Fundamentos de Estadística

UOC

Apuntes

PEC 0-1: TEORÍA + R

1. INTRODUCCIÓN A LA ESTADÍSTICA

Estadística:

Población: conjunto de todos los individuos o unidades que son objeto principal del estudio.
Muestra: subconjunto de unidades o individuos de la población de estudio.
Individuo: cada uno de los elementos que componen una población o muestra, ya sean personas, objetos...
Parámetro: aquel valor que resume una determinada información referente a la población.
Estadístico: aquel valor que expresa una determinada información referente a una muestra.

Variable: Característica observable que cambia entre los elementos de una población. Característica que puede ser medida o contada y que puede adoptar más de un valor.

Variables cualitativas o categóricas: son aquellas que recogen una característica que no se puede expresar mediante una cantidad, aunque si con una categoría. Ej.: Sexo (varón/mujer), color de ojos (azules/marrones/verdes...), etc. (datos no métricos)

- **Nominales:** Son aquellas variables que tienen un conjunto de categorías que no tienen ninguna tipo de jerarquía.
- **Ordinales:** Son aquellas variables donde el conjunto de categorías tienen una jerarquía u orden.

Variables cuantitativas o numéricas: aquellas variables que recogen como información una medida (una cantidad numérica) de lo que se está observando. Ej.: Edad, peso, tensión arterial, número de hijos, hermanos, etc. (datos métricos).

- **Discretas:** Son aquellas variables que tienen un conjunto finito de valores y en caso infinito solamente toman los números enteros
- **Continuas:** Son aquellas variables donde el conjunto de posibles valores entre dos números fijos es infinito.

2. ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

TABLA DE FRECUENCIAS

Frecuencia absoluta (n_i): número de veces que se repite un determinado valor. Puede ser individual o acumulada (N_i) (en les variables nominales no tiene sentido acumular).

Frecuencia relativa (f_i): proporción que representa las apariciones de ese valor respecto al total $f_r = \frac{n_i}{n}$

También puede ser individual o acumulada (F_i). Si se multiplica por 100 nos dará un porcentaje.

Ejemplo numérico discreto, número de hijos por familia: 2, 1, 1, 0, 1, 2, 2, 0, 1, 1.

x_i	n_i	N_i	f_i	F_i
0				
1				
2				

Ejemplo R Commander:

```
counts:
Hijos
0 1 2
2 5 3
```

```
percentages:
Hijos
0  1  2
20 50 30
```

```
$breaks
0 2 4 6 8 10
$counts
1 2 3 2 2 2
```

GRÁFICOS

Diagrama de sectores

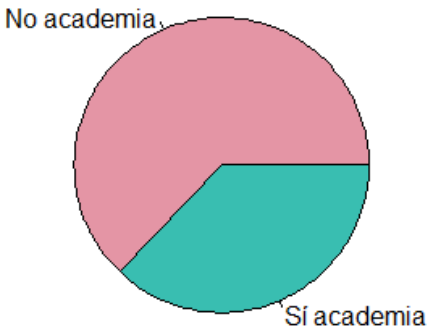
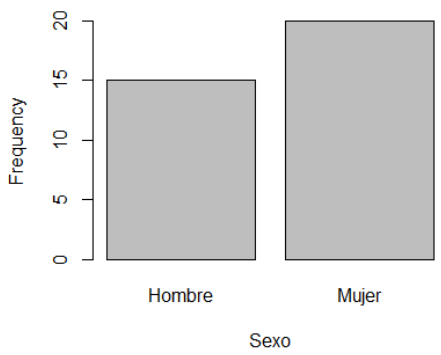
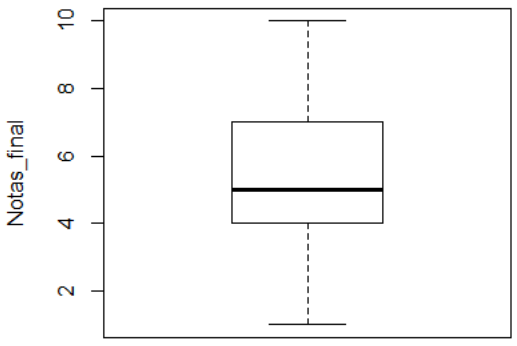


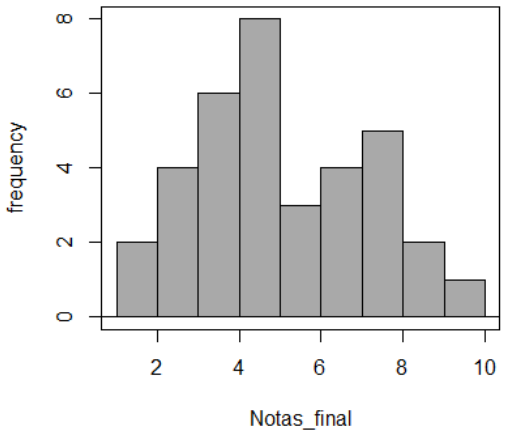
Diagrama de barras



Boxplot



Histograma



MEDIDAS DE CENTRALIZACIÓN

Media aritmética:

$$\bar{x} = \frac{\sum x_i}{n}$$

Mediana:

$$Pos_{Md} = \frac{n+1}{2}$$

Moda:

MEDIDAS DE DISPERSIÓN

Recorrido o rango:

$$R = Vmàx - Vmín \qquad R_{Intercuartílico} = Q3 - Q1$$

Varianza:

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n}$$

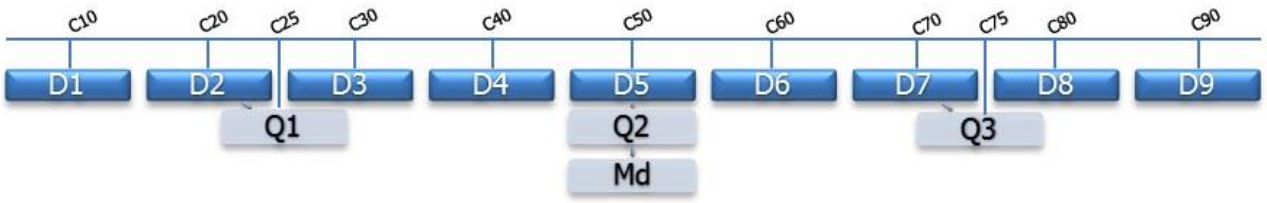
Desviación típica:

$$S = \sqrt{S^2}$$

Coeficiente de variación:

$$CV = \frac{S}{\bar{x}} \cdot 100$$

MEDIDAS DE POSICIÓN



ASIMETRÍA

S>0 asimetría positiva

S=0 simetría

S<0 asimetría negativa

TRANSFORMACIONES

Ejemplo R Commander:

```
     mean       sd IQR        cv  skewness 0% 25% 50% 75% 100%  n
 5.514286 2.160766   3 0.3918487 0.1370465  1   4   5   7   10 35
```

3. ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

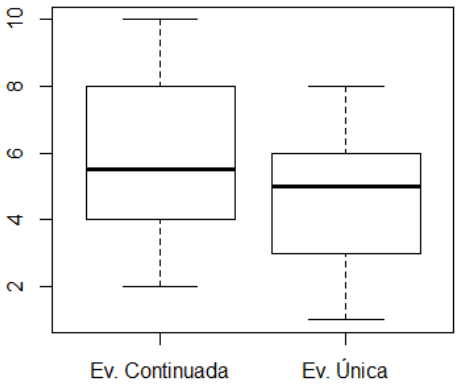
DOS VARIABLES CUALITATIVAS

```
Frequency table:
             Sexo
Academia      Hombre Mujer
  No academia     11    11
  Sí academia      4     9

Row percentages:
             Sexo
Academia      Hombre Mujer Total Count
  No academia   50.0  50.0   100    22
  Sí academia   30.8  69.2   100    13

Column percentages:
             Sexo
Academia      Hombre Mujer
  No academia   73.3    55
  Sí academia   26.7    45
  Total        100.0   100
  Count         15.0    20
```

UNA CUANTITATIVA Y UNA CUALITATIVA




```
                mean       sd  IQR       cv       skewness   0% 25% 50% 75% 100% Notas_final:n
Ev. Continuada 6.000000 2.160247 3.75 0.3600411  0.12470230  2 4.25 5.5   8   10            22
Ev. Única      4.692308 1.974192 3.00 0.4207295 -0.03135556  1 3.00 5.0   6    8            13
```

DOS VARIABLES CUANTITATIVAS

Diagrama de dispersión

Covarianza (S_{xy})

Correlación (r_{xy})

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

Coeficiente de determinación (R^2)

$$R^2 = {r_{xy}}^2$$

Modelo de regresión

$$\hat{y} = \beta_o + \beta_1 \cdot x$$

$$\beta_1 = \frac{S_{xy}}{S_x^2} = r_{xy} \cdot \frac{S_y}{S_x} \qquad \beta_o = \bar{y} - \beta_1 \cdot \bar{x}$$

Ejemplo R Commander:

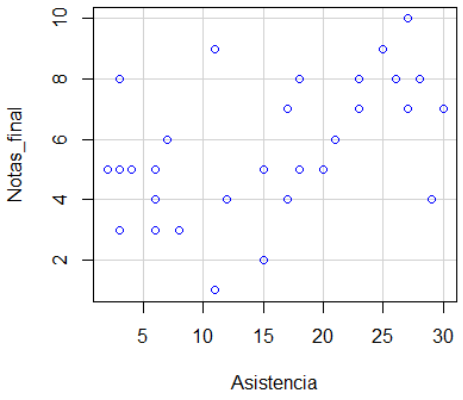
```
             Asistencia  Edad      Notas_final
Asistencia    1.0000000 -0.3601305   0.4851734
Edad         -0.3601305  1.0000000  -0.2902839
Notas_final   0.4851734 -0.2902839   1.0000000

Call:
lm(formula = Notas_final ~ Asistencia, data = Estadistica)

Residuals:
   Min     1Q Median     3Q    Max
-3.987 -1.109 -0.201  1.149  4.013

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  3.65183    0.66822   5.465 0.00000467 ***
Asistencia   0.12139    0.03808   3.187    0.00314 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.918 on 33 degrees of freedom
Multiple R-squared:  0.2354,  Adjusted R-squared:  0.2122
F-statistic: 10.16 on 1 and 33 DF,  p-value: 0.003136
```



4. TUTORIAL R COMMANDER

1. Carga la base de datos TutorialRcommanderFE.xlsx y codifica correctamente las variables categóricas:

Edad: edad al empezar el curso.
Sexo: 1.- Hombre, 2.-Mujer
Evaluación: 1.- Ev. continuada, 2.- Ev. única
Asistencia: número de clases a las que ha asistido el sujeto. Se realizaron 30 clases en total.
Academia: 1.- Ha ido a alguna academia a lo largo del curso, 2.- No ha ido a ninguna academia.
Notas_1: nota del examen parcial de mitad de curso.
Notas_2: nota del examen final de todos los alumnos.

2. Genera una variable nueva que llamaremos "asistenciacuali" que sea si la asistencia ha sido baja, menos de 10 clases asistidas, media, entre 10 y 20 clases asistidas y alta, más de 20 clases asistidas.

3. Recodifica la variable Sexo para poner "mujer" como primera categoría.

4. Calcula la nota final de curso "notacurso" teniendo en cuenta que el parcial contaba un 30% de la nota y el final el otro 70%.

5. Realiza el análisis descriptivo univariante que consideres adecuado de las variables solo para los alumnos que han ido a academia:
- Edad
- Sexo
- Asistenciacuali

6. Analiza las notas del examen final en función de si el alumno ha ido a academia o no ha ido.

7. Analiza las clases asistidas de los alumnos que han hecho evaluación continua.

8. ¿Analiza la relación que hay entre "sexo" y "academia"?

9. Estudia la relación entre "asistencia" y "notas_2"

PEC 0-1: RESOLUCIÓN

MODELO DE PEC 0

Descargar archivo EPFH_2020.RData

Pregunta 1
Sin responder aún
Puntúa como 1,00

Calculad el resultado de la siguiente operación:

$$0.0006 + 0.3 \cdot 0.11$$

Dad el resultado con un mínimo de tres decimales y tres cifras significativas. Tened en cuenta el orden de las operaciones y recordad que la respuesta no puede estar mal por exceso de decimales.

Respuesta:

Pregunta 2
Sin responder aún
Puntúa como 1,00

Calculad el resultado de la siguiente división: $\frac{25}{300}$

Dad el resultado con un mínimo de tres decimales y tres cifras significativas. Recordad que la respuesta no puede estar mal por exceso de decimales.

Respuesta:

Las variables con las que trabajaremos en esta PEC son:

- **CCAA**: La variable se corresponde con la lista codificada de las diferentes comunidades autónomas, eliminando Ceuta y Melilla por su escaso número de datos.
- **SECTOR**: hace referencia al sector donde trabaja el sustentador principal del hogar. Se trata de una variable categorizada con dos valores, dependiendo de si trabaja en el sector PÚBLICO o PRIVADO.
- **SEXOSP**: hace referencia al género del sustentador principal. Se trata de una variable categorizada con dos valores, HOMBRE o MUJER.
- **TIPOCONT**: hace referencia al tipo de contrato que obstenta el sustentador principial. Se trata de una variable categorizada con dos valores, INDEFINIDO o TEMPORAL.

Indicaciones

1. Tenéis que cargar el paquete R-Commander. Podéis hacerlo con los siguientes menús: Packages –> Load Package… –> Rcmdr

Alternativamente, podéis escribir library(Rcmdr) en la consola de R. Para que este paso funcione tenéis que tener previamente instalado R-Commander. Si no lo habéis hecho, podéis instalarlo con install.packages("Rcmdr") o bien usar el menú Packages. Tenéis instrucciones detalladas para la instalación de R y R-Commander en el aula.

2. Tenemos que cargar los datos en R-Commander, que están en el fichero EPFH_2020.RData que os podéis bajar del aula. Para cargarlo tenéis que hacer: Datos —> Cargar conjunto de datos…

Si los menús de R-Commander os salen en otro idioma y esto es un problema para vosotros, mirad la indicación #7

3. Podéis comprobar que la tabla está bien cargada clicando el botón "Visualizar conjunto de datos".
4. Ahora tenemos cargados todos los datos de la muestra, con todos los hogares respondientes, pero nosotros queremos trabajar solo con los de los parámetros que nos corresponda y que el enunciado indica. Podemos hacerlo con Datos > Conjunto de datos activo > Filtrar conjunto de datos activo.

Así, por ejemplo, en el campo «Expresión de selección» ponemos CCAA==NÚMERO DE CCAA QUE SE OS HAYA ASIGNADO (escrito exactamente como aparece en la tabla de datos y con comillas, dado que la codificación es alfanumérica aunque veamos un número). Por ejemplo, alguien a quien le haya tocado la comunidad de Castilla y León, podemos ver en el listado de campos de codificación que su código es "07", por lo que escribiríamos CCAA=="07".

5. Podemos volver a comprobar el contenido de la tabla de datos clicando el botón "Visualizar conjunto de datos". Ahora tendríamos que ver todos los registros de la misma comunidad autónoma, o sea, todos con el mismo valor en la variable CCAA. Si no es así, repasad los pasos anteriores.
6. Debemos repetir el paso 4 (y la validación explicada en el punto 5) para la variable SECTOR, según se indique en el ejercicio. Tened en cuenta que aquí los valores también son alfanuméricos y por lo tanto también tienen que ponerse entrecomillados, o sea, SECTOR=="1" o SECTOR=="6".
7. Una de las formas de cambiar el idioma de R-Commander, en Windows, es la siguiente:

- Abrid R (todavía no R-Commander).
- En los menús de R vais a Edit > GUI preferences.
- En la casilla "Language for menús and messages" (arriba a la derecha) poned *es* (o la abreviación del idioma que os apetezca).
- Si el programa os lo pide, cerrad R y volved a abrirlo.
- Cuando abráis R-Commander se tiene que abrir en castellano.

Pregunta (a)

Vamos a trabajar con un hogar correspondiente a la Comunidad 05 ("*Canarias*"), donde el sustentador principal trabaja en el sector "*PRIVADO*".

Una vez que hayáis cargado y filtrado la base de datos según se indica en el ejercicio, realiza una tabla de doble entrada con las variables SEXOSP y TIPOCONT (siguiendo la siguiente ruta en R-Commander: ***Tabla de contingencia/Tabla de doble entrada***) y responde a la siguiente pregunta: ¿Qué porcentaje de hogares de entre los que el sustentador principal tiene contrato INDEFINIDO son HOMBRES?

Observación: Una tabla de contingencia es una forma de representar dos variables estadísticas categóricas (cualitativas) mediante una tabla en la que se organizan sus categorías por filas y columnas. Las frecuencias de la tabla representan el número de individuos que tienen de forma conjunta una determinada categoría de la variable representada en filas y una determinada categoría correspondiente a la variable representada en columnas. Las sumas de los valores por filas y columnas nos dan los totales de cada una de las categorías de las variables.

Pregunta (b)

Continuamos trabajando con los mismos datos que en la cuestión anterior, es decir: un hogar correspondiente a la Comunidad 05 ("*Canarias*"), donde el sustentador principal trabaja en el sector "*PRIVADO*".

¿Qué porcentaje sobre el total corresponde a hogares cuyo sustentador principal son MUJERES y tienen contrato TEMPORAL?

- Observación1: debemos excluir del total aquellas observaciones sin información de esta variable.
- Observación2: indiquemos la solución en **%** y con **dos decimales** y **sin el signo de %**. Por ejemplo, si la frecuencia relativa es 0,322189 entonces le corresponde el porcentaje 32,22%, por lo que la respuesta que indicaríamos sería 32,22.

MODELO DE PEC 1

Pregunta 1
Sin responder aún
Puntúa como 1,00
Marcar pregunta

Para medir la dispersión de una variable estadística utilizamos:

Seleccione una:
- a. El coeficiente de variación de Pearson
- b. La desviación típica
- c. Todas las demás opciones son ciertas
- d. El rango intercuartílico

Pregunta 2
Sin responder aún
Puntúa como 1,00
Marcar pregunta

Si los 42 trabajadores de un hotel tienen todos el mismo sueldo bruto, de 1993 u.m., menos uno que tiene un sueldo bruto de 1915 u.m., la mayoría de los trabajadores tiene unos ingresos inferiores a la media.

Seleccione una:
- Verdadero
- Falso

Pregunta 3
Sin responder aún
Puntúa como 1,00
Marcar pregunta

Un estudiante ha hecho 7 trabajos en un curso y ha obtenido una nota media de 6,6. Posteriormente, se revisan las notas y un en uno de los trabajos donde había obtenido un 8,7 la nota se cambia por un 6,7. ¿Cuál será ahora la nota media?

Responded con una exactitud mínima de dos decimales.

Respuesta:

Pregunta 4
Sin responder aún
Puntúa como 1,00
Marcar pregunta

Calculad los estadísticos del conjunto de observaciones (1, 6, 13, 21, 28, 32, 34, 41, 51). Tened en cuenta que en esta pregunta se ha utilizado la varianza y la desviación estándar poblacionales (tal como se definen en el material escrito) mientras que Rcommander utiliza la varianza y la desviación estándar muestrales. Igualmente, la definición de cuartiles empleada aquí es la del material escrito, que da un resultado ligeramente diferente que la de RCommander.

mínimo	Elegir...
máximo	Elegir...
media	Elegir...
percentil 25%	Elegir...
mediana	Elegir...
percentil 75%	Elegir...
desviación estándar	Elegir...
varianza	Elegir...

Pregunta 5

Sin responder aún

Puntúa como 6,00

Marcar pregunta

Hemos recogido el número de asignaturas a las que se han matriculado este año los estudiantes de estadística de una cierta universidad: 7 estudiantes hacen 2 asignaturas, 10 estudiantes hacen 3 asignaturas, 15 estudiantes hacen 4 asignaturas, 14 estudiantes hacen 5 asignaturas i 10 estudiantes hacen 6 asignaturas.

Calculad los estadísticos que se piden a continuación:

- media:
- Mediana:
- Primer cuartil:
- Tercer cuartil:
- Varianza:
- Desviación estándar:

Indicacionse:

- Dad todos los resultados con una precisión de tres cifras significativas o tres decimales.
- Para los cuartiles podéis utilizar tanto la definición del material de la asignatura como la que utiliza RCommander, que dan resultados ligeramente diferentes.
- Para la varianza y la desviación estándar utilizad la varianza muestral o corregida, o sea, con el denominador $n-1$ en vez de n.

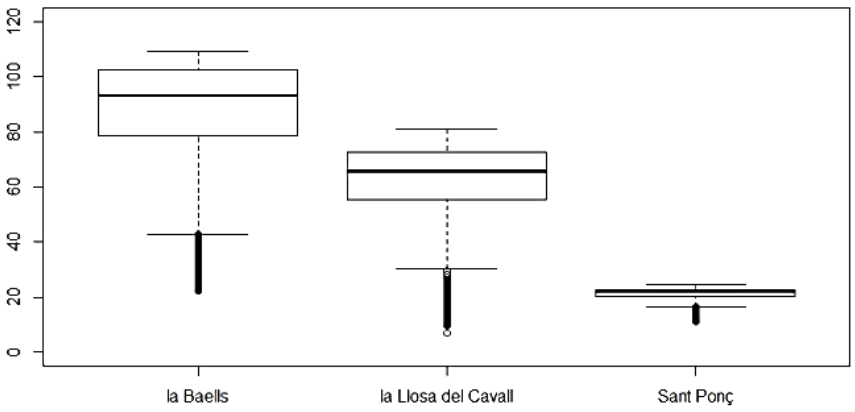
Pregunta 6

Sin responder aún

Puntúa como 1,00

Marcar pregunta

El siguiente gráfico representa el volumen embalsado en los embalses de la cuenca del Llobregat entre 2007 y 2017, o sea, la variable representada es la medida diaria de la cantidad de agua que hay en cada embalse.



Indicad cuál de las siguientes opciones es FALSA. Si alguna cuestión no se puede responder a partir de estos gráficos, consideradla falsa. Si todas las opciones son ciertas o todas son falsas, marcad "Ninguna de las demàs".

Seleccione una:

- a. En estos boxplots no está representada la media.
- b. Ninguna de las otras (seleccionad esta opción si creéis que las otras son todas ciertas o todas falsas)
- c. El embalse de la Baells no ha estado totalmente vacío en ningún momento de los 10 años.
- d. La media del volumen embalsado en Sant Ponç es inferior a la media del volumen embalsado en la Llosa del Cavall.
- e. La distribución del volumen embalsado en la Baells es asimétrica a la izquierda.
- f. Más del 50% de los días había más agua en la Baells que en la Llosa del Cavall.
- g. Si en estos diez años todos los embalses han llegado alguna vez a llenarse hasta su máxima capacidad, el mayor embalse de los tres es el de la Baells.

Pregunta 7

Sin responder aún

Puntúa como 1,00

Marcar pregunta

Hemos calculado los siguientes estadísticos de cuatro muestras diferentes:

```
##    Muestra 1       Muestra 2        Muestra 3        Muestra 4
##  Min.   :105.0   Min.   : 77.27   Min.   : 95.12   Min.   : 65.21
##  1st Qu.:127.4   1st Qu.: 95.27   1st Qu.:104.92   1st Qu.: 77.93
##  Median :146.3   Median :109.70   Median :129.02   Median : 99.17
##  Mean   :142.7   Mean   :111.77   Mean   :127.04   Mean   : 99.17
##  3rd Qu.:158.1   3rd Qu.:130.38   3rd Qu.:146.75   3rd Qu.:119.66
##  Max.   :174.8   Max.   :146.70   Max.   :160.64   Max.   :132.13
```

¿A cuál de estas cuatro muestras pertenecen los siguientes cuantiles?

```
##       20%       30%       40%       60%       70%       80%
##  76.01354  83.28016  92.27060 105.07353 115.31085 122.76723
```

Seleccione una:

- a. Muestra 1
- b. Muestra 2
- c. Muestra 3
- d. Muestra 4

Pregunta 8

Sin responder aún

Puntúa como 1,00

Marcar pregunta

La precipitación acumulada diaria es la cantidad total de lluvia caída en un lugar a lo largo de un día y se mide en mm o en litros por metro cuadrado, que es lo mismo. Si un día no llueve, la precipitación acumulada de ese día es cero. A partir de una serie de 10 años de la precipitación caída en tres localidades catalanas hemos obtenido el siguiente sumario:

L'Espunyola (Berguedà):

```
 Min.  1st Qu. Median Mean  3rd Qu. Max.
0.000  0.000   0.000  1.789 0.200 114.600
```

Piera (Anoia):

```
 Min.  1st Qu. Median Mean  3rd Qu. Max.
0.000  0.000   0.000  1.356 0.100   135.485
```

Santa Maria de Miralles (Anoia):

```
 Min.  1st Qu. Median Mean  3rd Qu. Max.
0.000  0.000   0.000  1.372 0.000   77.700
```

Indicad cuál de las siguientes opciones sobre la lluvia en estas tres localidades es CIERTA:

Seleccione una:

- a. Ninguna de las otras es cierta.
- b. Todas las demás son ciertas (excepto la que dice que son todas falsas).
- c. En promedio, donde llueve más es en Piera.
- d. El día que más llovió en l'Espunyola, llovió más que cualquier día en cualquiera de las tres localidades.
- e. Donde llueve menos días es en Santa Maria de Miralles.

Pregunta 9

Sin responder aún

Puntúa como 1,00

Marcar pregunta

Empareja los siguientes valores del coeficiente de correlación entre las variables X e Y con su interpretación:

0,6	Elegir...
4	Elegir...
-0,19	Elegir...

Pregunta 10

Sin responder aún

Puntúa como 1,00

Marcar pregunta

Un hotel oferece habitaciones individuales, dobles, triples y cuádruples. Para estancias de dos días, cual será el coeficiente de correlación entre las variables "gasto por persona" y "número de personas en la habitación"?

Seleccione una:

- a. Entre -1 y 0
- b. Entre 0 y 1
- c. Próximo a 0
- d. 1 o muy cercano a 1
- e. -1 o muy cercano a -1

Pregunta 11

Sin responder aún

Puntúa como 6,00

Marcar pregunta

Estamos estudiando la relación entre dos variables x e y. Disponemos de los siguientes estadísticos:

$\overline{x} = 130.3749$
$\overline{y} = 51.8573$
$s_x = 11.4199$
$s_y = 15.4527$
$s_{xy} = -167.239$

Responded a las siguientes preguntas con un mínimo de tres decimales.

¿Cuánto vale el coeficiente de correlación?

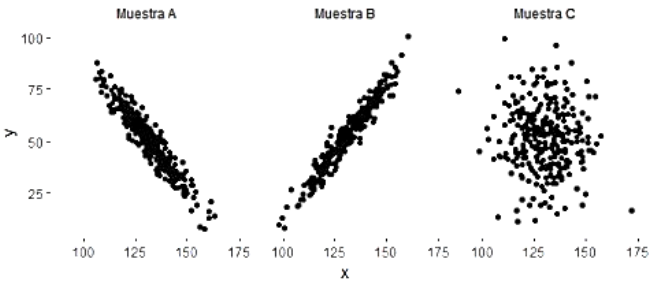
¿Y la pendiente de la recta?

¿Y el término independiente?

Calculad el coeficiente de determinación o de bondad de ajueste.

¿Qué valor de y predeciríamos para $x = 154.8225$?

De acuerdo con los resultados anteriores, decidid cuál de los diagramas de dispersión siguientes corresponde a las variables x e y.



El gráfico de la .

Pregunta 12

Sin responder aún

Puntúa como 5,00

Marcar pregunta

Hemos recogido datos de los beneficios de una muestra de grandes empresas a partir de la lista Forbes de las empresas más grandes del mundo del año 2004. Cuando el valor de los beneficios es negativo significa que la empresa tiene pérdidas.

A continuación tenemos algunos percentiles de la variable beneficios en miles de millones de dólares, para las empresas de tres países diferentes:

Malasia:

```
##     0%    10%    20%    30%    40%    50%    60%    70%    80%    90%
## -0.010  0.085  0.130  0.140  0.170  0.205  0.250  0.270  0.290  0.325
##   100%
##  0.530
```

Japón:

```
##      0%     10%     20%     30%     40%     50%     60%     70%     80%
## -20.110  -0.245  -0.050   0.010   0.030   0.070   0.110   0.160   0.250
##     90%    100%
##   0.425   7.990
```

Reino Unido:

```
##      0%     10%     20%     30%     40%     50%     60%     70%     80%
## -16.030  -0.695   0.020   0.120   0.160   0.205   0.280   0.345   0.520
##     90%    100%
##   0.825  10.270
```

En qué país hay un porcentaje mayor de empresas que tienen pérdidas?

Seleccione una:

- a. Malasia
- b. Japón
- c. Reino Unido
- d. Con estos datos no lo podemos saber

Este dosier está hecho para seguir la clase de prueba.

Más información en:

www.estadistix.com

Y si tienes cualquier consulta,
escríbenos un whatsapp al 644310902