UNIDAD 1: Análisis exploratorio de datos univariados y bivariados
TRABAJO PRÁCTICO

1) Señale en qué caso es más conveniente estudiar la población o una muestra. Razone la respuesta.
   a) La longitud de los tornillos que fabrica una máquina de manera continua durante un día.
   b) La estatura de los turistas extranjeros que visitan España en un año.
   c) El peso y la altura de un grupo de cinco amigos.
   d) La duración de una bombilla hasta que se funde.
   e) El sueldo de empleados de una empresa.

2) Dé una posible muestra de tamaño 4 de cada una de las siguientes poblaciones y clasifíquelas en "concretas" ó "hipotéticas"
   a) Todos los periódicos on-line en Argentina.
   b) Todas las compañías listadas en la Bolsa de Valores de Argentina
   c) Todos los estudiantes de la Facultad de Economía y Administración de la UNCo.
   d) Todas las distancias que podrían resultar cuando usted lanza una pelota de fútbol.
   e) Todas las mediciones de intensidades posibles de terremotos (escala de Richter) que pudieran registrarse en América de aquí a 5 años.
   f) Todos los posibles rendimientos (en gramos) de una cierta reacción química realizada en un laboratorio.

3) La siguiente información se tuvo en cuenta en una encuesta realizada a estudiantes cuando salían de la librería de la universidad durante la primera semana de clases.
   a) Cantidad de tiempo dedicada a comprar en la librería
   b) Número de libros de texto adquiridos
   c) Carrera que estudia
   d) Género
   e) Si el estudiante cuenta con un perfil de Facebook actualmente
   f) Número de WhatsApp enviados en el día anterior
   g) Tiempo en horas por semana dedicado a navegar en internet

Clasifique cada una de estas variables como categórica o numérica. Si la variable es numérica, establezca si es discreta o continua.

4) El tratamiento de los niños con desórdenes de la conducta puede ser complejo. El tratamiento se puede proveer en una variedad de escenarios dependiendo de la severidad de los comportamientos. Además del reto que ofrece el tratamiento, se encuentran la falta de cooperación del niño/niña y el miedo y la falta de confianza de los adultos. Para poder diseñar un plan integral de tratamiento, el siquiatra de niños y adolescentes puede utilizar información del niño, la familia, los profesores y de otros especialistas médicos para entender las causas del desorden. Para ello, un siquiatra local ha considerado una muestra aleatoria de 20 niños, anotando el tiempo (en horas) necesario que requiere en cada niño para lograr un plan integral del tratamiento, obteniéndose

6	7	7	8	8	8	8	9	9	9
9	9	9	9	10	10	10	10	10	11

   a) Describa los objetivos generales del estudio.
   b) Defina la variable de interés y la unidad estadística.
   c) Calcule los estadísticos de posición y dispersión.

d) Realice el diagrama de puntos y el diagrama de cajas. Escriba un párrafo de texto caracterizando a los niños en función del tiempo que necesitan para lograr un plan integral del tratamiento (tenga en cuenta los cinco números resumen)

e) Comente acerca de la forma de la distribución del tiempo necesario que requiere cada niño. Conoce algún estadístico que calcule la asimetría? Qué valor tiene el estadístico. Concuerda con lo observado en los gráficos anteriores.

f) Calcule el percentil 60 e interprete.

g) Hay observaciones atípicas?

5) Los siguientes datos fueron obtenidos observando el número de lesiones en la piel de una muestra de 30 manzanas:

0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6

a) Defina la variable "x", la unidad estadística e indique tipo de variable.

b) Realice la tabla de distribución de frecuencias.

c) Interprete y represente en el gráfico correspondiente: f_2, h_1%, F_4, H_3%, 100-H_3%.

d) Calcule el número medio de golpes por manzana.

e) Calcule la mediana y la moda. Interprete ambos estadísticos.

f) Calcule rango, varianza, desvío estándar muestral y coeficiente de variación.

g) Realice el diagrama de puntos.

h) Realice el diagrama de cajas y comente sobre la forma de la distribución.

i) A los datos originales restar la media y hallar media y desvío estándar de la variable transformada $y = (x - \bar{x})$.

j) Dividir los datos obtenidos en el inciso anterior por el desvío estándar, y volver a calcular media y desvío estándar $z = {}^{y}/_{s(y)}$. ¿Qué propiedad está verificando?

6) La central generadora de Energía eléctrica, realiza un estudio sobre los niveles de tensión recibida en las redes de distribución media de tensión (cliente industrial) para ello se registraron los siguientes datos que corresponden a los niveles de tensión (KV) de una muestra de 50 clientes industriales:

15,30	12,32	13,13	13,25	14,19	13,84	13,73	13,41	13,12	14,29
13,42	14,81	12,57	13,98	14,63	14,83	13,55	15,22	12,83	14,51
13,82	13,82	12,26	14,67	13,79	15,18	14,51	12,46	13,23	13,84
13,75	14,23	13,84	14,41	14,07	13,74	12,91	12,63	11,70	13,64
12,83	13,33	14,35	14,46	13,75	13,27	14,32	14,13	13,63	13,89

a) Defina la variable en estudio y la unidad estadística. Indique tipo de variable.

b) Realice la tabla de frecuencias. Represente gráficamente frecuencias simples y acumuladas.

c) Interprete e indique en el gráfico correspondiente: f_2, F_3, h_5%, 100 – H_3%.

d) Calcule gráfica y analíticamente la moda.

e) Calcule e interprete la media y la mediana. Calcule el desvío estándar.

f) Calcule e interprete el cuartil superior.

g) Represente la distribución de los niveles de tensión mediante un diagrama de cajas y comente sobre las características de la distribución.

h) Si se supone un aumento de 10 KV en cada uno de los clientes industriales, cómo se modifican la media y el desvío estándar? El coeficiente de variación cambia? Qué propiedades aplicó.

i) Ídem al punto anterior suponiendo un aumento del 5% sobre el nivel de tensión.

7) Los transductores[1] de temperatura de cierto tipo se envían en lotes de 50. Se seleccionó una muestra de 80 lotes y se determinó el número de transductores en cada lote que no cumplen con las especificaciones de diseño. La información recolectada quedó almacenada en el archivo *transductores.xlsx*

a) Indique la unidad estadística y defina la variable en estudio clasificándola según su naturaleza.
b) Realice la tabla de distribución de frecuencias y grafique.
c) Halle e interprete cuando corresponda: media, moda, mediana, desvío estándar y coeficiente de variación.
d) ¿Qué proporción de lotes muestreados tienen a lo sumo cinco transductores que no cumplen con las especificaciones? ¿Qué proporción tiene menos de cinco? ¿Qué proporción tienen por lo menos cinco unidades que no cumplen con las especificaciones?
e) Realice el diagrama de caja correspondiente y comente sobre las características de la distribución.

8) En el departamento de Control de Calidad de cierta fábrica de alambre de cobre se observan al microscopio diariamente al azar 100 porciones de 1 mm de longitud de dicho alambre y se cuenta la cantidad de fallas en el mismo. El archivo *NroFallas.xlsx* contiene la cantidad de fallas encontradas en las 2000 observaciones realizadas al cabo de un mes (20 días laborables).

a) Defina la unidad estadística y variable bajo estudio.
b) Realice la tabla de distribución de frecuencias y grafique.
c) Halle los estadísticos descriptivos: media, mediana, moda y desvío estándar.
d) Un estudio equivalente realizado en EEUU sobre porciones de alambre de 1/8 pulgadas de longitud arrojó una media de 4,13 fallas por porción y un desvío estándar igual a 2,64. Cuál de las dos distribuciones de frecuencias es más homogénea?

9) 120 probetas de un determinado material fueron sometidas a una carga de 1 tonelada midiéndose su deflexión en milímetros. El archivo *probeta.xlsx* contiene dicha información.

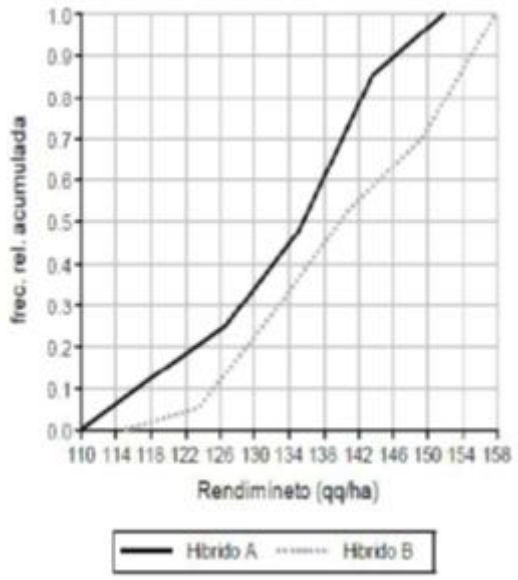
a) Indique unidad estadística. Defina la variable en estudio. Indique tipo de variable.
b) Realice la tabla de distribución de frecuencias y grafique.
c) A partir del diagrama de caja comente sobre las características de la distribución.
d) Indique el valor del coeficiente de asimetría. Es compatible con lo que se observa en el diagrama de caja?
e) Interprete la media y la mediana.
f) Interprete el cuartil inferior y el superior.
g) Calcule el desvío estándar y el coeficiente de variación.
h) Si el estadístico de apuntamiento es -0,8531 qué nos indica?

10) El número de plantas Larrea divaricata encontradas en cada uno de 48 bloques de muestreo según el informe del artículo “Some Sampling Characteristics of Plants and Arthropods of the Arizona Desert” (Ecology, 1962: 567-571), se encuentran en el archivo *Larrea.xls:*

a) Indique unidad estadística. Defina la variable en estudio e indique su tipo.
b) Construya la tabla de Frecuencias correspondiente.
c) Realizar las gráficas correspondiente a frecuencias absolutas fi y frecuencias absolutas acumuladas Fi
d) ¿Qué porcentaje de bloques tiene 3 plantas?
e) ¿En cuántos bloques se registró una planta o más?
f) ¿Qué cantidad de bloques registró entre 2 y 6 plantas inclusive? Represente.
g) Calcule los estadísticos de posición: media, mediana y moda. Interprete.
h) Calcule: rango, variancia, desvío estándar y coeficiente de variación.
i) Calcule primer y tercer cuartil.
j) Realizar el diagrama de cajas e indique la forma de la distribución.

[1] Dispositivos que convierten el fenómeno físico de la temperatura en una señal eléctrica normalizada

11) Un semillero decidió poner a prueba el rendimiento de dos híbridos experimentales de sorgo granífero bajo riego. Se estudiaron dos muestras, una del híbrido A y otra del híbrido B. Indique V o F.



El 40% de los valores obtenidos con el híbrido A son superiores a 138 qq/ha.	☐
Con el híbrido A aproximadamente el 80% de los rendimientos fueron de 142 qq/ha	☐
La proporción de rendimientos entre 122 y 146 qq/ha fue superior con el híbrido A	☐
El cuantil 0,50 en el híbrido B es superior al cuantil 0,50 del híbrido A	☐
La mediana mediana fue mayor en el híbrido B	☐
En el híbrido A hubo mayor cantidad de parcelas con rendimientos de hasta 146 qq/ha	☐
El P(70) del híbrido B es de aproximadamente 150 qq/ha	☐

12) El archivo bombas.xlsx contiene los datos de dos muestras independientes, que representan la duración de la vida, en años, de 30 bombas de combustible similares a temperaturas extremas. Describa el patrón del comportamiento de los datos de ambos tipos de bombas.

13) Los datos del archivo *resistenciaflexion.xlxs* corresponden a la resistencia a la flexión de dos muestras independientes de 20 juntas seleccionadas al azar sin recubrimiento y 25 juntas seleccionadas al azar con recubrimiento.

a) Defina la variable en estudio e indique la unidad estadística.

b) Calcule los estadísticos de la distribución de la resistencia a la flexión de las juntas según tipo de recubrimiento.

c) Analice las gráficas de diagramas de cajas y compare ambas distribuciones según forma, dispersión y localización.

d) Redacte un informe resaltando las características más sobresalientes de la resistencia a la flexión de las juntas según tipo de recubrimiento.

14) Considere la representación gráfica, que se muestra en la siguiente figura, de los diagramas de caja, correspondientes a la medida de los gramos de fibra que tienen los cereales de las marcas G, K, N, P, Q y R.

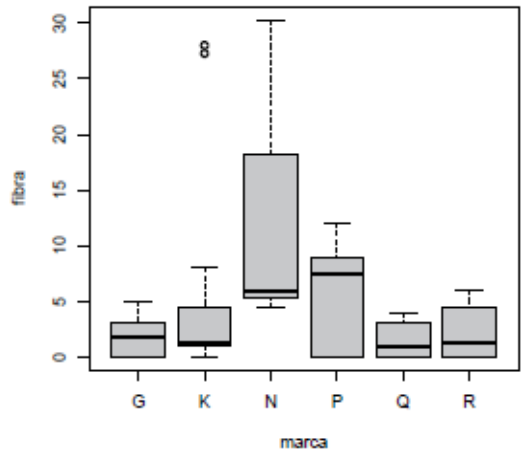


Figura: Diagramas de caja de los gramos de fibra en varias marcas de cereales

a) Cuántos gramos de fibra tiene que tener un cereal de la marca N para que el 50% de los de su marca tengan menos fibra?

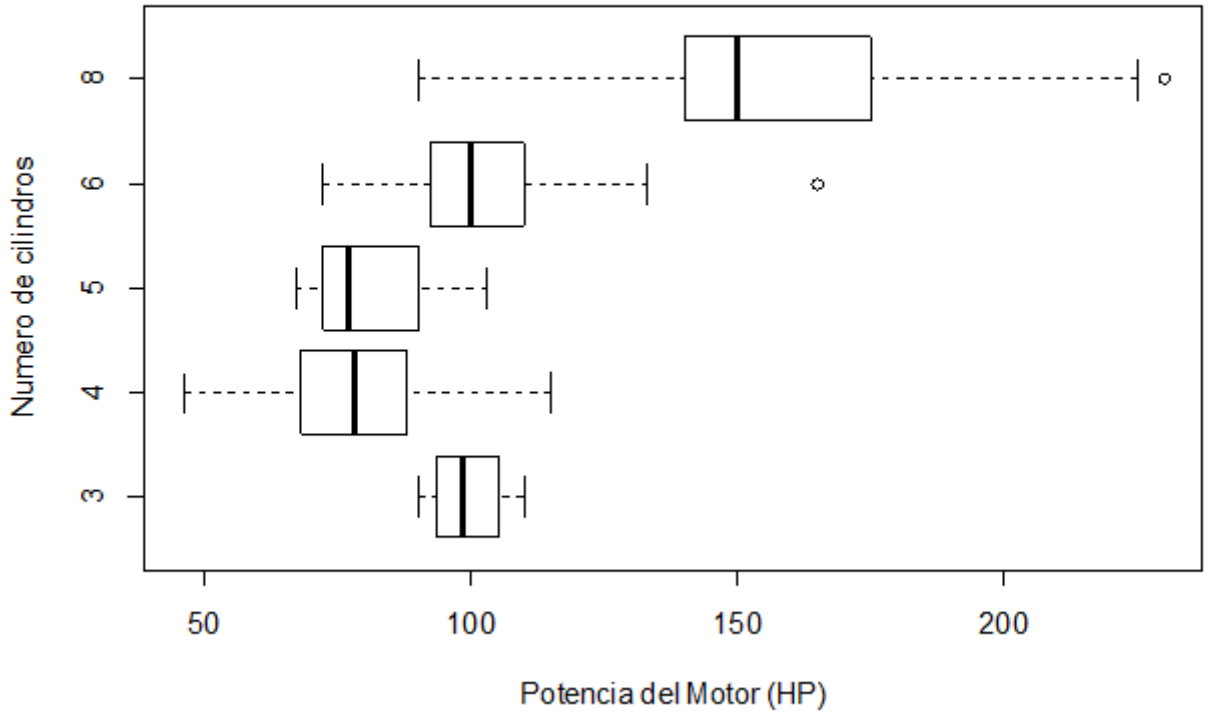
b) ¿Qué marca tiene más variabilidad? ¿Por qué?

c) ¿Cuánto vale el primer cuartil de la marca G? ¿Qué interpretación tiene ese valor? ¿Hay otras marcas con este mismo valor?

d) ¿Cuánto vale el rango intercuartílico para la marca N? ¿Qué porcentaje de paquetes se encuentran dentro de la caja de la marca N?

e) ¿Qué son los dos círculos que aparecen en el gráfico? Explique brevemente como se determinan.

15) Considere la población correspondiente a los 397 modelos de automóviles que se comercializaron en los EEUU en el año 1983 y la siguiente representación mediante diagramas de caja de la distribución de la potencia del motor (en HP) clasificada según el número de cilindros del mismo:



a) Defina la población. Indique la unidad estadística y el tipo de variables estudiadas según su naturaleza.

b) Considerando el número de cilindros, describa la ***forma*** de cada distribución de la potencia del motor.

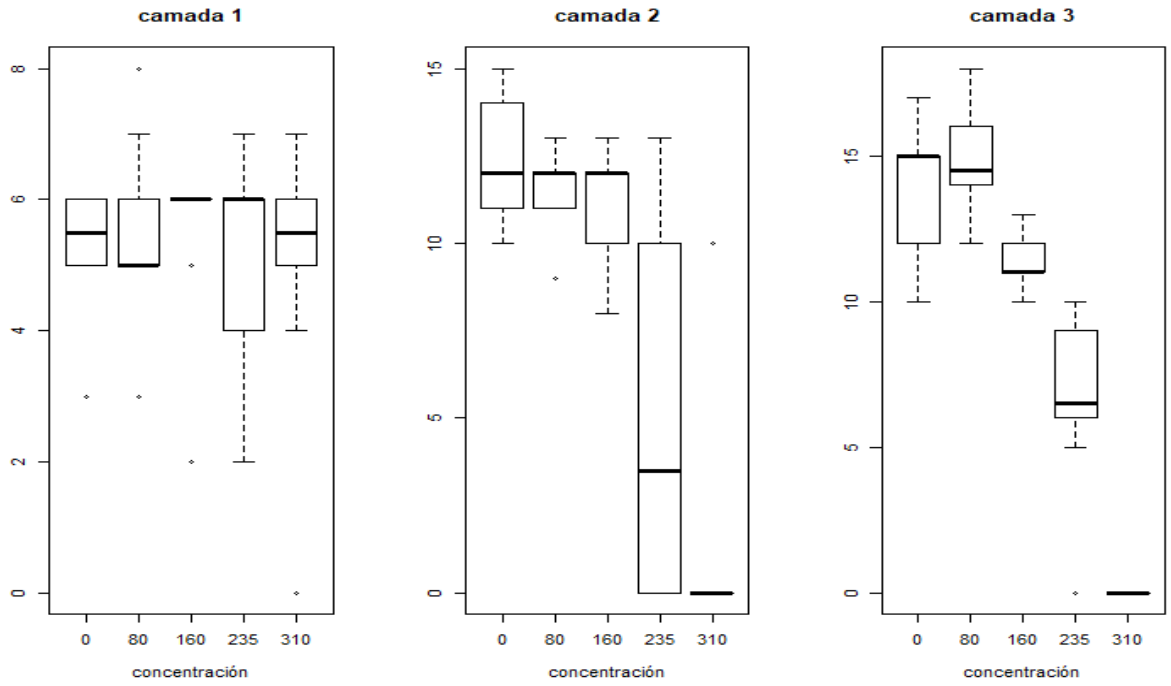
c) Compare las distribuciones de los motores de 4 y 5 cilindros (***forma***, ***dispersión*** y ***centro***)

d) Compare las distribuciones de los motores de 4 y 6 cilindros (***forma***, ***dispersión*** y ***centro***)

e) Observar el ***centro*** de cada distribución de acuerdo al número de cilindros: podría formar grupos de motores con una potencia similar? Cuáles son esos grupos?

16) Se quiere medir la efectividad de un herbicida, nitrofen, [2] en el zooplancton. Un total de 50 especímenes se distribuyen aleatoriamente en grupos de 10 y se introducen en una solución a distintas concentraciones de nitrofen: 0, 80,160, 235 y 310 mg/l. A continuación se contabiliza el número de crías vivas en tres camadas sucesivas. Los datos obtenidos se resumen en los siguientes diagramas.

[2] Los datos están disponibles en el cuadro de datos denominado nitrofen del paquete boot de R.



a) ¿A qué camada hace referencia las medidas que se presentan a continuación? ¿Por qué?

	mean	sd	IQR	cv	skewness	kurtosis
0	13.9	2.183	2.50	0.157	-0.722	-0.355
80	14.8	1.751	2.00	0.118	0.223	-0.062
160	11.5	0.971	1.00	0.084	0.453	-0.516
235	6.7	2.945	2.75	0.439	-1.197	2.304
310	0.0	0.000	0.00	NA	NaN	NaN

	0%	25%	50%	75%	100%	data:n
0	10	12.5	15.0	15.00	17	10
80	12	14.0	14.5	16.00	18	10
160	10	11.0	11.0	12.00	13	10
235	0	6.0	6.5	8.75	10	10
310	0	0.0	0.0	0.00	0	10

b) Interprete las medidas que se han calculado en la fila correspondiente a la concentración de 235 mg/l de nitrofen.

17) Considere los siguientes datos: 2.2, 7.6, 2.9, 4.6, 4.1, 3.9, 7.4, 3.2, 5.1, 5.3, 20.1, 2.3, 5.5, 32.7, 9.1, 1.7, 3.2, 5.8, 16.3, 15.9, 5.9, 6.7, 3.4 y 40.

a) Represente el diagrama de caja identificando atípicos

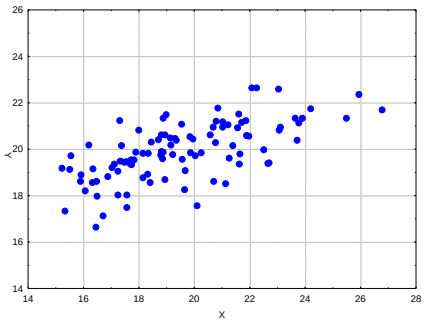
b) Transforme los datos por el logaritmo neperiano y construya nuevamente el diagrama de caja. ¿Qué observa?

c) Calcule los estadísticos descriptivos que considere más importantes incluyendo el coeficiente de asimetría y apuntamiento para las dos variables.

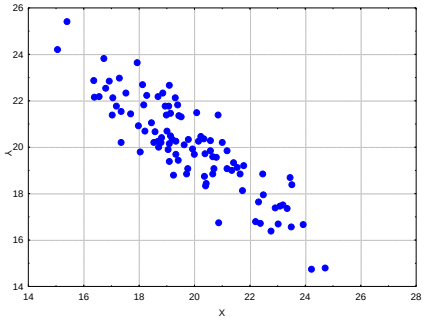
18) Si en una parcela hay 45 árboles de una altura media de 4,75 mt y en otra parcela de 64 árboles el promedio de altura es de 5,25 mt. ¿Cuál es la altura media del conjunto de árboles de ambas parcelas?

19) Dadas las siguientes representaciones gráficas de la relación entre dos variables aleatorias, asignarle a cada una la matriz de variancias-covariancias correspondiente:

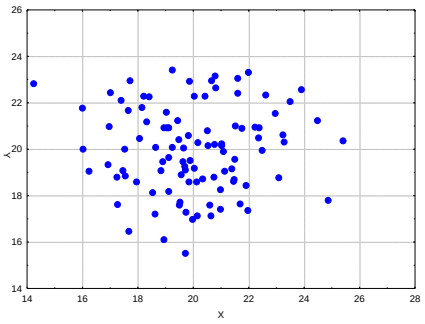
A



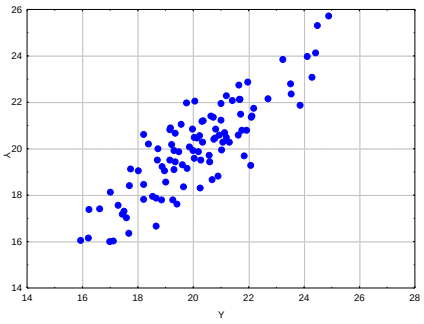
B



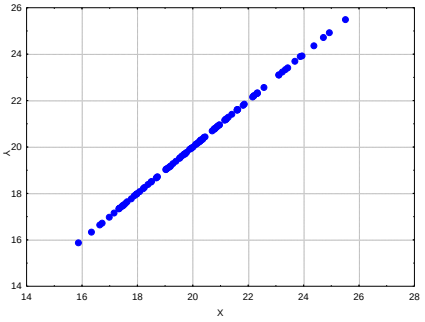
C



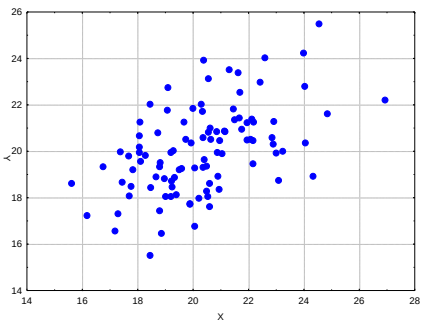
D



E



F



$\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$ $\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$ $\begin{bmatrix} 4 & 3.5 \\ 3.5 & 4 \end{bmatrix}$ $\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$ $\begin{bmatrix} 8 & 3 \\ 3 & 2 \end{bmatrix}$ $\begin{bmatrix} 4 & -3.5 \\ -3.5 & 4 \end{bmatrix}$

AYUDA: mirar también las escalas de los ejes