Curso 2018-2019

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

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.

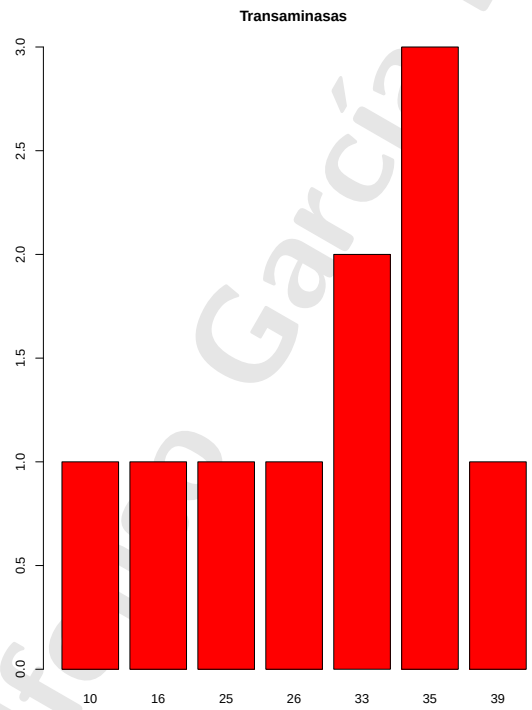

Figura 0.1 : Diagrama de barras

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.