Primer PBL. 26 de Febrero de 2009

Estadística I. GDO-ADE 73, 74, 75, 79 y 80

Alumnos (por orden alfabético): Grupo pequeño n^o:

1.

2.

3.

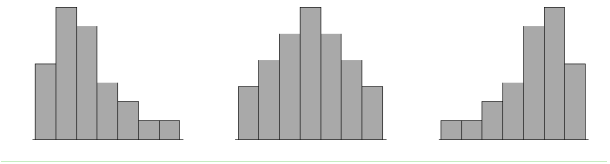
Puntuación sobre 7.5 puntos. Tiempo de realización: 1 hora
Todas las respuestas finales y los cálculos intermedios deben escribirse en esta misma hoja en los espacios reservados para ello. Expresad el resultado final con 2 decimales.

1. **(2 pts.)** La siguiente salida de R corresponde al resumen numérico de dos variables cuantitativas continuas, X e Y.

```
> numSummary(Datos[,"X"], statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
     mean       sd   0%  25%  50%  75% 100%   n NA
 3.528782 1.761974 1.19 1.95 3.07 4.98    8 197 10

> numSummary(Datos[,"Y"], statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
     mean       sd 0% 25% 50% 75% 100%   n NA
 43.47761 38.75604  2  12  30  66  169 201  6
```

a) **(0.75 pts.)** De los tres siguientes representaciones gráficas, ¿cuál representaría mejor la distribución de Y? ¿Qué nombre recibe ese tipo de forma de distribución? Razonad la respuesta.



Solución. *La mediana de Y, $Me_Y = 30$, es bastante menor que su media, $\overline{y} = 43.48$, lo cual indica que la distribución de Y es asimétrica a la derecha. Es decir, la forma de Y sería aproximadamente la del primer histograma.*

b) **(0.5 pts.)** Calculad el rango y el rango intercuartílico de Y.

Solución. *Rango de Y: $R_Y = y_{(n)} - y_{(1)} = 169 - 2 = 167$.*
Rango intercuartílico de Y: $RI_Y = Q_{3y} - Q_{1y} = 66 - 12 = 54$.

c) **(0.75 pts.)** ¿Cuál de las dos variables es más dispersa? Razonad la respuesta.

Solución. *Para poder comparar la dispersión de dos variables distintas necesitamos utilizar una medida de dispersión relativa como es el coeficiente de variación:*

$$CV_x = \frac{s_x}{\overline{x}} = \frac{1.761974}{3.528782} = 0.50 \qquad CV_y = \frac{s_y}{\overline{y}} = \frac{38.75604}{43.47761} = 0.89$$

El coeficiente de variación de Y es mucho mayor que el de X, y por tanto Y es más dispersa que X.

2. **(1.5 pts.)** Los siguientes datos se refieren a la duración (en semanas), X, de un producto de limpieza en un hogar de cuatro personas:

7.4	4.5	8.0	9.5	8.4	8.2	7.8	8.2	8.9	7.5	7.6	8.1
8.5	7.5	8.9	7.6	8.9	9.9	7.1	7.7	5.5	8.5	7.7	8.7

a) **(0.5 pts.)** Calculad la mediana de este conjunto de datos.

Solución. *Primero ordenamos los datos de menor a mayor:*

4.5	5.5	7.1	7.4	7.5	7.5	7.6	7.6	7.7	7.7	7.8	8.0
8.1	8.2	8.2	8.4	8.5	8.5	8.7	8.9	8.9	8.9	9.5	9.9

Como el número de datos, 24, es par, la mediana será $\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$*, es decir, la media de los dos valores centrales. En nuestro caso:* $Me_x = \frac{8.0 + 8.1}{2} = 8.05$.

b) **(0.25 pts.)** Completad la siguiente tabla con la distribución de frecuencias absolutas de X.

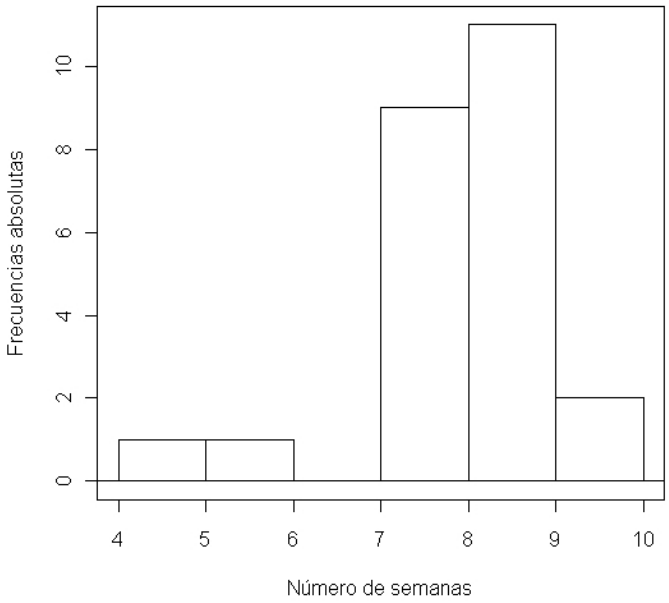
Solución.

Clases	*Frecuencias absolutas*
$[4, 5)$	*1*
$[5, 6)$	*1*
$[6, 7)$	*0*
$[7, 8)$	*9*
$[8, 9)$	*11*
$[9, 10)$	*2*

c) **(0.75 pts.)** Utilizando el resultado del apartado anterior, realizad en esta cuadrilla un gráfico que represente la distribución de frecuencias de X. Indicad claramente qué se representa en cada eje así como las unidades de división de la cuadrilla. ¿Qué nombre recibe este tipo de gráfico?

Solución.

Este tipo de gráfico recibe el nombre de histograma*, y representa la distribución de frecuencias de una variable cuantitativa continua.*



3. **(1.25 pts.)** En la siguiente tabla aparece la distribución conjunta de frecuencias relativas de las variables X, que representa el número de miembros de una familia, e Y, que representa el número de veces a la semana que las familias van a la compra. Los datos han sido obtenidos a partir de una muestra de 200 familias de dos o más miembros.

# compras \ # miembros :	2	3	4	5	6
1	0.12	0.10	0.12	0.14	0.13
2	0.03	0.04	0.06	0.08	0.07
3	0.01	0.00	0.02	0.03	0.05

a) **(0.25 pts.)** ¿Cuántas familias de cinco miembros hacen la compra una vez a la semana?

Solución. *La frecuencia absoluta es igual a la frecuencia relativa multiplicada por el número de observaciones en la muestra.*
$0.14 \times 200 = 28$ *familias de cinco miembros hacen la compra una vez a la semana.*

b) **(1 pto.)** Calcula la distribución del número de miembros en la familia para familias que realizan tres compras semanales.

Solución.

$X_{\|Y=3}$	*2*	*3*	*4*	*5*	*6*	*Total*
$f_{x_i\|Y=3}$	*0.09*	*0.00*	*0.18*	*0.27*	*0.46*	*1*

4. **(2.75 pts.)** Con el objetivo de comparar ocho marcas de producto lavajillas, se quiere estudiar la relación que hay entre el precio del producto y el poder de limpieza que tiene. Para ello se ha realizado la siguiente prueba: en condiciones estándar se ha procedido al lavado de una serie de platos con los diferentes productos y se ha contado el número de platos que se pueden lavar con 10 ml de producto de cada una de las marcas. Los datos obtenidos aparecen en la siguiente tabla junto con el precio de 100 ml de producto de cada marca.

X: Precio (euros/100 ml)	1.30	1.27	1.23	1.24	1.40	1.36	1.18	1.38
Y: Nº de platos lavados con 10 ml	26	25	22	32	33	29	15	33

a) **(0.5 pts.)** Sabiendo que el número medio de platos lavados con 10 ml de producto lavavajillas es 26.88 y que $\sum_{i=1}^{8} y_i^2 = 6053$, calculad la varianza de Y.

Solución.

$$s_y^2 = \frac{1}{8}\sum_{i=1}^{8} y_i^2 - \overline{y}^2 = \frac{6053}{8} - 26.88^2 = 34.09$$

b) **(0.75 pts.)** A continuación se muestran la salida de R con el modelo de regresión ajustado y el diagrama de dispersión para estas dos variables.

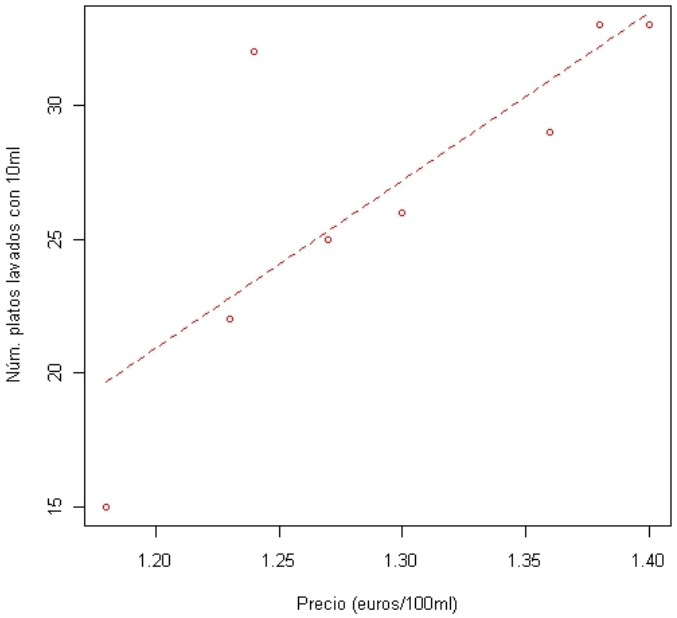
```
Call:
lm(formula = y ~ x, data = pbl)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6347 -1.3842 -0.6342 -0.0324  8.5877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -54.66      25.62  -2.133   0.0768 .
x              62.96      19.75   3.188   0.0189 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.124 on 6 degrees of freedom
Multiple R-squared: 0.6287,Adjusted R-squared: 0.5668
F-statistic: 10.16 on 1 and 6 DF,  p-value: 0.01889
```

Los coeficientes de la recta de regresión vienen recogidos en la salida de R en la columna *Estimate* del apartado *Coefficients* como *Intercept* (ordenada en el origen) y x (pendiente).



Escribid la ecuación de la recta de regresión de y sobre x y trazadla sobre el diagrama de dispersión anterior.

Solución. *La ecuación de la recta de regresión es:* $y = -54.66 + 62.96x$.

c) **(0.75 pts.)** Calculad el coeficiente de correlación de x e y e interpretadlo en referencia al tipo de relación existente entre las dos variables.

Solución. *El coeficiente de correlación es la raíz del coeficiente de determinación. El coeficiente de determinación viene recogido en la salida de R como* Multiple R-squared. *Por tanto:*

$$r_{(x,y)} = \sqrt{r^2_{(x,y)}} = \sqrt{0.6287} = 0.79.$$

Como es un valor relativamente alto y positivo, podemos decir que existe una relación lineal moderada o fuerte y positiva entre x e y. Lo podemos corroborar a la vista del diagrama de dispersión, donde en efecto los datos parecen tener una tendencia lineal.

d) **(0.75 pts.)** A partir del modelo de regresión ajustado, ¿cuál sería el "poder de lavado"de un producto con un precio de 1.35 euros/100 ml? ¿y el de un producto con un prcio de 0.70 euros/100ml?

Solución. *En el primer caso:* $y(1.35) = -54.66 + 62.96 \times 1.35 = 30.33$. *Es decir, con 10 ml de un producto de 1.35 euros/100 ml esperamos lavar unos 30 platos.*
En el segundo caso no podemos hacer la predicción, ya que 0.70 euros/100 ml no está dentro del rango de valores de la variable x usados para construir el modelo de regresión. De hecho, si hiciéramos lo mismo que en el caso anterior nos saldría un número de platos negativo.