Práctica 7

Estadística descriptiva bidimensional: diagrama de dispersión y coeficiente de correlación

En esta práctica aprenderemos a usar R Commander para representar gráficamente dos variables de forma simultánea, en particular, variables de tipo cuantitativo. Además, estudiaremos la relación entre las variables haciendo uso del coeficiente de correlación.

Contenido de la práctica

7.1. Diagrama de dispersión

La forma más habitual de describir gráficamente dos variables cuantitativas, X e Y, es el diagrama de dispersión. En él se representa mediante puntos en el plano cartesiano cada una de las observaciones de las variables, (x_i, y_i). Para realizar un diagrama de dispersión con R Commander utilizamos la opción:

Gráficas > Diagrama de dispersión...

En la pestaña *Datos* se eligen las dos variables que se desea representar. Hay que tener en cuenta, que si las dos variables consideradas pueden distinguirse como explicativa (x) y explicada (y), entonces cada una de ellas hay que seleccionarla en el cuadro correspondiente. En la pestaña *Opciones* existen diversas características del gráfico que permiten modificar distintos aspectos: las etiquetas de los ejes, el título del gráfico, que aparezca el diagrama de caja de las distribuciones marginales, etc.

Ejercicio 45: El archivo *Zaragoza_15_20.xlsx* contiene información sobre la temperatura (^oC) media y máxima mensual en la ciudad de Zaragoza, así como algunos aspectos demográficos: el número de nacimientos, el número de defunciones y el número de bodas (fuente: INE y IAEST).

1. Realiza un pequeño análisis descriptivo de las variables.

2. Construye un diagrama de dispersión que represente la temperatura media y la temperatura máxima de cada mes. ¿Qué tipo de relación observas entre las variables?

3. Selecciona la opción *Cajas de dispersión marginales* que aparece en la pestaña *Opciones*. ¿Qué puedes decir de la temperatura media mensual en Zaragoza?

4. En la pestaña *Opciones* también puedes especificar el tipo de símbolo utilizado para representar cada punto escribiendo un número en el cuadro *Dibujar los caracteres*. Observa qué ocurre si escribes cualquier número entre 0 y 25 (en particular, puedes probar con los números entre 15 y 20).

5. Para identificar a qué caso corresponde cada uno de los puntos dibujados en el diagrama de dispersión puedes hacerlo marcando, en *Identificar observaciones*, o bien la opción *Automáticamente* o bien la opción *Interactivamente con el ratón*. Observa qué ocurre si seleccionas la opción *Automáticamente* ¿por qué crees que marca esos puntos?

6. En lugar de utilizar números para identificar los distintos casos (filas) del conjunto de datos, es posible asignar nombres más representativos como, en este caso, pueden ser los valores contenidos en la variable *Fecha*. Hazlo utilizando la opción:

   Datos > *Conjunto de datos activo* > *Establecer nombres de casos...*

7. Realiza un diagrama de dispersión que represente la temperatura máxima y el número de bodas de cada mes.
   - ¿Qué relación observas entre las variables?
   - ¿Qué tienen en común los meses con más bodas?
   - ¿Qué tienen en común los meses con menos bodas?

   Utiliza la opción de indentificar observaciones *Interactivamente con el ratón* para responder a las dos últimas preguntas.

8. Como puedes observar, los datos no son homogéneos, ya que se mezclan observaciones *prepandémicas* y *pandémicas*. Esto puede provocar que la información aparezca distorsionada si se visualizan los datos de forma conjunta. Crea una nueva variable denominada *Pandemia* que tome los valores "No" y "Sí", según si los datos son anteriores a 2020 o no.

9. Realiza un diagrama de dispersión que represente la temperatura media y el número de bodas de cada mes, distinguiendo si los datos corresponde a meses de pandemia o no. Para ello, pulsa el botón *Gráfica por grupos...* de la pestaña *Datos* y selecciona la variable *Pandemia* (en este caso, para utilizar un tipo concreto de carácter para representar los puntos, tendrás que escribir tantos números como *grupos* haya, por ejemplo: *16,17*).

10. Antes de la pandemia, ¿qué relación observas entre la temperatura media y el número de bodas? ¿Y durante los meses de pandemia?

7.1.1. Matriz de diagramas de dispersión

Para estudiar, visualmente, la relación entre varios pares de variables al mismo tiempo utilizaremos la matriz de diagramas de dispersión, que se realiza accediendo al menú:

Gráficas > Matriz de diagramas de dispersión...

En la pestaña *Datos* se seleccionan 3 o más variables, de manera que R Commander dibujará todos los posibles diagramas de dispersión con las variables elegidas. Dado que el diagrama de dispersión de una variable consigo misma no tiene tanto interés, en la pestaña *Opciones* se puede especificar qué gráfico queremos colocar en su lugar (son los gráficos que aparecerán en la diagonal de la matriz). Por ejemplo, puede ser interesante representar el histograma de la variable o su diagrama de caja.

Ejercicio 46: Filtra el conjunto de datos contenido en el archivo *Zaragoza_15_20.xlsx*, para mantener únicamente los datos anteriores a 2020 (en este ejercicio y los posteriores utilizaremos el conjunto filtrado). Después, responde a las siguientes preguntas utilizando el archivo filtrado:

1. Dibuja una matriz de diagramas de dispersión con las variables: *Año, Bodas, Defunciones, Nacimientos* y *Temperatura.media*. Coloca en la diagonal el diagrama de caja correspondiente a cada variable.
2. ¿Entre qué variables observas una relación lineal positiva?
3. ¿Entre que variables observas una relación lineal negativa?
4. ¿Entre que variables observas otro tipo de relación o ausencia de relación?
5. Completa la siguientes frases sobre la relación que se observa entre los datos:
   - En general, a mayor temperatura media mensual, _____ número de bodas.
   - En general, a _____ temperatura media mensual, mayor número de defunciones.
   - En general, a _____ número de bodas, _____ número de defunciones.
   - En general, a mayor _____, menor número de nacimientos.
   - No se observa ningún tipo de relación entre la temperatura media y la variable _____.
6. A la vista de los diagramas de dispersión, trata de intuir el signo y la magnitud de los coeficientes de correlación correspondientes a cada par de variables con el fin de ordenarlos de menor a mayor.

7.2. Coeficiente de correlación

El coeficiente de correlación, r_{XY}, se calcula a través del menú:

Estadísticos > Resúmenes > Matriz de correlaciones...

En la ventana que aparece, se seleccionan las dos variables para las que queremos calcular el coeficiente de correlación (el resto de opciones se dejan como aparecen por defecto). Como resultado, nos proporcionará una matriz con los coeficientes:

$$\begin{matrix} r_{XX} & r_{XY} \\ r_{YX} & r_{YY} \end{matrix} \xrightarrow{\text{es decir}} \begin{matrix} 1 & r_{XY} \\ r_{XY} & 1 \end{matrix}$$

Como puede observarse, la matriz tendrá unos en la diagonal y será simétrica.

R Commander no permite calcular la covarianza de forma directa. Sin embargo, existen varias formas sencillas de hacerlo:

- Calculando las desviaciones típicas de las dos variables involucradas y, después, haciendo el cálculo: $s_{XY} = r_{XY} \cdot s_X \cdot s_Y$
- Modificando el comando que escribe R Commander en la ventana *R Script* para hacer el cálculo del coeficiente de correlación. Donde pone `cor(...)` habría que escribir `cov(...)`, es decir, simplemente cambiar `r` por `v` y ejecutar el comando completo.

Ejercicio 47: Realiza los siguientes apartados con el mismo conjunto de datos utilizando en el ejercicio 46:

1. Calcula el coeficiente de correlación de las variables *Temperatura.media* y *Temperatura.máxima*.
2. Observa la relación entre el coeficiente de correlación y el gráfico de dispersión.
3. Calcula la covarianza de las variables *Temperatura.media* y *Temperatura.máxima*.

7.2.1. Matriz de coeficientes de correlación

De forma análoga a la matriz de diagramas de dispersión, también es posible calcular una matriz con los coeficientes de correlación de varias parejas de variables. La forma de calcular es, de nuevo, mediante el menú:

Estadísticos > *Resúmenes* > *Matriz de correlaciones...*

En la ventana que aparece, se seleccionan todas las variables para las que queremos calcular el coeficiente de correlación (el resto de opciones se dejan como aparecen por defecto).

Ejercicio 48: Realiza los siguientes apartados utilizando el mismo conjunto del ejercicio 46:

1. Calcula una matriz de coeficientes de correlación con las variables: *Año, Bodas, Defunciones, Nacimientos* y *Temperatura.media*.
2. Observa la relación entre los distintos coeficientes de correlación y la matriz de gráficos de dispersión.
3. ¿Qué variables tienen mayor grado de relación lineal (positiva o negativa)?
4. ¿Que variables tienen menor grado de relación lineal?
5. Ordena los coeficientes de correlación de menor a mayor: ¿coincide la ordenación con la respuesta dada en el ejercicio 46?
6. Calcula la covarianza de cada pareja de variables. ¿Qué valor de la covarianza está más alejado del cero?
7. Hasta el año 2020, el mes con más defunciones había sido enero de 2017, cuyo número de defunciones estaba por encima de cualquier mes en los 50 años anteriores (en los gráficos que involucran el número de defunciones puedes observar que dicho dato se desmarca del resto). Observa cómo se modifica la matriz de coeficientes de correlación si filtras los datos para que dicho dato no sea tenido en cuenta.

8. A la vista de los coeficiente de correlación obtenidos:
   - ¿Podemos decir una de las causas de que un mes haya más defunciones es la disminución de la temperatura media de dicho mes?
   - ¿Podemos decir una de las causas de que un mes haya más bodas es una mayor temperatura media de dicho mes?
   - ¿Podemos decir que una de las causas de que un mes haya menos defunciones es el aumento en el número de bodas?

7.3. Ejercicios propuestos

Ejercicio 49: El conjunto de datos *Vino.RData* proporciona datos sobre el consumo de vino (en litros de alcohol, procedente del vino, por cada 100 000 personas) y sobre las muertes anuales por ataques al corazón (muertos por cada 100 000 personas) en 19 países desarrollados.

Las variables que contiene este fichero son:

- *pais*: país
- *cons*: consumo de alcohol procedente del vino (litros por cada 100 000 habitantes).
- *nmac*: tasa de muertes por ataques al corazón (muertos por cada 100 000 personas).

1. Dibuja un diagrama de dispersión que muestre cómo el consumo nacional de vino ayuda a explicar las muertes por ataques al corazón. Describe la forma de la relación. ¿Crees que existe una relación lineal? Dicha relación, ¿es positiva o negativa? Explica de forma simple qué dice la relación sobre el consumo de vino y los ataques al corazón.
2. ¿Proporcionan estos datos una clara evidencia de que tomar vino causa una reducción de las muertes por ataques al corazón? ¿Por qué?

Ejercicio 50: El archivo *Variables.RData* contiene la información de 5 variables numéricas. Trata de responder a las siguientes preguntas (las que puedas) primero a partir de la matriz de coeficientes de correlación y, depués, a partir de la matriz diagramas de dispersión:

1. ¿Para que variables existe una estrecha relación lineal positiva? ¿Y negativa?
2. ¿Para qué variables no se observa ningún tipo de relación lineal (positiva o negativa)?
3. ¿Para qué variables no se observa ningún tipo de relación?
4. ¿Para qué variables se observa algún tipo de relación no lineal?