mujer Y de forma significativa.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2014-2015

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

La duración en días X de un nueva componente electrónica sigue una distribución exponencial de función de densidad

$$f(x) = \frac{1}{200} e^{-x/200} \qquad , \qquad x > 0.$$

Se pide determinar,

(a) La función de distribución.

(b) La desviación típica.

(c) La probabilidad de que la componente electrónica dure más de 300 días.

(d) Si una componente electrónica de este modelo ha durado ya 150 días, ¿cuál es la probabilidad de que dure más de 200 días?

Problema 2

Los siguientes datos fueron utilizados por "Student" en su artículo en el que desarrolló la distribución t que lleva su nombre (Student, 1908). Corresponden al número de horas de sueño ganados por 10 pacientes usando, primero, D. hyoscyamine hydrobromide (grupo I) y, después, utilizando L. hyoscyamine hydrobromide (grupo II)

Grupo I	**0′7**	**−1′6**	**−0′2**	**−1′2**	**−0′1**	**3′4**	**3′7**	**0′8**	**0**	**2**
Grupo II	**1′9**	**0′8**	**1′1**	**0′1**	**−0′1**	**4′4**	**5′5**	**1′6**	**4′6**	**3′4**

Supuesto que puede admitirse la normalidad de los datos anteriores, se quiere averiguar si se ganan significativamente más horas de sueño con el segundo tratamiento que con el primero.

Problema 3

La siguiente tabla muestra la valoración subjetiva del número de personas reflejado en la tabla, acerca de su estado de salud (Turrall, 1992) en cinco regiones:

	Bueno	Aceptable	Malo
Southampton	**954**	**444**	**78**
Swindon	**985**	**504**	**87**
Jersey	**459**	**175**	**43**
Guernsey	**377**	**176**	**35**
West Dorset	**926**	**503**	**109**

(Es decir, por ejemplo, de las 1476 encuestadas en Southampton, 954 dijeron que su estado de salud es Bueno.)

Analizar si pueden considerarse equivalentes las cinco zonas muestreadas.

Problema 1

Este problema es muy parecido al Problema 2.12 de PREB y al ejercicio 4.4 de EBR.

(a) La función de distribución será (EBR-página 109)

$$F(x) = \int_0^x f(u)du$$

ya que la función f es distinta de cero sólo para $x > 0$. Es decir, si es $x > 0$ será

$$F(x) = \int_0^x f(u)du = \int_0^x \frac{1}{200} e^{-u/200} du = -e^{-u/200}\Big|_0^x = 1 - e^{-x/200}$$

siendo $F(x) = 0$, si es $x \leq 0$.

(b) La varianza de la distribución exponencial es (EBR-página 127) igual al parámetro al cuadrado, por lo que la desviación típica será

$$D[X] = 200.$$

(c) La probabilidad pedida será

$$P\{X > 300\} = 1 - F(300) = 1 - 1 + e^{-300/200} = e^{-1'5} = 0'2231.$$

(d) La probabilidad pedida será la probabilidad condicionada (EBR-página 100),

$$P\{X > 200/X > 150\} = \frac{P\{X > 200 \bigcap X > 150\}}{P\{X > 150\}}.$$

La intersección de los dos sucesos del numerador (que la duración sea mayor de 200 y mayor que 150) es el suceso primero por estar incluido en el segundo, por lo que

$$P\{X > 200/X > 150\} = \frac{P\{X > 200\}}{P\{X > 150\}} = \frac{1 - F(200)}{1 - F(150)} = \frac{e^{-200/200}}{e^{-150/200}} \frac{e^{-1}}{e^{-0'75}} = 0'7788.$$

Problema 2

Si resumimos ambas poblaciones por sus medias, se pide contrastar la hipótesis nula $H_0 : \mu_I \geq \mu_{II}$, frente a la alternativa $H_1 : \mu_I < \mu_{II}$.

Dado que las observaciones del Grupo I y las del Grupo II se han obtenido en las mismas personas, se trata de datos dependientes o apareados (EBR-sección 5.10) y cuyo tratamiento implica el uso de la variable diferencia D. La

hipótesis a contrastar se transforma en $H_0 : \mu_D \geq 0$, frente a la alternativa, $H_1 : \mu_D < 0$. Al ser tamaños muestrales pequeños, es necesario analizar la posible normalidad de los datos observados la cual suponemos por el enunciado del problema.

Ahora, contrastar las hipótesis mencionadas se llevará a cabo con el estadístico de contraste

$$\frac{\overline{d}}{S_d/\sqrt{n}}$$

que, para los datos del problema toma el valor: $\overline{d} = -1'58$ y $S_d = 1'23$, con lo que el estadístico del contraste toma el valor

$$\frac{\overline{d}}{S_d/\sqrt{n}} = \frac{-1'58}{1'23/\sqrt{10}} = -4'06$$

Como la región crítica de este test es la cola de la izquierda (EBR-sección 7.2, página 201), el p-valor del test será

$$P\{t_9 < -4'06\} < 0'0025$$

indicando el rechazo de la hipótesis nula y concluyendo, por tanto, que con el tratamiento segundo se ganan más horas de sueño de forma significativa.

Problema 3

Como los datos aportados son recuentos de observaciones clasificados por clases, comparar las cinco poblaciones debe hacerse mediante un test de la χ^2 de *homogeneidad de varias muestras* (EBR-sección 8.2.3), en donde la hipótesis nula que se establece es que las cinco poblaciones de las que proceden los datos son homogéneas. Esta hipótesis nula se rechazará cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i\, m_j/n)^2}{n_i\, m_j/n}$$

el estadístico de Pearson. En nuestro caso, toma el valor $\lambda = 18'4548$.

De la Tabla 4 de la χ^2 de Pearson vemos que el p-valor es

$$P\{\chi^2_{(r-1)(s-1)} > 18'4548\} = P\{\chi^2_8 > 18'4548\}$$

que está comprendido entre $0'01$ y $0'025$, suficientemente pequeño como para rechazar la hipótesis nula.

Las frecuencias esperadas son

```
        [,1]     [,2]     [,3]
[1,] 932.9933 454.2702 88.73646
[2,] 996.2043 485.0473 94.74842
[3,] 427.9380 208.3611 40.70094
[4,] 371.6803 180.9694 35.35030
[5,] 972.1841 473.3520 92.46388
```

todas mayores que 5.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 41206AD01A01).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada** (2014). Alfonso García Pérez, A. (2014). Editorial UNED, Colección Temática (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica** (1998). Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada** (2008). Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2015-2016

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

El cinco por ciento de las personas a las que se les administra un determinado medicamento sufre efectos secundarios tras su administración. Si el medicamento fue aplicado a 10 personas, determinar la probabilidad de que

(a) Exactamente uno padezca efectos secundarios.

(b) Lo padezca más de dos personas.

(c) Si se suministrase el antibiótico a 1000 individuos, ¿cuál sería el número medio de personas con efectos secundarios?.

(d) Calcular la probabilidad de que, de esas mil personas, padezcan efectos secundarios más de 40.

Problema 2

Una persona quiere hacer una inversión en donde la volatilidad, medida por la desviación típica de las ganancias mensuales, sea menor de 100 euros. El banco en donde tiene sus ahorros le ofrece un producto en el que las ganancias en euros de los últimos meses para el capital que él quiere invertir han sido

$$100\ ,\ 50\ ,\ -10\ ,\ 55\ ,\ -25\ ,\ 75\ ,\ 90$$

Si, como le dice su asesor financiero, las ganancias mensuales de este producto siguen una distribución normal, ¿es la volatilidad del producto financiero en cuestión menor de 100 con un nivel de significación de $0'05$? Después de llegar a una conclusión con ese nivel de significación, acote el p-valor y comente el resultado obtenido.

Problema 3

En un trabajo de Davidson (1993) se recogen los niveles de ozono en el aire de la costa sur de California entre los años 1976 y 1991. El autor del trabajo cree que el número de días que el nivel de ozono supera los niveles de 0′2 ppm (partes por millón) depende del denominado Índice Meteorológico Estacional que es el promedio estacional de temperatura a 850 milibares de presión. Parte de los datos son:

Índice	**16′7**	**17′1**	**18′2**	**18′1**	**17′2**	**18′2**
Días	**91**	**105**	**106**	**108**	**88**	**91**

Determinar la recta de regresión de la variable dependiente Y = Número de días, explicada por la variable independiente X = Índice estacional y analizar si es significativa para explicar Y en función de X.

Problema 1

a) Este problema, muy parecido al problema 2.13 del texto PREB, se puede formalizar mediante un modelo binomial (EBR-sección 4.4.1) en donde cada prueba de Bernoulli sea el administrar el medicamento en cuestión y el suceso *éxito*, el que el paciente sufra efectos secundarios. De esta forma, la variable *número de pacientes, de entre los diez, que padecieron efectos secundarios*, se puede modelizar mediante una variable X con distribución binomial $B(10, 0'05)$, al ser $p = 0'05$ la probabilidad de que se dé el suceso *éxito*.

La probabilidad pedida será ahora, utilizando la Tabla 1 de la distribución binomial (ADD),

$$P\{X = 1\} = 0'3151.$$

b) En la misma situación que en el apartado anterior, la probabilidad pedida será, por el suceso complementario

$$\begin{aligned} P\{X > 2\} &= 1 - P\{X \leq 2\} = 1 - [P\{X = 0\} + P\{X = 1\} + P\{X = 2\}] \\ &= 1 - [0'5987 + 0'3151 + 0'0746] = 0'0116 \end{aligned}$$

o directamente,

$$P\{X > 2\} = P\{X = 3\} + P\{X = 4\} + P\{X = 5\} = 0'0105 + 0'001 + 0'0001 = 0'0116.$$

c) Ahora lo que ocurre es que se aumenta el número de pruebas de Bernoulli, modelizándose el problema con una variable $X \rightsquigarrow B(1000, 0'05)$. La media de esta distribución es el producto de los dos parámetros, es decir,

$$E[X] = n \cdot p = 1000 \cdot 0'05 = 50.$$

Por tanto, el número medio o número esperado de personas con efectos secundarios, de entre los mil, sería 50.

d) El cálculo de probabilidades de distribuciones binomiales para un gran número de ensayos, como aquí ocurre, se realiza aproximando dicha distribución mediante el *teorema central del límite* (EBR-sección 4.7).

En el caso de una distribución binomial $X \rightsquigarrow B(n, p)$, su aproximación mediante una normal $Y \rightsquigarrow N(np, \sqrt{np(1-p)})$ es válida (EBR-sección 4.7) cuando supuesto sea $p \leq 0'5$ (como aquí ocurre) entonces sea también $np > 5$ (como aquí ocurre).

Por tanto, aproximaremos la $X \rightsquigarrow B(1000, 0'05)$, por una

$$Y \rightsquigarrow N\left(1000 \cdot 0'05\,,\, \sqrt{1000 \cdot 0'05 \cdot 0'95}\right) = N(50\,,\, 6'892)$$

quedando la probabilidad pedida igual a

$$P\{X > 40\} = P\left\{\frac{X-50}{6'892} > \frac{40-50}{6'892}\right\} = P\{Z > -1'45\} = 1 - P\{Z \leq -1'45\} =$$

$$= 1 - P\{Z > 1'45\} = 1 - 0'0735 = 0'9265$$

siendo Z una variable aleatoria $N(0,1)$ y en donde la última probabilidad la hemos calculado utilizando la Tabla 3 de dicha distribución, ADD.

Problema 2

El problema 4.4 de PREB determinaba un intervalo de confianza en situaciones distintas, pero puede ser una buena guía. En este caso, del enunciado se desprende que el cliente debe efectuar un test de hipótesis sobre la desviación típica. Como siempre que se pueda, la hipótesis a analizar $H_1 : \sigma < 100$ debe de establecerse como hipótesis alternativa para poder controlar el error cometido en la decisión tomada. Además, otro criterio muy tosco es que el signo "igual" debe de ir en la hipótesis nula.

Por tanto, se trata de un test de $H_0 : \sigma \geq 100$ frente a $H_1 : \sigma < 100$ para el caso de una población normal de media desconocida (EBR-sección 7.4). Como allí aparecen las hipótesis en términos de las varianzas, podemos establecerlas diciendo que se trata de un contraste de $H_0 : \sigma^2 \geq 10000$ frente a $H_1 : \sigma^2 < 10000$, rechazándose H_0 cuando y sólo cuando sea

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1;1-\alpha}$$

Al ser

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{6 \cdot 2323'81}{10000} = 1'39$$

y el punto crítico, a partir de la Tabla 4 (ADD) de la distribución χ^2, ser $\chi^2_{n-1;1-\alpha} = \chi^2_{6;0'95} = 1'635 > 1'39$ se deberá rechazar la hipótesis nula al caer en la región crítica el valor del estadístico del contraste.

Para valorar la decisión tomada debemos calcular el p-valor. Con los datos observados, el p-valor del test será $P\{\chi^2_6 < 1'39\}$, por ser la región crítica la cola de la izquierda.

A partir de la Tabla 4 de la distribución χ^2 (ADD) podemos acotarlo de la siguiente manera,

$$P\{\chi^2_6 < 1'237\} < P\{\chi^2_6 < 1'39\} < P\{\chi^2_6 < 1'635\}$$

$$1 - P\{\chi^2_6 < 1'237\} > 1 - P\{\chi^2_6 < 1'39\} > 1 - P\{\chi^2_6 < 1'635\}$$

$$P\{\chi^2_6 \geq 1'237\} > 1 - P\{\chi^2_6 < 1'39\} > P\{\chi^2_6 \geq 1'635\}$$

$$0'975 > 1 - P\{\chi^2_6 < 1'39\} > 0'95$$

$$1 - 0'975 < P\{\chi^2_6 < 1'39\} < 1 - 0'95$$

$$0'025 < P\{\chi^2_6 < 1'39\} < 0'05$$

con lo que el p-valor está acotado por

$$0'025 < \text{p-valor} < 0'05$$

lo que indica que la decisión tomada no es muy concluyente ya que suele decirse (EBR-página 194) que un p-valor menor que $0'01$ lleva a conclusiones claras de rechazo de la hipótesis nula y que p-valores mayores de $0'2$ a conclusiones claras de aceptación de la hipótesis nula, sugiriendo valores intermedios la repetición del experimento.

Observamos también que al ser el p-valor menor que $0'05$, que es el nivel de significación del test, ya podíamos haber concluido que el valor del estadístico cae en la región crítica, rechazando la hipótesis nula, y habernos ahorrado el camino de la determinación del punto crítico.

Problema 3

La recta de regresión que se obtiene es (EBR-sección 10.2)

$$\text{Días} = -1'602 + 5'674\ \text{Índice}$$

y uno de los dos posibles tests para analizar su significación consiste en contrastar la hipótesis nula de que es cero el coeficiente de regresión de la variable independiente X y está basado en la siguiente tabla

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.602    109.712  -0.015    0.989
Indice         5.674      6.236   0.910    0.414
```

que proporciona un p-valor igual a $0'414$ y que sugiere, por tanto, que aceptemos la hipótesis nula de ser cero el coeficiente de regresión asociado a la variable independiente, es decir, que la recta no es significativa para explicar a la variable dependiente en función de la independiente.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2015-2016

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres apartados puntúan lo mismo.

Problema 1

La estatura de los individuos de una determinada población sigue una distribución normal de desviación típica igual a 10 cm. Calcular el número mínimo de individuos que habrá que seleccionar de dicha población para que la probabilidad de que la estatura media de dicha muestra difiera de la poblacional en menos de 5 cm., sea 0′9.

Problema 2

Se cree que el Tarbosaurus Bataar es solamente un representante asiático del género norteamericano Tyrannosaurus Rex, de manera que el género Tarbosaurus no existiría en realidad como género diferenciado. Para analizar si existen o no diferencias significativas entre ambos grupos de dinosaurios, se midieron las longitudes (en m.) de los cráneos de 7 ejemplares fósiles encontrados en rocas del Cretácico Superior en Mongolia de Tarbosaurus Bataar y las de 9 ejemplares de Tyrannosaurus Rex de un yacimiento del oeste de Estados Unidos (Hurum y Sabath, 2003). Los resultados aparecen a continuación:

Tarbasaurus: 0′76 , 0′73 , 0′82 , 0′68 , 0′73 , 0′69 , 0′84

Tyrannosaurus: 0′82 , 0′90 , 0′88 , 0′74 , 0′79 , 0′84 , 0′79 , 0′82 , 0′80

Analizar si hay diferencias significativas entre los dos grupos en cuanto a las longitudes de sus cráneos, mediante

a) El test de la t de Student.

b) El test de Wilcoxon-Mann-Whitney.

Problema 1

Estamos en una situación como la de la Sección 5.11 de EBR. Se pide determinar el tamaño muestral necesario para que se verifique la condición expresada en el enunciado; en concreto, si la *estatura en cm. de los individuos de la población*, la representamos por la variable aleatoria X y admitimos que es $X \rightsquigarrow N(\mu, 10)$, se pide determinar n de forma que sea

$$P\{|\overline{x} - \mu| < 5\} = 0'9.$$

En estas condiciones sabemos (EBR-sección 5.4) que la media muestral se distribuye como $\overline{x} \rightsquigarrow N(\mu, 10/\sqrt{n})$. Por tanto, tipificando en la condición anterior y si, como siempre, Z representa una variable $N(0,1)$, será

$$P\left\{\frac{|\overline{x} - \mu|}{10/\sqrt{n}} < \frac{5}{10}\sqrt{n}\right\} = P\left\{|Z| < \frac{5}{10}\sqrt{n}\right\} = 0'9.$$

Ahora vamos a buscar en la Tabla 3 de la normal $N(0,1)$ un valor z tal que sea

$$P\{|Z| < z\} = 0'9$$

es decir, un valor tal que la $N(0,1)$ deje un área de probabilidad $0'9$ entre $-z$ y z, la Tabla 3 nos da como solución $z = 1'645$.

Por tanto, deberá ser

$$\frac{5}{10}\sqrt{n} = 1'645$$

de donde se obtiene el valor $n = 10'8241$, aunque como habrá que elegir un tamaño de muestra entero y como a medida que aumenta n, aumenta la probabilidad del suceso puesto como condición (al aparecer n en el numerador) tomaremos $n = 11$ como n mínimo que mantenga la precisión exigida en el enunciado.

Problema 2

a) Para ejecutar el test de la t de Student de comparación de dos poblaciones, necesitamos comprobar si puede aceptarse o no que las varianzas de ambas poblaciones puedan considerarse iguales (EBR-sección 7.6), es decir, necesitamos contrastar la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ frente a la alternativa de ser distintas (EBR-sección 7.5), contraste basado en el estadístico S_1^2/S_2^2. De hecho, aceptaremos esta hipótesis nula cuando y sólo cuando sea,

$$\frac{S_1^2}{S_2^2} \in [\ F_{n_1-1,n_2-1;1-\frac{\alpha}{2}}\ ,\ F_{n_1-1,n_2-1;\frac{\alpha}{2}}\ ].$$

De los datos obtenemos que es $n_1 = 7$, $n_2 = 9$, $\overline{x}_1 = 0'75$, $S_1^2 = (0'061)^2 = 0'0037$, $\overline{x}_2 = 0'82$, $S_2^2 = (0'0487)^2 = 0'00237$, por lo que es $S_1^2/S_2^2 = 1'57$.

Si consideramos un nivel de significación $\alpha = 0'1$, será, a partir de la Tabla 6 de la F de Snedecor, $F_{6,8;0'95} = 1/F_{8,6;0'05} = 1/4'1468 = 0'24115$ y $F_{6,8;0'05} = 3'5806$, con lo que la región de aceptación, a nivel $\alpha = 0'1$, es $[0'241\,,\,3'581]$, que contendrá al valor del estadístico por lo que se aceptará la hipótesis nula de ser iguales ambas varianzas poblacionales, a ese nivel de significación, suficientemente grande.

Por tanto, estaremos en un caso de comparación de las medias de dos poblaciones normales independientes con muestras pequeñas y con varianzas desconocidas pero que pueden suponerse iguales (EBR-sección 7.6), para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$, aceptándose H_0 cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\,\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha/2}$$

Como es

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\,\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|0'75 - 0'82|}{\sqrt{\dfrac{6\cdot 0'00373 + 8\cdot 0'002375}{14}}\,\sqrt{\dfrac{1}{7}+\dfrac{1}{9}}} = 2'554$$

y, a partir de la Tabla 5 de la t de Student, vemos que es $0'01 < P\{t_{14} > 2'554\} < 0'025$, por lo que el p-valor de este test bilateral estará entre $0'02$ y $0'05$, es decir, no queda clara la decisión aunque se rechaza para niveles de significación mayores que el p-valor, por lo que para un nivel $\alpha = 0'05$ se rechaza la hipótesis nula, existiendo, por tanto, diferencias significativas entre ambos grupos de dinosaurios en cuanto a las longitudes de sus cráneos.

b) Para utilizar el test de Wilcoxon-Mann-Whitney, EBR-sección 8.4.1, las hipótesis a contrastar harán referencia a las medianas poblacionales M_1 y M_2 y serán $H_0 : M_1 = M_2$ frente a la alternativa $H_1 : M_1 \neq M_2$, aceptándose H_0 cuando y sólo cuando sea

$$m \cdot n - u_{m,n;\alpha/2} < U < u_{m,n;\alpha/2}$$

siendo U el número de valores de la segunda muestra que preceden estrictamente a cada valor fijo de la primera muestra.

Si subrayamos los valores de la segunda muestra en la siguiente unión de ambas muestras ordenadas, en donde los de la segunda muestra que son iguales a los de la primera los hemos situado detrás para hacer más simple el recuento, tenemos la siguiente secuencia:

$0'68, 0'69, 0'73, 0'73, \underline{0'74}, 0'76, \underline{0'79}, \underline{0'79}, \underline{0'80}, 0'82, \underline{0'82}, \underline{0'82}, 0'84, \underline{0'84}, \underline{0'88}, \underline{0'90}$

Ahora, para calcular el valor de U nos fijamos en cada valor de la primera muestra (es decir, cada valor no subrayado) y vemos cuantos valores de la segunda muestra (es decir, cuántos valores subrayados) le preceden. Es decir, miramos el $0'68$ (primer valor no subrayado) y vemos que no hay ningún valor subrayado que lo preceda, por lo que el primer sumando de U es 0; así sucesivamente, vemos que U toma el valor,

$$U = 0 + 0 + 0 + 0 + 1 + 4 + 6 = 11.$$

En la determinación del punto crítico y el p-valor utilizaremos la aproximación normal ya que los tamaños muestrales son mayores que 5. En concreto, si el nivel de significación es $\alpha = 0'05$, será $z_{\alpha/2} = z_{0'025} = 1'96$ y

$$u_{m,n;\alpha/2} = u_{7,9;0'025} = \frac{7 \cdot 9}{2} + 1'96 \sqrt{\frac{7 \cdot 9 \cdot (9 + 7 + 1)}{12}} = 50'02$$

y la región de aceptación,

$$(m \cdot n - u_{m,n;\alpha/2}\ ,\ u_{m,n;\alpha/2}) = (7 \cdot 9 - 50'02\ ,\ 50'02) = (12'98\ ,\ 50'02).$$

Como $U = 11$ no pertenece a ella, deberemos rechazar la hipótesis nula de igualdad de ambas poblaciones, con un p-valor menor que el nivel de significación $0'05$. En concreto, el p-valor (aproximado por utilizar la aproximación normal) será, utilizando la cola inferior, al ser el p-valor el menor nivel de significación para el que se rechaza la hipótesis nula,

$$2 \times P\{U < 11\} \simeq 2 \times P\left\{ Z < \frac{11 - 7 \cdot 9/2}{\sqrt{7 \cdot 9(9 + 7 + 1)/12}} \right\} =$$

$$= 2 \times P\{Z < -2'17\} = 2 \times P\{Z > 2'17\} = 2 \times 0'015 = 0'03$$

que indica rechazar la hipótesis nula de igualdad aunque no con mucha seguridad.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2015-2016

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Se quiere determinar un intervalo de confianza para el tiempo medio en días que transcurre desde el inicio de los síntomas hasta la completa desaparición del herpes zóster. Para ello se eligieron al azar 40 individuos en los que se observó una duración media de 21′3 días y una cuasivarianza muestral igual a 9 días.

Determinar el mencionado intervalo de confianza para un coeficiente de confianza de 0′95.

Problema 2

Se quiere analizar si existe una regresión lineal simple entre la Estatura en cm., X, y el Número de Pulsaciones por minuto, Y. Para ello se recogieron los siguientes datos en 8 personas:

X	167	160	167	162	175	185	162	173
Y	62	68	65	64	80	80	64	76

Determinar la recta de regresión (de Y sobre X) y analizar si es significativa.

Problema 3

Se está estudiando la distribución de los cuatro grupos sanguíneos O, A, B, AB en una población. Extraída una muestra aleatoria de ella, de $n = 353$ individuos, se obtuvieron las siguientes frecuencias absolutas

Grupo sanguíneo	Frecuencias absolutas
O	121
A	120
B	79
AB	33

Un modelo teórico (principio de Hardy-Weinberg) asigna las siguientes probabilidades a cada uno de los grupos

Grupo sanguíneo	Probabilidades
O	r^2
A	$p^2 + 2pr$
B	$q^2 + 2qr$
AB	$2pq$

con $p + q + r = 1$.

A partir de los datos de la muestra se obtuvieron las siguientes estimaciones de los parámetros: $\hat{p} = 0'2465$ y $\hat{q} = 0'1732$.

Contrastar la hipótesis de que los datos se ajustan al modelo teórico.

Problema 1

Se trata de determinar el intervalo de confianza para la media de una población no necesariamente normal y muestras grandes (EBR-sección 6.3), para el caso de varianza desconocida,

$$\left[\overline{x} - z_{\alpha/2}\,\frac{S}{\sqrt{n}}\,,\overline{x} + z_{\alpha/2}\,\frac{S}{\sqrt{n}}\right].$$

Como de la Tabla 3 de la distribución normal $N(0,1)$ obtenemos que es $z_{\alpha/2} = z_{0'05/2} = z_{0'025} = 1'96$, el intervalo de confianza buscado será,

$$\begin{aligned}\left[\overline{x} - z_{\alpha/2}\,\frac{S}{\sqrt{n}}\,,\overline{x} + z_{\alpha/2}\,\frac{S}{\sqrt{n}}\right] &= \left[21'3 - 1'96\,\frac{3}{\sqrt{40}}\,,21'3 + 1'96\,\frac{3}{\sqrt{40}}\right] \\ &= [20'37\ ,\ 22'23].\end{aligned}$$

Problema 2

El estudio de la Regresión Lineal Simple aparece en EBR-capítulo 10. La recta de regresión solicitada es

$$y = -57'0265 + 0'7515\,x.$$

Para analizar su significación lo mejor es hacer un test sobre el coeficiente de regresión β_1 de la variable independiente X, EBR-sección 10.3.2. El p-valor de dicho test, que tiene como hipótesis nula que esta recta no es significativa, o equivalentemente que $\beta_1 = 0$ resulta ser muy pequeño, $0'0099$, concluyéndose, por tanto, que la recta sí es válida para explicar la variable dependiente en función de la independiente.

Aunque lógicamente no es viable en un examen, por completar el ejercicio, su ejecución y análisis con R aparece en EBR-sección 10.4. Éste se haría incorporando los datos con `(1)`, obteniendo la recta ejecutando `b(2)`, que se obtiene en `(3)`.

El p-valor del test antes mencionado se obtiene en `(4)`.

```
> x<-c(167,160,167,162,175,185,162,173)                    (1)
> y<-c(62,68,65,64,80,80,64,76)                            (1)
> recta<-lm(y~x)                                           (2)

> recta

Call:
lm(formula = y ~ x)
```

```
Coefficients:
(Intercept)            x
   -57.0265       0.7515                                              (3)

> summary(recta)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4660 -2.3606 -0.7088  3.4675  5.5224

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.0265    34.1879  -1.668   0.1464
x             0.7515     0.2022   3.716   0.0099 **
                                          (4)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.499 on 6 degrees of freedom
Multiple R-squared:  0.6971,    Adjusted R-squared:  0.6466
F-statistic: 13.81 on 1 and 6 DF,  p-value: 0.0099
```

Problema 3

Se trataría de contrastar la hipótesis nula de que los datos observados se ajustan al modelo teórico propuesto en el enunciado. Estas diferencias se contrastan mediante un test de bondad del ajuste de la χ^2 de Pearson (EBR-sección 8.2.2). Para poder ejecutar este test necesitamos construir la tabla de frecuencias esperadas $n \cdot p_i$, si fuera cierto el modelo teórico, para compararla con la de frecuencias observadas n_i, mediante el estadístico del test.

Dado que es $p+q+r=1$ deberá ser $\hat{r} = 1 - 0'2465 - 0'1732 = 0'5803$, con lo que las probabilidades esperadas para cada una de las cuatro clases será

Grupo sanguíneo	Probabilidades	
O	r^2	$0'3367481$
A	$p^2 + 2pr$	$0'3468502$
B	$q^2 + 2qr$	$0'2310142$
AB	$2pq$	$0'0853876$
		1

Si ahora multiplicamos las probabilidades esperadas para cada grupo sanguíneo por la frecuencia total $n = 353$, tendremos la tabla de frecuencias esperadas,

Grupo sanguíneo	Frecuencias esperadas
O	118′87
A	122′44
B	81′55
AB	30′14
	353

El valor del estadístico λ de Pearson será igual a

$$\lambda = \sum_i \left(\frac{n_i^2}{n \cdot p_i} \right) - n = \left(\frac{121^2}{118'87} + \frac{120^2}{122'44} + \frac{79^2}{81'55} + \frac{33^2}{30'14} \right) - 353$$

$$= 353'4379 - 353 = 0'4379.$$

Como hemos estimado dos parámetros a partir de la muestra, reduciremos dos grados más los de libertad de la χ^2 que serán por tanto, $4 - 1 - 2 = 1$.

Ya podemos plantear el test en donde la hipótesis nula es que los datos se ajustan bien al modelo y la alternativa es que se ajustan mal. Si fijamos un nivel de significación de 0′05, el punto crítico obtenido de unas tablas de la χ^2 será $\chi^2_{1;0'05} = 3'841 > 0'4379 = \lambda$ por lo que aceptaremos la hipótesis nula concluyendo con que los datos se ajustan bien al modelo teórico.

El p-valor del test es

$$P\{\chi_1^2 > 0'4379\} > 0'3$$

el cual confirma la decisión de aceptación de la hipótesis nula.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 41206AD01A01).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada** (2014). Alfonso García Pérez, A. (2014). Editorial UNED, Colección Temática (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica** (1998). Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada** (2008). Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2016-2017

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los siguientes datos corresponden a la marca obtenida en los juegos olímpicos de Seúl de 1988 en Salto de Altura por 25 atletas femeninas de heptatlón:

$$1'86\ ,\ 1'80\ ,\ 1'83\ ,\ 1'80\ ,\ 1'74\ ,\ 1'83\ ,\ 1'80\ ,\ 1'80\ ,\ 1'83$$
$$1'77\ ,\ 1'86\ ,\ 1'80\ ,\ 1'86\ ,\ 1'83\ ,\ 1'80\ ,\ 1'86\ ,\ 1'80$$
$$1'83\ ,\ 1'71\ ,\ 1'77\ ,\ 1'77\ ,\ 1'71\ ,\ 1'68\ ,\ 1'71\ ,\ 1'50$$

Se pide determinar: La Distribución de Frecuencias Absolutas, el Diagrama de Barras, la Media, la Mediana, la Moda, El Primer Cuartil, El Tercer Cuartil, la Desviación Típica, el Recorrido, y el Coeficiente de Asimetría de Pearson.

Problema 2

Utilizando los datos del problema anterior y el test de los signos, ¿puede admitirse, a un nivel de significación $\alpha = 0'05$, la hipótesis alternativa de que la marca mediana es mayor que $1'82$?

Problema 3

Los siguientes datos (Mazess et al, 1984), corresponden al Porcentaje de `Grasa` corporal y la `Edad` de hombres entre 23 y 61 años.

Grasa	$15'5$	$20'9$	$18'6$	$28'0$	$21'3$	$9'5$	$7'8$	$17'8$	$25'9$	$27'4$
Edad	24	37	41	60	58	23	27	27	41	45

Determinar la recta de regresión de la variable dependiente $Y =$ Grasa, explicada por la variable independiente $X =$ Edad y analizar si es significativa para explicar Y en función de X.

Problema 1

La distribución de frecuencias absolutas (EBR-sección 2.3) corresponderá a la de un carácter cuantitativo sin agrupar y será

X_i	n_i
$1'50$	1
$1'68$	1
$1'71$	3
$1'74$	1
$1'77$	3
$1'80$	7
$1'83$	5
$1'86$	4
	25

El Diagrama de Barras es el de la Figura 0.1.

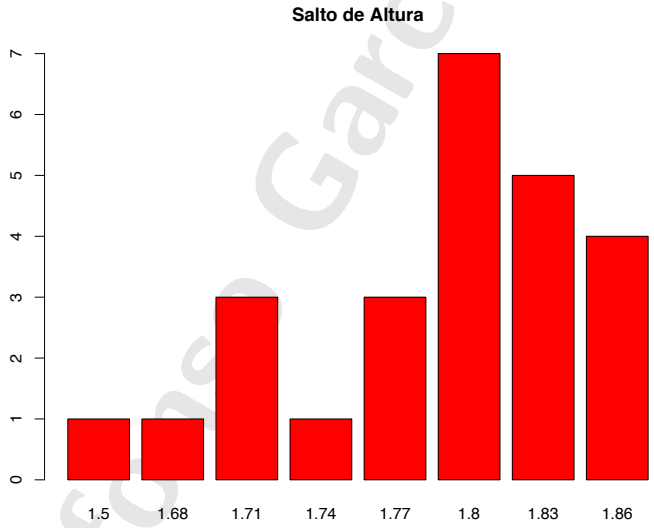


Figura 0.1 : Diagrama de barras

Las Medidas de Posición (EBR-sección 2.3.2) serán, la Media

X_i	n_i	$X_i \cdot n_i$
$1'50$	1	$1'50$
$1'68$	1	$1'68$
$1'71$	3	$5'13$
$1'74$	1	$1'74$
$1'77$	3	$5'31$
$1'80$	7	$12'6$
$1'83$	5	$9'15$
$1'86$	4	$7'44$
	25	$44'55$

$$\overline{x} = \frac{44'55}{25} = 1'782$$

La Moda (el valor más frecuente), $M_d = 1'80$.

Como la distribución de frecuencias acumuladas es

X_i	n_i	N_i
$1'50$	1	1
$1'68$	1	2
$1'71$	3	5
$1'74$	1	6
$1'77$	3	9
$1'80$	7	16
$1'83$	5	21
$1'86$	4	25
	25	

será

$$9 < 12'5 < 16$$

con lo que la Mediana corresponderá al valor asociado a la frecuencia absoluta acumulada 16, es decir, $M_e = 1'80$.

Por otro lado, al ser

$$6 < 25/4 = 6'25 < 9$$

el primer cuartil será $p_{1/4} = 1'77$ y, al ser

$$16 < 3 * 25/4 = 18'75 < 21$$

el tercer cuartil será $p_{3/4} = 1'83$.

En cuanto a las medidas de dispersión (EBR-sección 2.3.3), al ser la varianza

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \overline{x}^2 = \frac{79'5339}{25} - 1'782^2 = 0'005832$$

la desviación típica será $s = \sqrt{s^2} = 0'07637$.

El Recorrido será $R = 1'86 - 1'5 = 0'36$.

Por último, el Coeficiente de Asimetría de Pearson, EBR-sección 2.3.4, será

$$A_p = \frac{\overline{x} - M_d}{s} = \frac{1'782 - 1'80}{0'07637} = -0'2357$$

mostrando los datos una cierta asimetría a la izquierda, como ya se podía deducir del gráfico de su distribución de frecuencias.

Aunque en el examen resulta imposible, si hubiéramos resuelto el problema con R hubiéramos ejecutado la siguiente secuencia de instrucciones obteniendo, lógicamente, los mismos resultados. Con `(1)` incorporamos los datos. Con `(2)` formamos la tabla de frecuencias absolutas. Con `(3)` el gráfico de barras. Con `(5)` varias de la medidas solicitadas que son completadas con `(6)`, `(7)`, `(8)` y `(9)`.

```
> Salto<-c(1.86,1.80,1.83,1.80,1.74,1.83,1.80,1.80,1.83,1.77,1.86,1.80,     (1)
+ 1.86,1.83,1.80,1.86,1.80,1.83,1.71,1.77,1.77,1.71,1.68,1.71,1.50)         (1)

> table(Salto)                                                              (2)

 1.5 1.68 1.71 1.74 1.77  1.8 1.83 1.86
   1    1    3    1    3    7    5    4

> valores<-c(1.5,1.68,1.71,1.74,1.77,1.8,1.83,1.86)                        (3)
> frecuencias<-c(1,1,3,1,3,7,5,4)                                          (3)

> barplot(frecuencias,names=valores,col=2,main="Salto de Altura")           (4)

> summary(Salto)                                                            (5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Mx.
  1.500   1.770   1.800   1.782   1.830   1.860

> library(modeest)

> mfv(Salto)                                                                (6)
[1] 1.8

> sqrt(24*var(Salto)/25)                                                    (7)
[1] 0.07636753

> (mean(Salto)-mfv(Salto))/sqrt(24*var(Salto)/25)                           (8)
[1] -0.2357023

> range(Salto)
```

```
[1] 1.50 1.86
> range(Salto)[2]-range(Salto)[1]                                    (9)
[1] 0.36
```

Problema 2

El test de los signos viene explicado en EBR-sección 8.3.1. Este problema es similar al Problema 11.3 del texto PREB.

Como se solicita en el enunciado, con objeto de comprobar si puede admitirse una marca mayor de $1'82$ se establece esta hipótesis como alternativa y se contrasta $H_0 : M \leq 1'82$ frente a $H_1 : M > 1'82$.

El estadístico T del test de los signos es el número de signos positivos de entre las diferencias $X_i - M_0$ rechazándose la hipótesis nula cuando sea $T \geq t_\alpha$, siendo t_α el menor número entero para el que

$$\sum_{t=t_\alpha}^{n} \binom{n}{t} (0'5)^n \leq \alpha.$$

Como el tamaño muestral es mayor que 12, podemos utilizar la aproximación normal en la determinación del punto crítico, siendo

$$t_\alpha = 0'5(z_\alpha \sqrt{n} + n + 1).$$

Al ser $\alpha = 0'05$ será $z_\alpha = 1'645$ y

$$t_\alpha = 0'5(1'645\sqrt{25} + 25 + 1) = 17'1125.$$

Como para los datos de este ejercicio es $T = 9 < 17'1125$ no podemos rechazar la hipótesis nula y debemos concluir que la mediana es menor o igual que $1'82$.

Problema 3

La recta de regresión que se obtiene es (EBR-sección 10.2)

$$\text{Grasa} = 3'9061 + 0'4011 \text{ Edad}$$

y uno de los dos posibles tests para analizar su significación consiste en contrastar la hipótesis nula de que es cero el coeficiente de regresión de la variable independiente X. Este test nos da un p-valor igual a $0'00902$ y que sugiere, por tanto, que rechacemos la hipótesis nula de ser cero el coeficiente de regresión asociado a la variable independiente, es decir, podemos concluir con que la recta es significativa para explicar a la variable dependiente en función de la independiente y que el Porcentaje de Grasa Corporal depende de la Edad.

Si hubiéramos utilizado R (cosa imposible en el examen pero que puede ser útil para conocer cómo resolverlo con este paquete en otras ocasiones), hubiéramos ejecutado los siguientes comandos:

```
> y
 [1] 15.5 20.9 18.6 28.0 21.3  9.5  7.8 17.8 25.9 27.4
> x
 [1] 24 37 41 60 58 23 27 27 41 45

recta<-lm(y~x)
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     3.9061       0.4011


> summary(recta)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9370 -3.1626  0.9958  2.8351  5.5469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9061     4.7263   0.826  0.43250
x             0.4011     0.1171   3.425  0.00902 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.706 on 8 degrees of freedom
Multiple R-squared:  0.5945,    Adjusted R-squared:  0.5438
F-statistic: 11.73 on 1 and 8 DF,  p-value: 0.009022
```

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2016-2017

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Si suponemos que el sexo del bebé que tienen las parejas es independiente en cada ocasión que tienen un bebé y que la probabilidad de varón es igual que la de hembra, determinar la probabilidad de que una pareja tenga cuatro niñas antes del segundo niño.

Problema 2

Se quiere comparar el Peso de dos razas independientes de animales de granja para averiguar si existen o no diferencias significativas entre ellas, suponiendo distribuciones normales para dichos Pesos. Para ello se seleccionaron al azar $n_1 = 5$ animales de la primera raza y $n_2 = 5$ de la segunda, obteniéndose los siguientes pesos en gramos:

Raza 1	**200**	**185**	**203**	**194**	**320**
Raza 2	**195**	**238**	**242**	**250**	**157**

¿Existen diferencias significativas entre los Pesos medios de ambas razas?

Problema 3

Los siguientes datos (Paul, 1968), corresponden a un estudio llevado a cabo con 1718 personas elegidas al azar en el que se las clasificó según fueran GCC = Grandes Consumidores de café (100 o más tazas al mes) o no lo fueran, NGCC y, según tuvieran una Enfermedad Coronaria, EC, o no la tuvieran, NEC. La tabla obtenida fue la siguiente:

	GCC	NGCC	
EC	**38**	**39**	
NEC	**752**	**889**	
			1718

Analizando estos resultados, ¿cree que existe relación significativa entre las dos variables que forman esta tabla?

Problema 1

Si denominamos *éxito* al suceso tener varón, podemos modelizar el experimento descrito en el enunciado mediante una distribución binomial negativa (EBR-sección 4.4.5) en donde la variable sería $X =$ *número de niñas antes del segundo varón*, siendo la probabilidad de varón $p = 0'5$ y donde la función de masa de esta variable aleatoria viene determinada mediante la expresión:

$$p_X(x) = \binom{n+x-1}{n-1}(1-p)^x p^n \qquad\qquad x = 0, 1, ...$$

siendo en nuestro caso

$$p_X(4) = \binom{2+4-1}{2-1}(1-0'5)^4 0'5^2 = 5 \cdot 0'5^6 = 0'078125.$$

Problema 2

Este problema es muy parecido al Problema 5.4 del texto "Problemas Resueltos de Estadística Básica". El enunciado nos sugiere un contraste de la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$.

Dado que estamos en una situación de comparación de dos poblaciones normales independientes de varianzas desconocidas y muestras pequeñas (EBR-sección 7.6) debemos valorar, en primer lugar, si es razonable suponer iguales o no las varianzas σ_1^2 y σ_2^2 de ambas poblaciones. Por ello, primero contrastaremos $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$. Dicha hipótesis nula se rechazará cuando y sólo cuando (EBR-sección 7.5) sea

$$\frac{S_1^2}{S_2^2} \notin \left[F_{(n_1-1,n_2-1);1-\alpha/2}\ ,\ F_{(n_1-1,n_2-1);\alpha/2}\right]$$

o, equivalentemente (ADD-página 18) cuando el intervalo

$$\left[\frac{S_1^2/S_2^2}{F_{(n_1-1,n_2-1);\alpha/2}}\ ,\ \frac{S_1^2/S_2^2}{F_{(n_1-1,n_2-1);1-\alpha/2}}\right]$$

no contenga al 1.

A partir de los datos del enunciado obtenemos que es $S_1^2 = 3147'3$ y $S_2^2 = 1559'3$; si además fijamos como nivel de significación $\alpha = 0'1$ será $F_{(n_1-1,n_2-1);\alpha/2} = F_{(4,4);0'05} = 6'3883$ y $F_{(n_1-1,n_2-1);1-\alpha/2} = F_{(4,4);0'95} = 1/F_{(4,4);0'05} = 1/6'3883 = 0'1565$, obteniéndose como región de aceptación, para ese nivel de significación, el intervalo

$$I = \left[F_{(n_1-1,n_2-1);1-\alpha/2}\ ,\ F_{(n_1-1,n_2-1);\alpha/2}\right] = [0'1565\ ,\ 6'3883].$$

Como es $S_1^2/S_2^2 = 2'0184 \in I$, aceptaremos H_0 con un p-valor mayor que $0'1$ confirmando la decisión de aceptación.

Equivalentemente, el intervalo región aceptación es

$$\left[\frac{S_1^2/S_2^2}{F_{(n_1-1,n_2-1);\alpha/2}}\ ,\ \frac{S_1^2/S_2^2}{F_{(n_1-1,n_2-1);1-\alpha/2}}\right] = \left[\frac{2'0184}{6'3883}\ ,\ \frac{2'0184}{0'1565}\right] = [0'31595\ ,\ 12'897].$$

Una vez admitida la igualdad de las varianzas poblacionales, podemos realizar ahora el contraste $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ para el caso de dos poblaciones normales de varianzas desconocidas, pero supuestamente iguales, y muestras pequeñas.

En esta situación, se acepta H_0 cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha/2}.$$

Como no nos dan un nivel de significación concreto, vamos a calcular el p-valor del test. Al ser el valor del estadístico de contraste

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|-0'7143|}{22'078} = \frac{4}{30'681} = 0'1304$$

el p-valor será

$$\text{p-valor} = 2 \cdot P\{t_8 > 0'1304\}$$

que a partir de la Tabla 5 de la t de Student es

$$\text{p-valor} > 2 \cdot P\{t_8 > 0'262\} = 2 \cdot 0'4 = 0'8$$

el cual es lo suficientemente grande como para concluir con la no existencia de diferencias significativas entre los pesos medios de ambas razas.

Si hubiéramos utilizado R (cosa imposible en el examen pero que puede ser útil para conocer cómo resolverlo con este paquete en otras ocasiones), hubiéramos ejecutado los siguientes comandos:

```
> x1<-c(200,185,203,194,320)                                    (1)
> x2<-c(195,238,242,250,157)                                    (2)
> var.test(x1,x2,conf.level=0.9)                                (3)

        F test to compare two variances

data:  x1 and x2
```

```
F = 2.0184, num df = 4, denom df = 4, p-value = 0.5131
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
  0.3159568 12.8940457                                                    (4)
sample estimates:
ratio of variances
          2.018406
> t.test(x1,x2,,var.equal=T)                                              (5)

        Two Sample t-test

data:  x1 and x2
t = 0.13037, df = 8, p-value = 0.8995
  (6)                          (7)
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -66.75037  74.75037
sample estimates:
mean of x mean of y
    220.4     216.4
```

En `(1)` y `(2)` incorporamos los datos, en `(3)` ejecutamos el test de igualdad de varianzas cuya región de aceptación, dada en `(4)` contiene al 1 aceptándose dicha igualdad.

Por último, el test de igualdad de medias para varianzas desconocidas pero supuestamente iguales (`var.equal=T`) es ejecutado en `(5)`. El valor del estadístico de contraste es obtenido en `(6)` y el p-valor en `(7)`. Lógicamente, los resultados y conclusiones coinciden con los antes obtenidos.

Problema 3

Se trata de un contraste de independencia de caracteres (EBR-sección 8.2.4) en donde la hipótesis nula es que ambas variables: El consumo de café y la presencia o no de enfermedad coronaria, son independientes y la hipótesis alternativa es que no son independientes.

El estadístico de Pearson toma el valor

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n} = 0'36792.$$

El p-valor del test será:

$$P\{\chi_1^2 > 0'36792\}$$

acotado entre

$$P\{\chi_1^2 > 1'074\} < P\{\chi_1^2 > 0'36792\} < P\{\chi_1^2 > 0'148\}$$

es decir, entre 0′3 y 0′7, en todo caso, suficientemente grande como para aceptar la hipótesis nula de independencia y concluir con que no existe relación significativa entre ambas variables.

Si pudiéramos calcular el valor del estadístico de Pearson con el software R utilizaríamos la siguiente secuencia:

```
> X<-matrix(c(38,752,39,889),ncol=2)                                    (1)
> X
     [,1] [,2]
[1,]   38   39
[2,]  752  889

> chisq.test(X,correct=FALSE)                                           (2)

        Pearson's Chi-squared test

data:  X
X-squared = 0.36792, df = 1, p-value = 0.5441                           (3)

> chisq.test(X)                                                         (4)

        Pearson's Chi-squared test with Yates' continuity correction

data:  X
X-squared = 0.23969, df = 1, p-value = 0.6244                           (5)
```

Primero introducimos los datos según se indica en `(1)`; después se ejecuta `(2)` y, finalmente obtenemos el valor del estadístico de contraste y el p-valor en `(3)`.

Observamos que, como la tabla es 2×2, si no decimos nada, como ocurre en `(4)`, R calcula el valor del estadístico con la corrección de Yates, como obtenemos en `(5)`.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2016-2017

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Se quiere estimar mediante un intervalo de confianza de coeficiente de confianza del 95 %, el tiempo medio en días que tardan los alumnos de un determinado curso a distancia en completar los 10 ejercicios del examen, conociendo por convocatorias anteriores que la varianza es de 300.

Para ello se eligieron al azar 10 alumnos que finalizaron dicho curso obteniéndose los siguientes tiempos en días de finalización del curso:

120 , 179 , 260 , 115 , 222 , 259 , 195 , 200 , 195 , 210

Sabiendo que dichos tiempos siguen una distribución normal, calcular el intervalo de confianza buscado.

Problema 2

Se quiere analizar si puede admitirse que los niveles medios de colesterol en una población determinada se encuentran por debajo de 200 mg/dl. Para ello se tomó una muestra de 50 personas de dicha población que proporcionó una media de 196 mg/dl. y una cuasivarianza muestral igual a 90. Calcule el p-valor del test y diga las conclusiones que obtendría.

Problema 3

Los siguientes datos (Anionwu et al., 1981) corresponden a niveles de hemoglobina en situación estable de diversos pacientes con tres diferentes tipos de enfermedad de célula falciforme:

HB-SS	HB-S/talasemia	HB-SC
7′2	8′1	10′7
7′7	9′2	11′3
8′0	10′0	11′5
8′1	10′4	11′6
8′3	10′6	11′7
8′4	10′9	11′8

Analizar la igualdad de los niveles medios de hemoglobina de los tres tipos de enfermedad con una ANOVA suponiendo que se verifican las condiciones necesarias para que dicho test sea válido. Si no se acepta la hipótesis nula de igualdad ejecutar un test de comparaciones múltiples de Tukey.

Problema 1

Se trata de la determinación del intervalo de confianza para la media de una población normal de varianza conocida $\sigma^2 = 300$, estudiando en EBR-sección 6.2.

Dicho intervalo de confianza tiene como expresión la siguiente:

$$\left[\overline{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]$$

De las tablas de la normal se obtiene que $z_{\alpha/2} = z_{0'025} = 1'96$ y de los datos observados se obtiene una media muestral de $\overline{x} = 195'5$. Por tanto, el intervalo buscado será

$$\left[\overline{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right] = \left[195'5 - 1'96\sqrt{\frac{300}{10}}\ ,\ 195'5 + 1'96\sqrt{\frac{300}{10}}\right] =$$

$$= [184'7646\ ,\ 206'2354].$$

Con la ayuda del paquete estadístico R (o con una calculadora) se podría haber resuelto ejecutando:

```
> x<-c(120,179,260,115,222,259,195,200,195,210)

> mean(x)
[1] 195.5

> qnorm(0.95+0.025)
[1] 1.959964

> 195.5-1.96*sqrt(300/10)
[1] 184.7646
> 195.5+1.96*sqrt(300/10)
[1] 206.2354
```

Problema 2

Del enunciado se desprende que se quiere contrastar la hipótesis nula $H_0 : \mu \geq 200$ frente a la alternativa $H_1 : \mu < 200$. Estamos en un caso de contrastes para la media de una población no necesariamente normal y muestras grandes (EBR-sección 7.3) rechazándose la hipótesis nula cuando y sólo cuando sea

$$\frac{\overline{x} - \mu_0}{S/\sqrt{n}} < z_{1-\alpha}.$$

Dado que no nos dan nivel de significación, vamos a calcular el p-valor del test. El estadístico del contraste toma el valor:

$$\frac{\overline{x}-\mu_0}{S/\sqrt{n}}=\frac{196-200}{\sqrt{90/50}}=-2'98$$

y como es un test unilateral con región crítica la cola de la izquierda, el p-valor será, a partir de una tablas de la distribución normal

$$\text{p-valor}=P\{Z<-2'98\}=P\{Z>2'98\}=0'0014$$

pudiendo rechazarse con bastante seguridad la hipótesis nula y concluir que el nivel medio de colesterol de la población en estudio sí puede establecerse en menos de 200.

Problema 3

Aunque en el examen habrá tenido que resolver el ejercicio con una calculadora, la tabla ANOVA (EBR-sección 9.2) que le saldrá será la dada por `(1)` obteniendo en `(2)` un p-valor tan bajo que podemos concluir con que existen diferencias significativas entre los niveles medios de hemoglobina de los tres tipos de enfermedad.

El test de Tukey de comparaciones múltiples (EBR-sección 9.5) se ejecuta en `(3)` indicando los p-valores de los tres tests, que aparecen en la columna `(4)` que rechacemos las tres hipótesis nulas, lo que indica que existen diferencias estadísticamente significativas entre los tres tratamientos.

```
> HB_SS<-c(7.2,7.7,8.0,8.1,8.3,8.4)

> HB_S_talasemia<-c(8.1,9.2,10.0,10.4,10.6,10.9)

> HB_SC<-c(10.7,11.3,11.5,11.6,11.7,11.8)

> niveles<-c(HB_SS,HB_S_talasemia,HB_SC)

> enfermedad<-factor(rep(LETTERS[1:3],c(6,6,6)))

> problema<-data.frame(enfermedad,niveles)

> result<-aov(niveles~enfermedad,problema)

> summary(result)                                                    (1)
            Df Sum Sq Mean Sq F value   Pr(>F)
enfermedad   2  36.52  18.262   37.83 1.38e-06 ***
                                     (2)
Residuals   15   7.24   0.483
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> TukeyHSD(result)                                                    (3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = niveles ~ enfermedad, data = problema)

$enfermedad
         diff       lwr      upr     p adj
B-A 1.916667 0.8746766 2.958657 0.0006728
C-A 3.483333 2.4413432 4.525323 0.0000009
C-B 1.566667 0.5246766 2.608657 0.0037777
                              (4)
```

_ _

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 41206AD01A01).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada** (2014). Alfonso García Pérez, A. (2014). Editorial UNED, Colección Temática (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica** (1998). Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada** (2008). Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2017-2018

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los siguientes datos (Fernández, 2015) corresponden a diámetros de borde, en centímetros, de fragmentos de dos tipos de cerámica ibérica gris del Cerro de las Nieves (situado en Pedro Muñoz, provincia de Ciudad Real) obtenidos en la campaña de 1986:

Tipo I: 14 , 18 , 20 , 16 , 24 , 18 , 22 , 21 , 18 , 22

Tipo II: 22 , 18 , 16 , 18 , 16 , 30 , 14 , 24 , 12 , 14

Para los 10 restos de Tipo I se pide determinar: La Distribución de Frecuencias Absolutas, el Diagrama de Barras, la Media, la Mediana, la Moda, El Primer Cuartil, El Tercer Cuartil, la Desviación Típica, el Recorrido, y el Coeficiente de Asimetría de Pearson.

Problema 2

Utilizando los datos del problema anterior y suponiendo que las observaciones siguen distribuciones normales independientes, ¿existen diferencias significativas entre las medias de ambos Tipos de restos a un nivel de significación $\alpha = 0'05$? Calcule los p-valores.

Problema 3

Los siguientes datos (Bennett, 1988), corresponden Consumo de Oxígeno X y Ventilación Espirada Y en un experimento en kinesiología,

X	574	592	664	667	718	770	927
Y	21'9	18'6	18'6	19'1	19'2	16'9	18'3

Determinar la recta de regresión de la variable dependiente Y explicada por la variable independiente X y analizar si es significativa para explicar Y en función de X.

Problema 1

La distribución de frecuencias absolutas (EBR-sección 2.3) corresponderá a la de un carácter cuantitativo sin agrupar y será

X_i	n_i
14	1
16	1
18	3
20	1
21	1
22	2
24	1
	10

El Diagrama de Barras es el de la Figura 0.1.

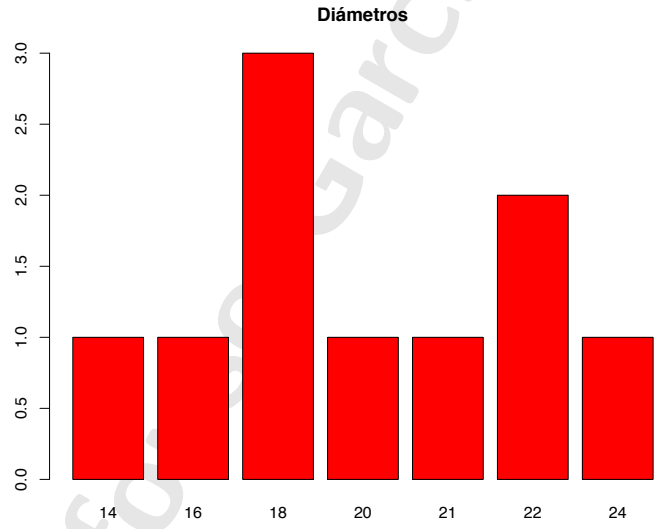


Figura 0.1 : Diagrama de barras

Las Medidas de Posición (EBR-sección 2.3.2) serán, la Media

X_i	n_i	$X_i \cdot n_i$
14	1	14
16	1	16
18	3	54
20	1	20
21	1	21
22	2	44
24	1	24
	10	193

$$\overline{x} = \frac{193}{10} = 19'3.$$

La Moda (el valor más frecuente), $M_d = 18$.

Como la distribución de frecuencias acumuladas es

X_i	n_i	N_i
14	1	1
16	1	2
18	3	5
20	1	6
21	1	7
22	2	9
24	1	10
	10	

será

$$5 \leq 10/2 = 5 < 6$$

con lo que la Mediana corresponderá al valor $(18 + 20)/2 = 19$.

Por otro lado, al ser

$$2 < 10/4 = 2'5 < 5$$

el primer cuartil será $p_{1/4} = 18$ y, al ser

$$7 < 3 \cdot 10/4 = 7'5 < 9$$

el tercer cuartil será $p_{3/4} = 22$.

En cuanto a las medidas de dispersión (EBR-sección 2.3.3), al ser la varianza

$$s^2 = 8'41$$

la desviación típica será $s = \sqrt{s^2} = 2'9$.

El Recorrido será $R = 24 - 14 = 10$.

Por último, el Coeficiente de Asimetría de Pearson, EBR-sección 2.3.4, será

$$A_p = \frac{\overline{x} - M_d}{s} = \frac{19'3 - 18}{2'9} = 0'4482$$

mostrando los datos una cierta asimetría a la derecha, como ya se podía deducir del gráfico de barras.

Aunque en la Prueba Presencial resulta imposible, si hubiéramos resuelto el problema con R hubiéramos ejecutado la siguiente secuencia de instrucciones obteniendo, lógicamente, los mismos resultados. Con `(1)` incorporamos los datos. Con `(2)` formamos la tabla de frecuencias absolutas. Con `(3)` y `(4)` el gráfico de barras. Con `(5)` varias de la medidas solicitadas que son completadas con `(6)`, `(7)`, `(8)` y `(9)` con la ligera diferencia habitual en el tercer cuartil.

```
> TipoI<-c(14,18,20,16,24,18,22,21,18,22)                          (1)

> table(TipoI)                                                     (2)

14 16 18 20 21 22 24
 1  1  3  1  1  2  1

> valores<-c(14,16,18,20,21,22,24)                                 (3)
> frecuencias<-c(1,1,3,1,1,2,1)                                    (3)

> barplot(frecuencias,names=valores,col=2,main="Diámetros")        (4)

> summary(TipoI)                                                   (5)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00   18.00   19.00   19.30   21.75   24.00

> library(modeest)

> mfv(TipoI)                                                       (6)
[1]  18

> sqrt(9*var(TipoI)/10)                                            (7)
[1]  2.9

> (mean(TipoI)-mfv(TipoI))/sqrt(9*var(TipoI)/10)                   (8)
[1]  0.4482759

> range(TipoI)
[1] 14 24

> range(TipoI)[2]-range(TipoI)[1]                                  (9)
[1] 10
```

Problema 2

Estamos ante un caso de comparación de dos poblaciones supuestamente normales independientes y muestras pequeñas, con varianzas desconocidas (EBR-sección 7.6).

Lo primero que necesitamos es averiguar si pueden suponerse iguales o no las varianzas poblacionales para lo que deberemos hacer primero un test de $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$, (EBR-sección 7.5).

Dicha hipótesis nula se rechazará cuando y sólo cuando (EBR-sección 7.5) sea

$$\frac{S_1^2}{S_2^2} \notin \left[F_{(n_1-1,n_2-1);1-\alpha/2}\ ,\ F_{(n_1-1,n_2-1);\alpha/2}\right]$$

o, equivalentemente (ADD-página 18) cuando el intervalo

$$\left[\frac{S_1^2/S_2^2}{F_{(n_1-1,n_2-1);\alpha/2}}\ ,\ \frac{S_1^2/S_2^2}{F_{(n_1-1,n_2-1);1-\alpha/2}}\right]$$

no contenga al 1.

Los dos extremos de la región de aceptación serán, a partir de la Tabla 6 de la F de Snedecor y de EBR-sección 5.3.3

$$F_{(n_1-1,n_2-1);1-\alpha/2} = F_{(9,9);1-0'025} = F_{(9,9);0'975} = 1/F_{(9,9);0'025} = 1/4'026 = 0'2484$$

y

$$F_{(n_1-1,n_2-1);\alpha/2} = F_{(9,9);0'025} = 4'026.$$

A partir de los datos del enunciado obtenemos que el cociente de cuasivarianzas muestrales es

$$S_1^2/S_2^2 = 9'34/30'04 = 0'311 \in [0'2484\,,\ 4'026]$$

con lo que se aceptará la igualdad de las varianzas poblacionales.

Para calcular el p-valor observemos que el valor del estadístico está muy cerca del extremo inferior por lo que utilizaremos éste en dicho cálculo con lo que el p-valor será

$$2 \cdot P\{F_{(9,9)} < 0'311\} = 2 \cdot P\{1/F_{(9,9)} < 0'311\} = 2 \cdot P\{F_{(9,9)} > 1/0'311\} =$$

$$= 2 \cdot P\{F_{(9,9)} > 3'2154\} > 2 \cdot P\{F_{(9,9)} > 4'026\} = 0'05$$

y también,

$$2 \cdot P\{F_{(9,9)} < 0'311\} = 2 \cdot P\{F_{(9,9)} > 3'2154\} < 2 \cdot P\{F_{(9,9)} > 3'1789\} = 0'1.$$

Luego el p-valor estará entre

$$0'05 < \text{p-valor} < 0'1$$

y se aceptaría la hipótesis nula como habíamos hecho.

Ahora ya podemos pasar a contrastar la igualdad de las medias de dos poblaciones normales independientes, muestras pequeñas y varianzas desconocidas pero iguales (EBR-página 220 arriba), $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$.

En esta situación, se acepta H_0 cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha/2}.$$

Al ser

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|19'3 - 18'4|}{\sqrt{\dfrac{9 \cdot 9'34 + 9 \cdot 30'04}{18}}\sqrt{\dfrac{1}{10}+\dfrac{1}{10}}} = \frac{0'9}{4'437342 \cdot 0'44721} = 0'4535.$$

Como, de las tablas de la t de Student (Tabla 5) es

$$t_{n_1+n_2-2;\alpha/2} = t_{18;0'025} = 2'101 > 0'4535$$

se aceptará la igualdad de las medias de ambas poblaciones.

El p-valor será

$$2 \cdot P\{t_{18} > 0'4535\} > 2 \cdot P\{t_{18} > 0'534\} = 2 \cdot 0'3 = 0'6$$

suficientemente grande como para aceptar la hipótesis nula de igualdad de medias.

Por completar el ejercicio, si hubiéramos resuelto el problema con R ejecutaríamos los siguientes comandos en donde hemos marcado con `(1)` los dos p-valores:

```
> TipoI<-c(14,18,20,16,24,18,22,21,18,22)
> TipoII<-c(22,18,16,18,16,30,14,24,12,14)

> var(TipoI)
```

```
[1] 9.344444
> var(TipoII)
[1] 30.04444

> var(TipoI)/var(TipoII)
[1] 0.3110207

> var.test(TipoI,TipoII)

        F test to compare two variances

data:  TipoI and TipoII
F = 0.31102, num df = 9, denom df = 9, p-value = 0.09687
                                                  (1)
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.07725314 1.25216756
sample estimates:
ratio of variances
         0.3110207

> t.test(TipoI,TipoII,var.equal=T)

        Two Sample t-test

data:  TipoI and TipoII
t = 0.45348, df = 18, p-value = 0.6556
                                (1)
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.269623  5.069623
sample estimates:
mean of x mean of y
     19.3      18.4
```

Problema 3

La recta de regresión que se obtiene es (EBR-sección 10.2)

$$Y = 23'904 - 0'00707\ X$$

y uno de los dos posibles tests para analizar su significación consiste en contrastar la hipótesis nula de que es cero el coeficiente de regresión de la variable independiente X. Este test nos da un p-valor igual a $0'187988$ y que sugiere, por tanto, que aceptemos la hipótesis nula de ser cero el coeficiente de regresión asociado a la variable independiente, es decir, podemos concluir con que la recta no es significativa para explicar a la variable dependiente en función de la independiente como vemos en la Figura 0.2.

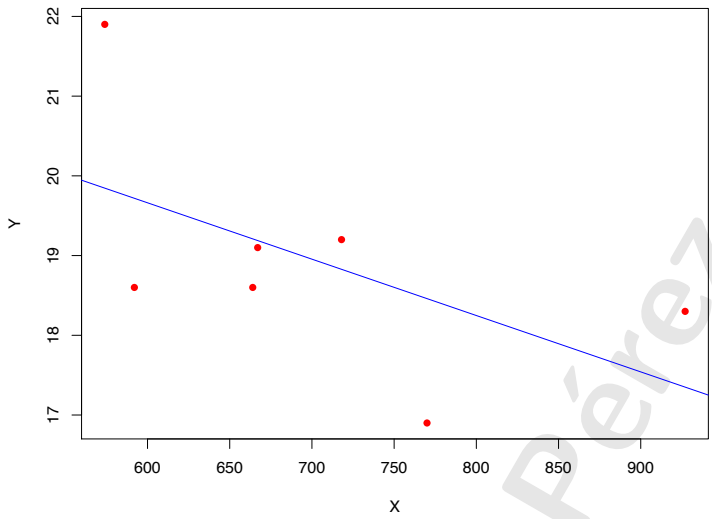


Figura 0.2 : Nube de puntos y recta de regresión

Si hubiéramos utilizado R (cosa imposible en el examen pero que puede ser útil para conocer cómo resolverlo con este paquete en otras ocasiones), hubiéramos ejecutado los siguientes comandos:

```
> X<-c(574,592,664,667,718,770,927)
> Y<-c(21.9,18.6,18.6,19.1,19.2,16.9,18.3)

> recta<-lm(Y~X)

> summary(recta)

Call:
lm(formula = Y ~ X)

Residuals:
       1        2        3        4        5        6        7
 2.05413 -1.11860 -0.60952 -0.08831  0.37229 -1.56004  0.95003

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.904352   3.295998   7.253 0.000778 ***
X           -0.007071   0.004639  -1.524 0.187988
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.366 on 5 degrees of freedom
Multiple R-squared:  0.3172,    Adjusted R-squared:  0.1807
F-statistic: 2.323 on 1 and 5 DF,  p-value: 0.188

> plot(X,Y,pch=16,col=2)
> abline(recta,type="l",col=4)
```

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2017-2018

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los siguientes datos (Pearson y Lee, 1903) son las longitudes en pulgadas de los antebrazos de 15 adultos varones elegidos al azar:

17′3 , 18′4 , 20′9 , 16′8 , 18′7 , 20′5 , 17′9 , 20′9

18′3 , 20′5 , 18′4 , 17′3 , 18′7 , 17′1 , 18′7

Se pide determinar: La Distribución de Frecuencias Absolutas, el Diagrama de Barras, la Media, la Mediana, la Moda, El Primer Cuartil, El Tercer Cuartil, la Desviación Típica, el Recorrido, y el Coeficiente de Asimetría de Pearson.

Problema 2

Se quiere analizar la probabilidad de infarto p en personas de una población homogénea. Para ello se seleccionaron $n = 200$ personas de esa población y se les preguntó si habían padecido alguna vez algún infarto, a lo que respondieron afirmativamente 2 de ellas. ¿Puede admitirse en esa población que es $p < 0'015$?

Problema 3

Se quieren comparar tres métodos de terapia para la mejora de la movilidad en enfermos de Parkinson. Para ello se dividieron al azar 12 enfermos de Parkinson en tres grupos, aplicándose a cada grupo uno de los tres métodos en comparación. Los resultados de mejora en segundos respecto a la movilidad en una prueba determinada, obtenidos por cada paciente, vienen recogidos en la siguiente tabla:

Método I	5′1	4′8	6′4	4′5
Método II	4′4	3′7	3′4	4′5
Método III	3′4	3′9	4′0	3′6

¿Existen diferencias significativas entre los tres Métodos a un nivel de significación de 0′01?

Problema 1

La distribución de frecuencias absolutas (EBR-sección 2.3) corresponderá a la de un carácter cuantitativo sin agrupar y será

X_i	n_i
$16'8$	1
$17'1$	1
$17'3$	2
$17'9$	1
$18'3$	1
$18'4$	2
$18'7$	3
$20'5$	2
$20'9$	2
	15

El Diagrama de Barras es el de la Figura 0.3.

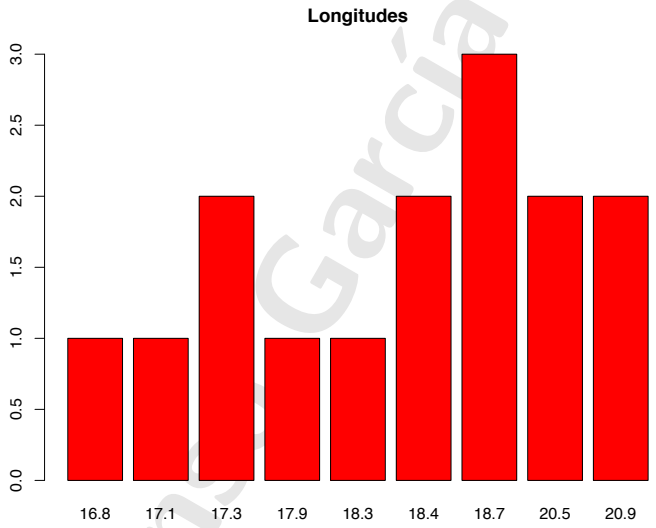


Figura 0.3 : Diagrama de barras

Las Medidas de Posición (EBR-sección 2.3.2) serán, la Media

$$\overline{x} = 18'69333.$$

La Moda (el valor más frecuente), $M_d = 18'7$.

Como la distribución de frecuencias acumuladas es

X_i	n_i	N_i
$16'8$	1	1
$17'1$	1	2
$17'3$	2	4
$17'9$	1	5
$18'3$	1	6
$18'4$	2	8
$18'7$	3	11
$20'5$	2	13
$20'9$	2	15
	15	

será

$$6 \leq 15/2 = 7'5 < 8$$

con lo que la Mediana corresponderá al valor $18'4$.

Por otro lado, al ser

$$2 < 15/4 = 3'75 < 4$$

el primer cuartil será $p_{1/4} = 17'3$ y, al ser

$$11 < 3 \cdot 15/4 = 11'25 < 13$$

el tercer cuartil será $p_{3/4} = 20'5$.

En cuanto a las medidas de dispersión (EBR-sección 2.3.3), al ser la varianza

$$s^2 = 1'815289$$

la desviación típica será $s = \sqrt{s^2} = 1'347327$.

El Recorrido será $R = 20'9 - 16'8 = 4'1$.

Por último, el Coeficiente de Asimetría de Pearson, EBR-sección 2.3.4, será

$$A_p = \frac{\overline{x} - M_d}{s} = \frac{18'69333 - 18'7}{1'347327} = -0'004950543$$

mostrando los datos una cierta asimetría a la izquierda, como ya se podía deducir del gráfico de barras.

Aunque en la Prueba Presencial resulta imposible, si hubiéramos resuelto el problema con R hubiéramos ejecutado la siguiente secuencia de instrucciones obteniendo, lógicamente, los mismos resultados. Con `(1)` incorporamos los datos. Con `(2)` formamos la tabla de frecuencias absolutas. Con `(3)` y `(4)` el

gráfico de barras. Con (5) varias de la medidas solicitadas que son completadas con (6), (7), (8) y (9) con la ligera diferencia habitual en los cuartiles.

```
> Longitudes<-c(17.3,18.4,20.9,16.8,18.7,20.5,17.9,                    (1)
+ 20.9,18.3,20.5,18.4,17.3,18.7,17.1,18.7)                            (1)

> table(Longitudes)                                                    (2)
Longitudes
16.8 17.1 17.3 17.9 18.3 18.4 18.7 20.5 20.9
   1    1    2    1    1    2    3    2    2

> valores<-c(16.8,17.1,17.3,17.9,18.3,18.4,18.7,20.5,20.9)            (3)
> frecuencias<-c(1,1,2,1,1,2,3,2,2)                                   (3)

> barplot(frecuencias,names=valores,col=2,main="Longitudes")           (4)

> summary(Longitudes)                                                  (5)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.80   17.60   18.40   18.69   19.60   20.90

> library(modeest)

> mfv(Longitudes)                                                      (6)
[1]  18.7

> sqrt(14*var(Longitudes)/15)                                          (7)
[1]  1.347327

> (mean(Longitudes)-mfv(Longitudes))/sqrt(14*var(Longitudes)/15)       (8)
[1]  -0.00494807

> range(Longitudes)
[1]  16.8 20.9

> range(Longitudes)[2]-range(Longitudes)[1]                            (9)
[1]  4.1
```

Problema 2

Podemos modelizar la variable en estudio X como una dicotómica que toma el valor 1 cuando el individuo padece infarto y que toma el valor 0 cuando no lo padece. Si p es la probabilidad de infarto, se puede modelizar X con una binomial $B(1, p)$, estando interesados en contrastar si puede admitirse que sea $p < 0'015$ en base a una muestra de X de tamaño $n = 200$.

Como la hipótesis de interés debemos establecerla, si es posible, como hipótesis alternativa, en especial en los tests unilaterales, estamos en un caso de un contraste de hipótesis sobre el parámetro p de la forma $H_0 : p \geq 0'015$

frente a la alternativa $H_1 : p < 0'015$. (EBR-página 207).

Rechazaremos H_0 cuando sea

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{1-\alpha}.$$

Como en el enunciado no se fija un nivel de significación, vamos a calcular el p-valor del test. El estadístico de contraste toma el valor

$$\frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0'01 - 0'015}{\sqrt{\frac{0'015 \cdot 0'985}{200}}} = -0'5817297.$$

El p-valor será, a partir de la Tabla 3 de la distribución normal (ADD-página 33)

$$\text{p-valor} = P\{Z < -0'58\} = P\{Z > 0'58\} = 0'281$$

suficientemente grande como para aceptar la hipótesis nula y concluir con que la probabilidad de infarto en la población de donde se extrajo la muestra no es menor que $0'015$.

Problema 3

La comparación de más de dos poblaciones se efectúa con un Análisis de la Varianza (EBR-sección 9.2) mediante el cual contrastamos la hipótesis nula de que no existen diferencias significativas entre las medias de los tres métodos, $H_0 : \mu_{MI} = \mu_{MII} = \mu_{MIII}$, frente a la alternativa de no ser iguales las tres medias.

Para ello debemos formar la tabla ANOVA, que para los datos de este ejercicio queda igual a

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadísticos
Método	$SST_i = 4'922$	2	$2'4608$	$F = 6'948$
Residual	$SSE = 3'187$	9	$0'3542$	
Total	$SST = 8'109$	11		

Ahora utilizaremos el estadístico de contraste F. Como, a partir de la Tabla 6 de la Adenda vemos que es $F_{(2,9);0'01} = 8'0215 > F = 6'948$, aceptamos H_0, es decir, concluimos que no existen diferencias significativas entre los tres métodos a nivel $\alpha = 0'01$. No obstante, observamos que el p-valor debe de estar muy próximo a $0'01$ por lo que la decisión no es muy fiable.

Si pudiéramos resolver el problema con R ejecutaríamos la siguiente secuencia: ejecutamos el ANOVA mediante `(1)` obteniendo en `(2)` el p-valor.

```
> metodo1<-c(5.1,4.8,6.4,4.5)

> metodo2<-c(4.4,3.7,3.4,4.5)

> metodo3<-c(3.4,3.9,4.0,3.6)

> metodos<-c( metodo1, metodo2, metodo3)

> tiempos<-factor(rep(LETTERS[1:3],c(4,4,4)))

> problema<-data.frame(tiempos,metodos)

> summary(aov(metodos~tiempos,problema))                                (1)
            Df Sum Sq Mean Sq F value Pr(>F)
tiempos      2  4.922  2.4608   6.948  0.015 *
                                        (2)
Residuals    9  3.187  0.3542
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el p-valor $0'015$ es mayor que el nivel de significación $0'01$, aceptaremos la hipótesis nula de que no existen diferencias significativas entre los tres métodos.

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2017-2018

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Calcular la media y la desviación típica de la variable aleatoria cuya función de distribución es la siguiente:

$$F(x) = \begin{cases} 0 & \text{si} \quad x < -2 \\ 1/4 & \text{si} \quad -2 \leq x < 1 \\ 7/12 & \text{si} \quad 1 \leq x < 1'5 \\ 1 & \text{si} \quad x \geq 1'5 \end{cases}$$

Problema 2

Los datos de la siguiente tabla corresponden al Consumo de Helados por persona (durante periodos de 4 semanas), Y y la Temperatura media en grados Fahrenheit, X, (Koteswara Rao Kadiyala, 1970).

X	41	56	63	68	69	65	61
Y	0'386	0'374	0'393	0'425	0'406	0'344	0'327

Analizar mediante una Regresión Lineal si la variable X es o no significativa para explicar a la variable dependiente Y. Dé finalmente la ecuación que permite hacer las predicciones.

Problema 3

Un programa informático se considera eficaz si, en promedio, proporciona una solución a un problema matemático (por supuesto correcta) en menos de 10 segundos. Para analizar si un nuevo programa es eficaz, se eligieron al azar seis problemas matemáticos y se anotaron los siguientes tiempos de resolución en segundos:

$$9'9\ ,\ 10'6\ ,\ 9'7\ ,\ 9'6\ ,\ 10'1\ ,\ 9'8$$

Analizar, a nivel $\alpha = 0'05$ y utilizando el test de rangos signados de Wilcoxon, si el programa resulta eficaz.

Problema 1

Este problema es muy parecido el Problema 2.8 del texto PREB. Al ser una función de distribución *en escalera*, la variable aleatoria que representa, X, es de tipo discreto, tomando esta variable valores en donde F tiene los saltos, que son las abscisas -2, 1 y $1'5$, y con probabilidades igual al valor del salto de F en esos puntos.

Es decir, la *función de masa* de X es

$$p_X(-2) = P\{X = -2\} = F(-2) - F(-2-) = \frac{1}{4} - 0 = \frac{1}{4}.$$

$$p_X(1) = P\{X = 1\} = F(1) - F(1-) = \frac{7}{12} - \frac{1}{4} = \frac{1}{3}.$$

$$p_X(1'5) = P\{X = 1'5\} = F(1'5) - F(1'5-) = 1 - \frac{7}{12} = \frac{5}{12}.$$

La *media* o *esperanza* de X, al ser de tipo discreto, es igual a la suma de los valores que toma por las probabilidades con que los toma; formalmente,

$$\mu_X = E[X] = \sum_x x\, p_X(x) = -2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{3} + 1'5 \cdot \frac{5}{12} = \frac{5'5}{12} = 0'458.$$

Para calcular la *desviación típica* de X, previamente calcularemos su *varianza*,

$$V(X) = \sum_x x^2\, p_X(x) - \mu_X^2 = (-2)^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{3} + 1'5^2 \cdot \frac{5}{12} - 0'458^2 = 2'061$$

con lo que la desviación típica será igual a

$$D(X) = \sqrt{V(X)} = \sqrt{2'061} = 1'4356.$$

Problema 2

La recta de regresión que se obtiene es (EBR-sección 10.2)

$$Y = 0'3453432 + 0'0005617\ X$$

y uno de los dos posibles tests para analizar su significación consiste en contrastar la hipótesis nula de que es cero el coeficiente de regresión de la variable independiente X. Este test nos da un p-valor igual a $0'7355$ y que sugiere, por tanto, que aceptemos la hipótesis nula de ser cero el coeficiente de regresión asociado a la variable independiente, es decir, podemos concluir con que la

recta no es significativa para explicar a la variable dependiente en función de la independiente como vemos en la Figura 0.4, en donde se aprecia un dato anómalo, el par $(41, 0'386)$.

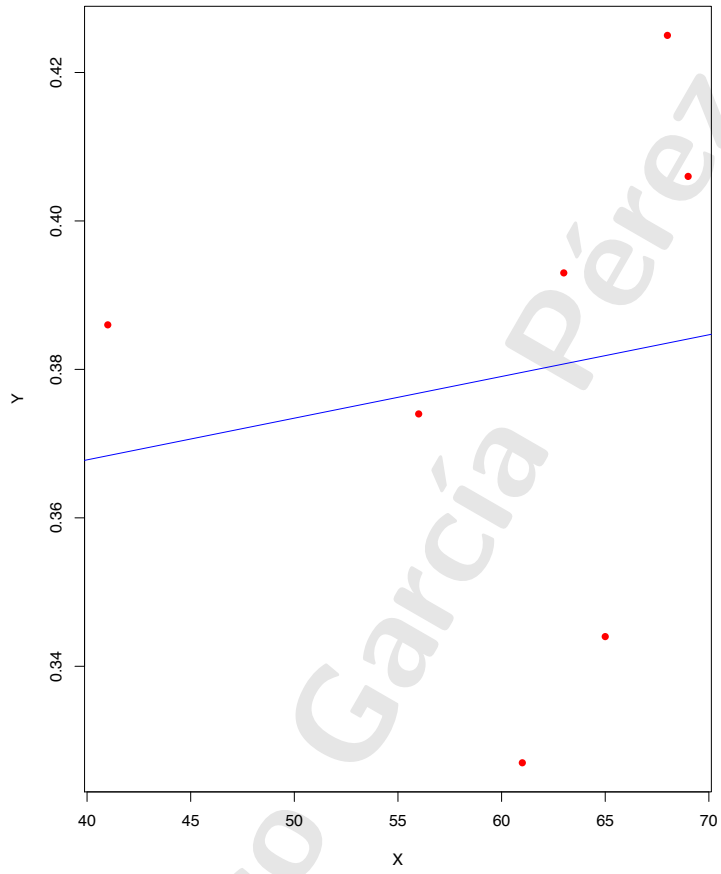


Figura 0.4 : Nube de puntos y recta de regresión

Si hubiéramos utilizado R (cosa imposible en el examen pero que puede ser útil para conocer cómo resolverlo con este paquete en otras ocasiones), hubiéramos ejecutado los siguientes comandos:

```
> X<-c(41,56,63,68,69,65,61)
> Y<-c(0.386,0.374,0.393,0.425,0.406,0.344,0.327)

> recta<-lm(Y~X)
```

```
> summary(recta)

Call:
lm(formula = Y ~ X)

Residuals:
        1         2         3         4         5         6         7
 0.017627 -0.002798  0.012270  0.041461  0.021900 -0.037853 -0.052607

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3453432  0.0960329   3.596   0.0156 *
X           0.0005617  0.0015722   0.357   0.7355
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03706 on 5 degrees of freedom
Multiple R-squared:  0.02489,   Adjusted R-squared:  -0.1701
F-statistic: 0.1276 on 1 and 5 DF,  p-value: 0.7355

> plot(X,Y,pch=16,col=2)
> abline(recta,type="l",col=4)
```

Problema 3

Este problema es muy parecido el Problema 8.3 del texto EEA. Se nos solicita analizar la hipótesis de que el promedio de la variable *Tiempo de resolución de un problema matemático*, es menor que 10. Como nos piden contrastar esta hipótesis utilizando el test de los rangos signados de Wilcoxon (EBR-sección 8.3.2), la formalizaremos utilizando la mediana de dicha variable, contrastando la hipótesis nula $H_0 : M \geq 10$ frente a la alternativa $H_1 : M < 10$.

El estadístico de este contraste es

$$T^+ = \sum_{i=1}^{n} z_i \, r(|D_i|)$$

Para el cálculo de T^+ necesitamos calcular primero las diferencias $D_i = X_i - 10$, luego sus valores absolutos $|D_i|$, los rangos de estos valores absolutos (es decir, el lugar que ocupan entre los seis valores), $r(|D_i|)$, siendo por último, T^+ la suma de los $r(|D_i|)$ cuyos D_i sean positivos. Estos cálculos aparecen en la siguiente tabla

X_i	$9'9$	$10'6$	$9'7$	$9'6$	$10'1$	$9'8$
D_i	$-0'1$	$0'6$	$-0'3$	$-0'4$	$0'1$	$-0'2$
$\lvert D_i\rvert$	$0'1$	$0'6$	$0'3$	$0'4$	$0'1$	$0'2$
$r(\lvert D_i\rvert)$	*1'5*	*6*	*4*	*5*	*1'5*	*3*

Observamos que al ser $|D_1| = 0'1 = |D_5|$ y ocupar el valor $0'1$ los lugares primero y segundo, asignamos a ambos el rango promedio $(1+2)/2 = 1'5$. El valor de T^+ será, por tanto,

$$T^+ = 6 + 1'5 = 7'5.$$

Ahora, rechazaremos $H_0 : M \geq 10$ frente a $H_1 : M < 10$, cuando y sólo cuando sea

$$T^+ \leq \frac{n(n+1)}{2} - \tau_{n;\alpha}$$

con $\tau_{n;\alpha}$ el menor número entero tal que

$$P\{T^+ \geq \tau_{n;\alpha}\} \leq \alpha$$

o más claro, un $\tau_{n;\alpha}$ el menor número entero tal que

$$P\{T^+ > \tau_{n;\alpha} - 1\} \leq \alpha.$$

La Tabla 14 de la Addenda nos dice que para un tamaño muestral $n = 6$ y un nivel de significación $\alpha = 0'05$ es

$$P\{T^+ > 18\} \leq 0'05$$

por lo que deberá ser $\tau_{n;\alpha} - 1 = 18$ y $\tau_{n;\alpha} = 19$.

Por tanto, al ser $T^+ = 7'5 > 21 - 19 = 2$, aceptaremos H_0, concluyendo con que el programa informático no es eficaz.

De la Tabla 14 del final del texto obtenemos que el p-valor es mayor que $0'1$, confirmando la decisión tomada de aceptación de la hipótesis nula.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 41206AD01A01).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada** (2014). Alfonso García Pérez, A. (2014). Editorial UNED, Colección Temática (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica** (1998). Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada** (2008). Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

Bennett, G.W. (1988). Determination of anaerobic threshold. *Canadian Journal of Statistics*, **16**, 307-310.

Fernández Martínez, V.M. (2015). *Arqueo-Estadística. Métodos Cuantitativos en Arqueología*. Alianza Editorial.

Koteswara Rao Kadiyala (1970). Testing for the independence of regression disturbances. *Econometrica*, **38**, 97-117.

Pearson, K. y Lee, A. (1903). On the laws of inheritance in man. I. Inheritance of physical characters. *Biometrika*, **2**, 357-462.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Primera semana. Curso 2018-2019

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los siguientes datos corresponden a valores de Transaminasas Alanina Amino Transferasa (ALT), en unidades por litro, en la sangre de 10 mujeres seleccionadas al azar.

16 , 25 , 39 , 33 , 35 , 10 , 35 , 35 , 33 , 26

Se pide determinar: La Distribución de Frecuencias Absolutas, el Diagrama de Barras, la Media, la Mediana, la Moda, el Primer Cuartil, el Tercer Cuartil, la Desviación Típica, el Recorrido, y el Coeficiente de Asimetría de Pearson.

Problema 2

Los datos que aparecen a continuación son porcentajes de proteínas contenidos en una muestra de trigo molido de tamaño $n = 10$, obtenida mediante el método de medición de Kjeldahl (Fearn, 1983). Determinar un intervalo de confianza de coeficiente de confianza del 95 % para la varianza de dicha variable, suponiendo la normalidad de los datos.

$9'23$, $8'01$, $10'95$, $11'67$, $10'41$, $9'51$, $8'67$, $7'75$, $8'05$, $11'39$

Problema 3

Se quiere analizar si existe realmente un incremento significativo de temperatura en el planeta para lo que se eligieron al azar 10 lugares L en los que se midió la temperatura en un día determinado y, en ese mismo lugar y día transcurridos exactamente 5 años. Los resultados obtenidos en grados centígrados fueron los siguientes:

	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$	$L7$	$L8$	$L9$	$L10$
Temp. inicial	12	22	32	38	22	10	9	29	22	15
Temp. tras 5 años	14	23	31	40	27	15	11	38	22	14

Determine, mediante el test de los rangos signados de Wilcoxon, si puede concluirse que hay un incremento significativo de la temperatura a nivel de significación $\alpha = 0'05$.

Problema 1

La distribución de frecuencias absolutas (EBR-sección 2.3) corresponderá a la de un carácter cuantitativo sin agrupar y será

X_i	n_i
10	1
16	1
25	1
26	1
33	2
35	3
39	1
	10

El Diagrama de Barras es el de la Figura 0.1.

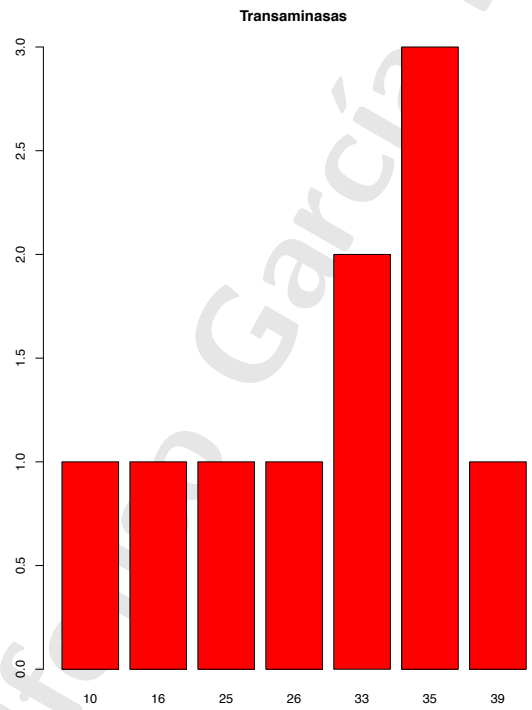


Figura 0.1 : Diagrama de barras

Las Medidas de Posición (EBR-sección 2.3.2) serán, la Media

$$\overline{x} = \frac{287}{10} = 28'7.$$

La Moda (el valor más frecuente), $M_d = 35$.

Como la distribución de frecuencias acumuladas es

X_i	n_i	N_i
10	1	1
16	1	2
25	1	3
26	1	4
33	2	6
35	3	9
39	1	10
	10	

será

$$4 \leq 10/2 = 5 < 6$$

con lo que la Mediana corresponderá al valor 33.

Por otro lado, al ser

$$2 < 10/4 = 2'5 < 3$$

el primer cuartil será $p_{1/4} = 25$ y, al ser

$$6 < 3 \cdot 10/4 = 7'5 < 9$$

el tercer cuartil será $p_{3/4} = 35$.

En cuanto a las medidas de dispersión (EBR-sección 2.3.3), al ser la varianza

$$s^2 = 79'41$$

la desviación típica será $s = \sqrt{s^2} = 8'91$.

El Recorrido será $R = 39 - 10 = 29$.

Por último, el Coeficiente de Asimetría de Pearson, EBR-sección 2.3.4, será

$$A_p = \frac{\overline{x} - M_d}{s} = \frac{28'7 - 35}{8'91} = -0'707$$

mostrando los datos una cierta asimetría a la izquierda, como ya se podía deducir del gráfico de barras.

Aunque en la Prueba Presencial resulta imposible, si hubiéramos resuelto el problema con R hubiéramos ejecutado la siguiente secuencia de instrucciones

obteniendo, lógicamente, los mismos resultados. Con `(1)` incorporamos los datos. Con `(2)` formamos la tabla de frecuencias absolutas. Con `(3)` y `(4)` el gráfico de barras. Con `(5)` varias de la medidas solicitadas que son completadas con `(6)`, `(7)`, `(8)` y `(9)` con la ligera diferencia habitual en los cuartiles.

```
> ALT<-c(16,25,39,33,35,10,35,35,33,26)                                    (1)

> table(ALT)                                                               (2)

10 16 25 26 33 35 39
 1  1  1  1  2  3  1

> valores<-c(10,16,25,26,33,35,39)                                         (3)
> frecuencias<-c(1,1,1,1,2,3,1)                                            (3)

> barplot(frecuencias,names=valores,col=2,main="Transaminasas")            (4)

> summary(ALT)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   25.25   33.00   28.70   35.00   39.00                            (5)

> library(modeest)

> mfv(ALT)                                                                 (6)
[1]  35

> sqrt(9*var(ALT)/10)                                                      (7)
[1] 8.911229

> (mean(ALT)-mfv(ALT))/sqrt(9*var(ALT)/10)                                 (8)
[1] -0.7069732

> range(ALT)
[1] 10 39

> range(ALT)[2]-range(ALT)[1]                                              (9)
[1]  29
```

Problema 2

Si denominamos X al porcentaje de proteínas, conocemos intervalos de confianza para la varianza de X en el caso de que esta variable se distribuya según una normal, EBR-sección 6.4 ó ID-sección 3.4, siendo en ese caso dicho intervalo (la media es desconocida):

$$I = \left[\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} \, , \; \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \right].$$

Para los datos del enunciado es,

$$I = \left[\frac{9 \cdot 2'1464}{\chi^2_{9;0'025}} \, , \; \frac{9 \cdot 2'1464}{\chi^2_{9;0'975}} \right] = \left[\frac{19'3176}{19'02} \, , \; \frac{19'3176}{2'7} \right] = \left[1'015647 \, , \; 7'154667 \right]$$

ya que es $S^2 = 2'1464$ y, por la Tabla 4 de ADD de la distribución χ^2 de Pearson, $\chi^2_{9;0'025} = 19'02$ y $\chi^2_{9;0'975} = 2'7$.

Problema 3

Se trata de comparar dos poblaciones, pero como los datos de temperatura se refieren a un mismo lugar en dos ocasiones, T_1 y T_2, las observaciones serán dependientes; es decir, no se trata de dos conjuntos de temperaturas independientes, sino que hacen referencia a un mismo lugar por lo que se trata de un problema de Datos Apareados. Por tanto, en primer lugar, calcularemos los valores de la variable diferencia $D = T_2 - T_1$

	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$	$L7$	$L8$	$L9$	$L10$
D_i	2	1	-1	2	5	5	2	9	0	-1

estando interesados en contrastar si puede admitirse que la mediana M_D de esta variable diferencia es positiva. Es decir, contrastaremos las hipótesis $H_0 : M_D \leq 0$ frente a la alternativa $H_1 : M_D > 0$, utilizando el test de los Rangos Signados de Wilcoxon (EBR-sección 8.3.2).

Ya para empezar, vemos que una diferencia es cero y, en ese caso, se indica que debemos reducir el tamaño de la muestra a $n = 9$ eliminando esta observación.

El estadístico del contraste es T^+ =*suma de los rangos de las diferencias positivas*.

De los datos obtenemos la siguiente tabla:

D_i	2	1	-1	2	5	5	2	9	-1
$\lvert D_i \rvert$	2	1	1	2	5	5	2	9	1
$r(\lvert D_i \rvert)$	5	2	2	5	$7'5$	$7'5$	5	9	2

con lo que el valor del estadístico de rangos signados de Wilcoxon, suma de los rangos de las diferencias positivas, será igual a

$$T^+ = \sum_{i=1}^{n} z_i \, r(|D_i|) = 5 + 2 + 5 + 7'5 + 7'5 + 5 + 9 = 41.$$

Mirando la Tabla 14 de ADD vemos el punto crítico para un nivel de significación $\alpha = 0'05$ es $t_{0'05} = 36$. Como el estadístico es mayor que el punto

crítico, $T^+ = 41 > 36 = t_\alpha$, debemos rechazar la hipótesis nula y concluir que, efectivamente, parece haber un calentamiento global.

El p-valor es $P\{T^+ \geq 41\}$ que, mirando la Tabla 14 de ADD es igual a $0'01$, suficientemente pequeño como para confirmar la decisión de rechazo de la hipótesis nula.

ESTADÍSTICA BÁSICA

Prueba Presencial de Febrero. Segunda semana. Curso 2018-2019

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Se supone que una variable aleatoria X asociada a un determinado experimento aleatorio sigue una distribución uniforme de parámetros $(-1, 2)$. Determinar su función de distribución y calcular las siguientes probabilidades:

(a) $P\{X < 0\}$.

(b) $P\{|X| < 1'3\}$.

(c) $P\{|1 - X| < 1'9\}$.

Problema 2

Los datos que aparecen a continuación son concentraciones en gramos por decilitro (g/dl) de hemoglobina Hbg en la sangre de 10 individuos, seleccionados al azar en la sección de pintura de una fábrica de coches (Royston, 1983).

$$13'4\ ,\ 15\ ,\ 16'4\ ,\ 15'5\ ,\ 14'8\ ,\ 14'7\ ,\ 15'1\ ,\ 16\ ,\ 14'8\ ,\ 14'5$$

Supuesto que dichas concentraciones siguen una distribución normal, a nivel de significación $\alpha = 0'05$, ¿cabe admitir un nivel medio de concentración de hemoglobina significativamente menor que el considerado saludable, que es de 15'5 g/dl?

Problema 3

Los datos que aparecen a continuación corresponden al Porcentaje de grasa corporal a diferentes valores de Edad en Hombres elegidos al azar, Mazess et al. (1984).

Edad	23	27	27	45
Porcentaje	9'5	7'8	17'8	27'4

Analizar la Regresión Lineal Simple de la variable dependiente, Porcentaje, en función de la independiente, Edad

Problema 1

La distribución de X es de tipo continuo, siendo función de densidad (EBR-sección 4.5.2)

$$f(x) = \frac{1}{2+1} = \frac{1}{3}$$

si es $-1 \leq x \leq 2$.

Por tanto, su función de distribución será (EBR-página 109)

$$F(x) = P\{X \leq x\} = \int_{-1}^{x} \frac{1}{3}\, dy = \frac{x+1}{3}$$

si es $-1 \leq x \leq 2$.

Si fuera $x < -1$ sería $F(x) = 0$ y, si fuera $x > 2$ sería $F(x) = 1$.

Las probabilidades pedidas serán, por tanto
(a)

$$P\{X < 0\} = F(0) = \frac{1}{3}.$$

(b)

$$P\{|X| < 1'3\} = P\{-1'3 < X < 1'3\} = F(1'3) - F(-1'3) = F(1'3) = \frac{1'3+1}{3} = \frac{2'3}{3}.$$

(c)

$$P\{|1-X| < 1'9\} = P\{|X-1| < 1'9\} = P\{-1'9 < X-1 < 1'9\} = P\{-0'9 < X < 2'9\} =$$

$$= F(2'9) - F(-0'9) = 1 - \frac{-0'9+1}{3} = \frac{2'9}{3}.$$

Problema 2

Si representamos por μ a la media de la variable Concentración de Hemoglobina en la población antes muestreada, la hipótesis que queremos contrastar será $\mu < 15'5$ que deberá ir como hipótesis alternativa por dos razones: uno, porque esta hipótesis es en la que estamos interesados y, dos, porque formalmente el "igual" no está incluido en esta hipótesis.

Por tanto, queremos contrastar la hipótesis nula $H_0 : \mu \geq 15'5$ frente a la hipótesis alternativa, $H_1 : \mu < 15'5$.

Estamos ante un test de hipótesis para la media μ de una población normal, EBR-sección 7.2, con varianza poblacional desconocida. En este caso, rechazaremos H_0 cuando y sólo cuando sea

$$\frac{\overline{x}-\mu_0}{S/\sqrt{n}} < t_{n-1;1-\alpha}.$$

Como es

$$\frac{\overline{x}-\mu_0}{S/\sqrt{n}} = \frac{15'02-15'5}{0'8297255/\sqrt{10}} = -1'829392$$

y, a partir de las tablas de la t de Student, ADD-Tabla 5, es $t_{n-1;1-\alpha} = t_{9;0'95} = -1'833 < -1'829392$, deberemos aceptar la hipótesis nula y concluir que los empleados de la sección de pintura de la fabrica no tienen un nivel de concentración de hemoglobina menor de lo saludable.

No obstante, vemos que el p-valor es aproximadamente $0'05$, muy dudoso para obtener conclusiones claras.

Aunque en la Prueba Presencial no se podía resolver el problema con R, por completar, su resolución con este paquete estadístico sería,

```
> x<-c(13.4,15,16.4,15.5,14.8,14.7,15.1,16,14.8,14.5)

> t.test(x,alternative="less",mu=15.5)

        One Sample t-test

data:  x
t = -1.8294, df = 9, p-value = 0.0503
alternative hypothesis: true mean is less than 15.5
95 percent confidence interval:
     -Inf 15.50098
sample estimates:
mean of x
    15.02
```

Problema 3

La recta de regresión es (EBR-sección 10.2)

$$\text{Porcentaje} = -9'1576 + 0'8125\,\text{Edad}$$

Uno de los dos posibles tests para analizar su significación consiste en contrastar la hipótesis nula de que es cero el coeficiente de regresión de la variable independiente, Edad. Este test nos da un p-valor igual a $0'109$ y que sugiere, por tanto, que aceptemos la hipótesis nula de ser cero el coeficiente de regresión asociado a la variable independiente, es decir, podemos concluir con que la recta no es significativa para explicar a la variable dependiente en función de la independiente.

Si hubiéramos utilizado R (cosa imposible en el examen pero que puede ser útil para conocer cómo resolverlo con este paquete en otras ocasiones), hubiéramos ejecutado los siguientes comandos:

```
> Porcentaje<-c(9.5,7.8,17.8,27.4)
> Edad<-c(23,27,27,45)

> ajuste1<-lm(Porcentaje~Edad)
> ajuste1
Call:
lm(formula = Porcentaje ~ Edad)

Coefficients:
(Intercept)         Edad
    -9.1576       0.8125
```

La recta de regresión para los hombres será, por tanto,

$$\text{Porcentaje} = -9'1576 + 0'8125\,\text{Edad}$$

El análisis de su significación lo podemos obtener ejecutando

```
> summary(ajuste1)

Call:
lm(formula = Porcentaje ~ Edad)

Residuals:
        1         2         3         4
-0.030928 -4.981100  5.018900 -0.006873

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.1576     9.2829  -0.987    0.428
Edad          0.8125     0.2931   2.772    0.109

Residual standard error: 5 on 2 degrees of freedom
Multiple R-squared:  0.7935,    Adjusted R-squared:  0.6902
F-statistic: 7.685 on 1 and 2 DF,  p-value: 0.1092
```

ESTADÍSTICA BÁSICA

Prueba Presencial de Septiembre. Curso 2018-2019

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

4) Los tres problemas puntúan lo mismo.

Problema 1

Los siguientes datos son los precios en dólares en 1961, de 8 botellas de Seagram's 7 Crown Whisky elegidas al azar en tiendas del Grupo I, correspondientes a estados americanos en donde había monopolio, y a 8 botellas del mismo licor en tiendas de estados del Grupo II, en donde las tiendas de licores eran de propiedad privada (Fuente revista Chance, 1991, volumen 4, número 1).

Grupo I	**4′11**	**4′15**	**4′20**	**4′55**	**3′80**	**4′00**	**4′19**	**4′75**
Grupo II	**4′89**	**4′95**	**4′55**	**4′90**	**5′25**	**5′30**	**4′29**	**4′85**

¿Se puede concluir que existen diferencias significativas entre los precios medios de ambos Grupos, admitiendo que los precios siguen distribuciones normales independientes?

Problema 2

Los datos que siguen, Shaw (1942), corresponden al número de icebergs observados en 1920, según el mes que se indica, al sur de Terranova (Canada), y en los Grandes Bancos (meseta submarina de la plataforma continental frente a la costa sudeste de Terranova en donde se encuentran la cálida corriente del Golfo y la fría corriente de Labrador),

	Mes											
	E	**F**	**M**	**A**	**M**	**Jn**	**Jl**	**A**	**S**	**O**	**N**	**D**
Terranova	**3**	**10**	**36**	**83**	**130**	**68**	**25**	**13**	**9**	**4**	**3**	**2**
G. Bancos	**0**	**1**	**4**	**9**	**18**	**13**	**3**	**2**	**1**	**0**	**0**	**0**

En base a estos datos, ¿existen o no diferencias significativas entre los avistamientos de icebergs desde uno y otro lugar?

Problema 3

Los siguientes datos corresponde a longitudes de aves elegidas al azar, de tres especies geográficamente aisladas:

Ave	Longitud			
Gorrión Molinero	13	12′5	14	12′5
Herrerillo Capuchino	12′5	11′5	10′5	11
Jilguero Común	12′5	13	13′5	12′8

A la vista de estos datos, ¿puede inferirse que existen diferencias significativas entre los tres tipos de aves, a nivel de significación $\alpha = 0'05$?

Problema 1

Estamos ante una situación de contraste para la diferencias de medias de dos poblaciones normales independientes, muestras pequeñas, con varianzas desconocidas (EBR-sección 7.6), por lo que debemos valorar primero si las varianzas, aunque desconocidas, pueden considerarse iguales o no. Para ello contrastaremos la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ frente a la alternativa de ser distintas (EBR-sección 7.5), contraste basado en el estadístico S_1^2/S_2^2. De hecho, aceptaremos esta hipótesis nula cuando y sólo cuando sea,

$$\frac{S_1^2}{S_2^2} \in [\ F_{n_1-1,n_2-1;1-\frac{\alpha}{2}}\ ,\ F_{n_1-1,n_2-1;\frac{\alpha}{2}}\ ].$$

A partir del enunciado se obtiene que es $\overline{x}_1 = 4'21875$, $S_1^2 = 0'0904125$, $\overline{x}_2 = 4'8725$, $S_2^2 = 0'1108786$.

Como es $S_1^2/S_2^2 = 0'8154191$, si consideramos un nivel de significación $\alpha = 0'1$, será, a partir de la Tabla 6 de la F de Snedecor, $F_{7,7;1-0'05} = 1/F_{7,7;0'05} = 1/3'787 = 0'264$, con lo que la región de aceptación, a nivel $\alpha = 0'1$, es $[0'264\ ,\ 3'787]$, contendrá al valor del estadístico y se aceptará la hipótesis nula de ser iguales ambas varianzas poblacionales, a ese nivel suficientemente alto, lo que lleva a que el p-valor también será alto y, por tanto, tendremos bastante confianza en la decisión de aceptación de la hipótesis nula.

Aunque en el examen es imposible, si hubiéramos podido resolver este apartado con R, con las siguientes sentencias obtenemos las medias y cuasivarianzas muestrales, así como el valor del estadístico del contraste S_1^2/S_2^2,

```
> x1<-c(4.11,4.15,4.20,4.55,3.80,4,4.19,4.75)
> x2<-c(4.89,4.95,4.55,4.9,5.25,5.30,4.29,4.85)
> mean(x1)
[1] 4.21875
> mean(x2)
[1] 4.8725
> var(x1)
[1] 0.0904125
> var(x2)
[1] 0.1108786
> var(x1)/var(x2)
[1] 0.8154191
```

De hecho, con R podemos obtener el p-valor ejecutando `(1)`

```
> 2*pf(0.8154191,7,7)                                    (1)
[1] 0.7946475
```

Si quisiéramos ejecutar este test directamente con R deberíamos ejecutar `(2)`, (EAR-sección 4.2.3), observando que aquí se analiza si la región de aceptación,

$$\left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}}\ ,\ \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;1-\alpha/2}}\right] = \left[\frac{0'8154191}{3'787}\ ,\ \frac{0'8154191}{0'264}\right] = [0'2153\,,\ 3'0887]$$

cociente contiene o no al 1. La región de aceptación se observa en (3) y el p-valor de este test, igual lógicamente al anterior, aparece en (4).

```
> var.test(x1,x2,conf.level=0.9)                                        (2)

        F test to compare two variances

data:  x1 and x2
F = 0.81542, num df = 7, denom df = 7, p-value = 0.7946                 (4)
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
 0.2153181 3.0880275                                                    (3)
sample estimates:
ratio of variances
         0.8154191
```

Apuntamos que, intercambiando los papeles de ambas poblaciones (que es lo que nos dice la ortodoxia, EBR-sección 7.5), hubiéramos obtenido las mismas conclusiones.

Por tanto, el test para contrastar la igualdad de las medias poblacionales; es decir, para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$ será el que acepte H_0 cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha/2}$$

Como es

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|4'21875 - 4'8725|}{\sqrt{\dfrac{7\cdot 0'0904125 + 7\cdot 0'1108786}{14}}\sqrt{\dfrac{1}{8}+\dfrac{1}{8}}} = 4'1214$$

y, a partir de la Tabla 5 de la t de Student, vemos que el p-valor del test es

$$2\cdot P\{t_{14} > 4'1214\} < 2\cdot P\{t_{14} > 3'326\} = 2\cdot 0'0025 = 0'005$$

suficientemente pequeño como para rechazar la hipótesis nula de igualdad en los precios de ambos Grupos.

Este test de igualdad de medias se puede resolver con R ejecutando (5) (véase EBR-sección 7.6), en donde indicamos que consideramos las varianzas

poblacionales como iguales. Como el 0 no está incluido en la región de aceptación dada en (6), rechazamos la hipótesis nula de igualdad de los niveles medios de ambas poblaciones. El p-valor $0'001038$ aparece en (7) e indica el rechazo de H_0.

```
> t.test(x1,x2,var.equal=T)                                              (5)

        Two Sample t-test

data:  x1 and x2
t = -4.1214, df = 14, p-value = 0.001038                                 (7)
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9939634 -0.3135366                                                   (6)
sample estimates:
mean of x mean of y
  4.21875   4.87250
```

Problema 2

Se trata de un contraste de homogeneidad de varias muestras, porque los datos de la tabla son recuentos de observaciones (EBR-sección 8.2.3), en donde la hipótesis nula es que las poblaciones de donde se obtuvieron ambas muestras son homogéneas y la hipótesis alternativa es que existen diferencias significativas entre ellas.

En principio, el estadístico de Pearson tomaría el valor

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n} = 3'7164$$

siendo el p-valor del test,

$$P\{\chi^2_{11} > 3'7164\}.$$

Utilizando la Tabla 4 de la distribución χ^2, el p-valor queda acotado por

$$P\{\chi^2_{11} > 3'7164\} > P\{\chi^2_{11} > 3'816\} = 0'975$$

es decir, mayor que $0'975$, suficientemente grande como para aceptar que no existen diferencias significativas entre los avistamientos desde uno u otro lugar.

No obstante, la tabla de frecuencias esperadas es

	E	F	M	A	M	Jn	Jl	A	S	O	N	D
Terranova	2.65	9.72	35.33	81.26	130.73	71.55	24.73	13.25	8.83	3.53	2.65	1.77
G. Bancos	0.35	1.28	4.67	10.74	17.27	9.45	3.27	1.75	1.17	0.47	0.35	0.23

apareciendo celdillas con frecuencias esperadas menores que 5, por lo que deberíamos agrupar columnas contiguas o utilizar la corrección de Yates.

Si agrupamos clases contiguas, uniendo las tres primeras columnas y las 6 últimas, la tabla de doble entrada será

	Mes				
	E a M	A	M	Jn	Jl a D
Terranova	49	83	130	68	56
G. Bancos	5	9	18	13	6

el estadístico de Pearson toma el valor

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n} = 2'4029$$

siendo el p-valor del test,

$$P\{\chi^2_4 > 2'4029\}.$$

Utilizando la Tabla 4 de la distribución χ^2, quedará acotado por

$$P\{\chi^2_4 > 4'878\} < P\{\chi^2_4 > 2'4029\} < P\{\chi^2_4 > 2'195\}$$

es decir,

$$0'3 < P\{\chi^2_4 > 2'4029\} < 0'7$$

muy cercano a $0'7$. En todo caso, mayor que $0'3$, suficientemente grande como para aceptar que no existen diferencias significativas entre los avistamientos desde uno u otro lugar.

Aunque no es posible en el examen, si quisiéramos ejecutar el test con el software R utilizaríamos la siguiente secuencia:

```
> X<-matrix(c(3,10,36,83,130,68,25,13,9,4,3,2,0,1,4,9,18,13,3,2,1,0,0,0),
+ ncol=12,byrow=T)
> colnames(X)<-c("E","F","M","A","M","Jn","Jl","A","S","O","N","D")
> rownames(X)<-c("Terranova","G. Bancos")

> X
          E  F  M  A  M Jn Jl  A S O N D
Terranova 3 10 36 83 130 68 25 13 9 4 3 2
G. Bancos 0  1  4  9  18 13  3  2 1 0 0 0

> chisq.test(X,correct=T)                                          (1)

        Pearson's Chi-squared test
```

```
data:  X
X-squared = 3.7164, df = 11, p-value = 0.9775                          (2)

Warning message:
In chisq.test(X, correct = T) : Chi-squared approximation may be incorrect
```

Primero introducimos los datos según se indica, después se ejecuta el test mediante `(1)` utilizando la corrección de Yates y, finalmente, obtenemos el valor del estadístico de contraste y el p-valor en `(2)`.

Le hemos pedido en `(1)` que nos ejecute la corrección de Yates porque algunas frecuencias esperadas son menores que 5:

```
> round(chisq.test(X)$expected,2)
             E    F     M     A      M    Jn    Jl    A    S    O    N    D
Terranova 2.65 9.72 35.33 81.26 130.73 71.55 24.73 13.25 8.83 3.53 2.65 1.77
G. Bancos 0.35 1.28  4.67 10.74  17.27  9.45  3.27  1.75 1.17 0.47 0.35 0.23
```

Si nos ayudamos de R en la ejecución del test agrupando clases contiguas, ejecutaríamos `(3` y `(4)`, obteniendo el p-valor en `(5)`. La tabla final de frecuencias esperadas muestra celdillas con valores mayores que 5,

```
> X2<-matrix(c(49,83,130,68,56,5,9,18,13,6),ncol=5,byrow=T)              (3)
> colnames(X2)<-c("E a M","A","M","Jn","Jl a D")
> rownames(X)<-c("Terranova","G. Bancos")
> chisq.test(X2,correct=F)                                               (4)

        Pearson's Chi-squared test

data:  X2
X-squared = 2.4029, df = 4, p-value = 0.6621                           (5)

> chisq.test(X2,correct=F)$expected
         E a M        A        M        Jn    Jl a D
[1,] 47.697941 81.26316 130.72769 71.546911 54.764302
[2,]  6.302059 10.73684  17.27231  9.453089  7.235698
```

Problema 3

Se trata de un *Análisis de la Varianza para un factor en un diseño completamente aleatorizado*, cuyos fundamentos y desarrollos teóricos aparecen en EBR-sección 9.2, con el que se quiere contrastar la hipótesis nula de igualdad de la longitud media de las tres especies de ave $H_0 : \mu_A = \mu_B = \mu_C$, frente a la alternativa de no ser las tres iguales.

Como en todos los contrastes de este tipo, lo primero que debemos determinar es la tabla de Análisis de la Varianza, la cual es

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Aves	$SST_i = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{T^2}{n}$	$r-1$	$\frac{SST_i}{r-1}$	$\frac{SST_i/(r-1)}{SSE/(n-r)}$
Residual	$SSE = SST - SST_i$	$n-r$	$\frac{SSE}{n-r}$	
Total	$SST = \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}$	$n-1$		

que, para los datos de nuestro problema resulta ser igual a

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Aves	$SST_i = 6'832$	2	$3'416$	$F = 7'289$
Residual	$SSE = 4'217$	9	$0'469$	
Total	$SST = 11'049$	11		

El estadístico F tiene, si es cierta la hipótesis nula de igualdad de los efectos medios de las longitudes, una distribución F de Snedecor con grados de libertad igual al par formado por los grados de libertad correspondientes a las fuentes de variación *Aves* y *Residual*, antes determinados, $(r-1, n-r) = (2, 9)$, por lo que para determinar el punto crítico, al nivel de significación requerido en el enunciado, $\alpha = 0'05$, buscaremos en la tabla de la F de Snedecor (Tabla 6) el valor $F_{(2,9);0'05} = 4'2563$. Al ser $F = 7'289$ mayor que dicho punto crítico, se rechaza H_0 a ese nivel de significación, concluyendo con la existencia de diferencias significativas entre las tres poblaciones de aves.

De dicha tabla también se obtiene una acotación del p-valor:

$$\text{p-valor} = P\{F_{(2,9)} > 7'289\} < P\{F_{(2,9)} > 5'7147\} = 0'025.$$

Aunque no es posible en el examen, para resolver este ejercicio con R, EBR-sección 9.3, incorporaríamos los datos ejecutando las tres siguientes sentencias,

```
> longitudes<-c(13,12.5,14,12.5,12.5,11.5,10.5,11,12.5,13,13.5,12.8)
> aves<-factor(rep(LETTERS[1:3],c(4,4,4)))
```

```
> datos<-data.frame(aves,longitudes)
```

para obtener la tabla ANOVA ejecutamos `(1)`

```
> summary(aov(longitudes~aves,datos))                                    (1)
            Df Sum Sq Mean Sq F value Pr(>F)
aves         2  6.832   3.416   7.289 0.0131 *                           (2)
Residuals    9  4.217   0.469
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al final de la fila `(2)` se observa un p-valor suficientemente bajo como para concluir con el rechazo de la igualdad de las tres poblaciones de ave.

– –

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 41206AD01A01).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada** (2014). Alfonso García Pérez, A. (2014). Editorial UNED, Colección Temática (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica** (1998). Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada** (2008). Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

Fearn, T. (1983). A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Applied Statistics*, **32**, 73-79.

Mazess, R.B., Peppler, W.W. y Gibbons, M. (1984). Total body composition by dual-photon (^{1}53Gd) absorptiometry. *American Journal of Clinical Nutrition*, **40**, 834-839.

Royston, J.P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Applied Statistics*, **32**, 121-133.

Shaw, N. (1942). *Manual of meteorology.* Vol. 2, London: Cambridge University Press.