Examen Final de Estadística I, 20 de mayo de 2011.

Grados ADE, DER-ADE.

NORMAS: 1) Entregar cada problema en un cuadernillo distinto, aunque esté en blanco.
2) Realizar los cálculos con al menos dos cifras decimales significativas.
3) No se podrá abandonar el examen hasta transcurridos 30 minutos depués de haber empezado.
4) No está permitido salir del aula sin entregar el examen, aunque esté en blanco.

Problema 1 (3 puntos)
Parte A La siguiente tabla de contingencia contiene datos sobre el gasto mensual en llamadas telefónicas (en euros) y el sexo de 500 encuestados:

Gasto	Hombres	Mujeres
$[0,6)$	59	76
$[6,12)$	42	62
$[12,30)$	54	51
$[30,60)$	67	39
$[60,120]$	28	22

a) Indicar el tipo de las variables que intervienen (cualitativa nominal/ordinal, cuantitativa continua/discreta). Hallar sus distribuciones marginales absolutas.

b) Justifica si es cierto que el gasto mensual medio de los hombres es superior al de las mujeres.

c) De los encuestados cuyo gasto mensual es inferior a 12 euros, ¿qué porcentaje son hombres?

Solución Ejercicio 1 Parte A:

a) La variable X =“Gasto mensual en llamadas telefónicas (en euros)” es cuantitativa continua, mientras que Y =“sexo del encuestado” es cualitativa nominal (binaria). Sus distribuciones marginales absolutas son:

X	$n_{i\bullet}$
$[0,6)$	135
$[6,12)$	104
$[12,30)$	105
$[30,60)$	106
$[60,120]$	50

Y	$n_{\bullet j}$
hombre	250
mujer	250

b) Deben compararse las medias de las distribuciones de X/Y ='hombre' y de X/Y ='mujer'. Empezamos especificando las distribuciones condicionadas:

$X/Y = hombre$	x_i	n_{i1}	$x_i\, n_{i1}$
$[0,6)$	3	59	177
$[6,12)$	9	42	378
$[12,30)$	21	54	1134
$[30,60)$	45	67	3015
$[60,120]$	90	28	2520
Total		250	7224

$X/Y = mujer$	x_i	n_{i2}	$x_i\, n_{i2}$
$[0,6)$	3	76	228
$[6,12)$	9	62	558
$[12,30)$	21	51	1071
$[30,60)$	45	39	1755
$[60,120]$	90	22	1980
Total		250	5592

El gasto medio para los hombres encuestados es de $\overline{x}_{|Y=hombre} = \frac{7224}{250} = 28.896$ euros, mientras que en las mujeres $\overline{x}_{|Y=mujer} = \frac{5592}{250} = 22.368$ euros. Luego es cierto que el gasto mensual medio de los hombre es superior al de las mujeres.

c) De los encuestados que cumplen la condición $X < 12$, hay que determinar qué porcentaje de ellos son hombres. Hay un total de $101 + 138 = 239$ encuestados que cumplen esta condición. De ellos, 101 son hombres, es decir el $101/239 \times 100\% = 42.259\% \approx 42.26\%$ son hombres.

$X < 12$	Hombres	Mujeres
$[0, 6)$	59	76
$[6, 12)$	42	62
Total	101	138

Parte B Se recoge una muestra del 1000 precios de hoteles en toda Europa. Teniendo en cuenta la siguiente salida de R Commander:

```
numSummary(precios.hoteles,statistics=c("mean","sd","quantiles"))
      mean        sd        0%       25%      50%      75%      100%     n
  147.9907  92.81217  50.15713  77.41145  116.7258  189.3803  523.5089  1000
```

d) ¿Qué se puede decir sobre la asimetría de la distribución? Dibuja un diagrama de caja.

e) Razona qué medida de centralidad crees que debería usarse para estos datos y di cuál es su valor.

f) ¿Qué medida de variabilidad aparece en la salida de R Commander? Propón otra medida alternativa e indica su valor.

Solución Ejercicio 1 Parte B:

d) Puesto que la media aritmética (147.9907) es muy superior a la mediana (116.7258) la distribución presenta asimetría positiva (o hacia la derecha).

La caja del diagrama de caja debe estra formada por los cuartiles inferior (77.41145) y superior (189.3803). La mediana debe situarse en el punto 116.7258, mientras que la media aritmética está fuera (hacia la derecha) de la caja (147.9907). La primera barrera exterior derecha está situada en el punto $189.3803 + 1.5 \times (189.3803 - 77.41145) = 357.33$ y la segunda barrera exterior derecha, en $189.3803 + 3 \times (189.3803 - 77.41145) = 525.29$. Puesto que el máximo de los datos es 523.5089, podemos concluir que se trata de un dato atípico (aunque no atípico extremo) al encontrarse entre las dos barreras exteriores derechas. Por otro lado, la primera barrera exterior izquierda estaría en $77.41145 - 1.5 \times (189.3803 - 77.41145) = -34.55$, aunque carece de sentido calcularla pues no hay datos negativos (el mínimo es 50.15713). Finalmente, los "bigotes" de la caja se alargarían, el izquierdo hasta 50.15713, mientras que el derecho llegaría hasta el dato anterior a 357.33.

e) Como hemos visto en (d), esta distribución es asimétrica y presenta datos atípicos. Por tanto, como medida de *centralidad* es mejor utilizar la mediana (que vale 116.7258) en lugar de la media aritmética.

f) La medida de *variabilidad* que aparece en R Commander es la desviación típica, pero ésta también se ve afectada por la presencia de datos atípicos. Sería más razonable utilizar el Rango Intercuartílico, que vale $RIC = 189.3803 - 77.41145 = 111.9689$.

Problema 2 (4 puntos) En *Estadística I*, el 4% de los estudiantes obtuvo una nota final de "10". De éstos, el 80% ya había obtenido un "10" en el parcial. Por otro lado, de los estudiantes que no obtuvieron una nota final de "10", el 5% de ellos sí que obtuvo un "10" en el parcial.

a) ¿Cuál es la probabilidad de que un estudiante escogido al azar obtuviera un "10" en el parcial?

b) Si un estudiante obtuvo un "10" en el parcial, ¿cuál es la probabilidad de que su nota final sea de "10"?

c) Se seleccionan 100 estudiantes con una nota final de "10" y se cuenta cuántos de ellos no obtuvieron un "10" en el parcial. Especifica la distribución de esta variable aletaoria, que llamaremos X, indicando su nombre y los valores de sus parámetros. Calcula la esperanza y desviación típica de X.

d) En la misma situación del apartado c), calcula (aproximadamente) la probabilidad de que más de 25 estudiantes no obtuvieran un "10" en el parcial. ¿Qué teorema permite realizar esta aproximación?

Solución Ejercicio 2: Let "$F10$" denote an event that a randomly selected student of *Est I* obtains a final grade of "10", and "$M10$" that a randomly selected student of *Est I* obtains a midterm grade of "10", and "$\overline{F10}$" and "$\overline{M10}$" the complements of those respectively.

a) We have

$$\begin{aligned} P(M10) &= P(M10|F10)P(F10) + P(M10|\overline{F10})P(\overline{F10}) \\ &= 0.8(0.04) + 0.05(0.96) = 0.08 \end{aligned}$$

b) Based on the Bayes' Theorem and the previous part

$$\begin{aligned} P(F10|M10) &= \frac{P(M10|F10)P(F10)}{P(M10)} \\ &= \frac{0.8(0.04)}{0.08} = 0.4 \end{aligned}$$

c) $X \sim Binomial(n = 100, p = 0.20)$ ("success" = failing to obtain a midterm grade of "10") hence

$$\begin{aligned} E[X] &= np = 20 \\ V[X] &= np(1-p) = 16 \\ SD[X] &= 4 \end{aligned}$$

d)

$$\begin{aligned} P(X > 25) &\approx P\left(Z > \frac{25-20}{4}\right) \\ &= P(Z > 1.25) = 0.1056 \end{aligned}$$

The approximation is based on the Central Limit Theorem.

COMMENTS:

- If the answer in 1) is incorrect, do not penalize for it in 2) only in 1).
- In 3): name for 0.2, parameters 0.2+0.2, expectation 0.2 and SD for 0.2.
- In 4): writing the probability for 0.2, standarization for 0.2, use of the table for 0.4, mentioning the CLT 0.2.

Problema 3 (3 puntos) Se dispone de una muestra de tamaño $n = 10$ de rendimientos de acciones de cierto sector:

rendimientos (en %)	7.3	5.6	8.3	8.3	6.3	6.7	7.2	8.2	7.5	7.7

Suponiendo que los rendimientos tienen una distribución normal con desviación típica 1,

a) Calcula un intervalo de confianza al 95% para el rendimiento medio poblacional.

b) Determina la desviación típica de la media muestral de rendimientos. ¿Qué ocurriría con el intervalo de confianza si el tamaño muestral fuera de $n = 100$? Razona tu respuesta.

c) Suponiendo que la media poblacional de rendimientos fuera igual a 6.5, calcula la probabilidad de que la media muestral de rendimientos (para $n = 10$) sea inferior a 6.

Solución Ejercicio 3 Consideramos la v.a. $X =$"rendimiento de acciones en cierto sector" y suponemos que tiene una ley $N(\mu, 1)$. Consideramos una muestra $X_1, X_2, \ldots, X_n$ de v.a. i.i.d. con la misma ley que X.

a) $IC(\mu)_{0.95} = [\overline{X} \mp z_{\alpha/2}\,\sigma/\sqrt{n}] = [7.31 \mp (1.96)\,1/\sqrt{10}] = [6.69, 7.93]$

b) La media muestral $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ tiene ley $N(\mu, \frac{\sigma}{\sqrt{n}})$. Por tanto, para $n = 10$ la desviación típica de $\overline{X}$ es $1/\sqrt{10} = 0.3162$.
Puesto que la amplitud del IC es $2\,z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$, si el tamaño muestral aumentara a $n = 100$, entonces suponiendo el mismo nivel de confianza, la amplitud del IC disminuiría. También es posible que al calcular la media muestral sobre esta nueva muestra de tamaño $n = 100$ el valor fuera diferente de 7.31, con lo que el nuevo IC estaría centrado en otro punto.

c) Hay que calcular la probabilidad $P(\overline{X} < 6)$ suponiendo que $\overline{X} \sim N(6.5, 1/\sqrt{10})$.
$P(\overline{X} < 6) = P\left(\frac{\overline{X}-6.5}{1/\sqrt{10}} < \frac{6-6.5}{1/\sqrt{10}}\right) = P(Z < -1.58) = 0.0571$, donde la variable $Z = \frac{\overline{X}-6.5}{1/\sqrt{10}} \sim N(0,1)$.