PRÁCTICA 1: GESTIÓN DE DATOS Y ESTADÍSTICA UNIVARIANTE

Atanasov Angelov, Daniel – daniel.atanasov24@estudiantes.uva.es

Sanz Tomé, Raúl – raul.sanz24@estudiantes.uva.es

Resumen:

En esta práctica de estadística con Python, analizamos un conjunto de datos mediante resúmenes estadísticos y tablas de frecuencias. Visualizamos la distribución con histogramas, diagramas de barras y sectores, y otros tipos de representaciones gráficas. Tras estos análisis, interpretamos los resultados obtenidos en cada paso y sacamos nuestras conclusiones relacionadas con el uso de los distintos elementos para comprender mejor la información.

1. Introducción

En esta práctica hemos trabajado con un conjunto de datos estadísticos de una sola variable, con el propósito de aprender a gestionarlos, manejar ficheros y analizar descriptivamente los resultados obtenidos. Estos datos son las estaturas y el sexo de 60 personas, estando ambas variables relacionadas de forma que el primer valor de cada variable es de la misma persona y así hasta los últimos datos.

2. Metodología

Los procesos que hemos seguido han sido los siguientes: comenzamos empleamos las bibliotecas necesarias para el manejo de esos datos, como numpy, pandas, matplotlib, entre otras.

Después hemos elaborado la tabla de frecuencias relacionada con los datos y, a continuación, hemos hecho el resumen estadístico con el que obtenemos los valores de la media, moda, mediana, y más.

También hemos realizado las correspondientes representaciones gráficas de los datos. Estas representaciones incluyen gráficos de dispersión, gráficos de caja y bigotes, histogramas, diagramas de barras y diagramas de sectores.

Para finalizar, hemos incorporado el dataset y trabajado con él de las distintas formas anteriormente empleadas.

3. Resultados

Para empezar, hemos realizado tablas de frecuencias, las cuales muestran de forma ordenada los datos y sus correspondientes frecuencias absolutas y relativas, tanto acumuladas como no. Junto a esto, hemos generado un gráfico de dispersión y otro de caja y bigotes. En el primero hemos visto como se dispersan los datos de cada una de las variantes, en concreto, las estaturas de las mujeres y de los hombres. Ambos poseen una mayor concentración en torno a los valores centrales o intermedios. Sin embargo, la distribución de las mujeres nos muestra que estas tienen más valores extremos y que, además, están más alejados de los valores centrales. En cuanto al segundo gráfico, observamos que el rango intercuartílico de la estatura de las mujeres se encuentra entre 1,62 y 1,82; mientras que en el caso de los hombres está entre 1,68 y 1, 85, aproximadamente en ambos casos. Además, observamos que la mediana en el caso de los hombres es menor que en el de las mujeres y, como veíamos en el gráfico anterior, los datos extremos en el caso de las mujeres llegan a estar más alejados de la zona central. Además, hemos realizado un histograma en el que vemos que hay más datos de mujeres, ya que la frecuencia es mayor, que los datos llegan hasta valores más extremos (1,4-2,0m) y que estos están más concentrados entorno a los 1,6 - 1,8 metros. En el caso de los hombres encontramos unas frecuencias menores, más dispersas y con una concentración también menor.

Después de realizar un resumen estadístico nuevo y otra tabla de frecuencias, hemos representado un diagrama de barras y otro de sectores. En el de barras vemos que las mujeres llegan a tener alturas ligeramente mayores, mientras que en el de sectores observamos que hay el doble de mujeres que de hombres en los datos.

A continuación, volvemos a calcular las medias de estatura para hombres y para mujeres y realizamos nuevamente los histogramas de cada uno, idénticos a los descritos anteriormente. Volvemos a realizar los histogramas, pero esta vez, les cambiamos ciertas características, como el color, que ahora es verde.

Para finalizar, después de cargar el dataset generamos una tabla de frecuencias, un diagrama de barras y otro de sectores de la variable “año de fabricación”. En ambos diagramas observamos que los datos tienen unas frecuencias bastante similares, es decir, que a lo largo de los años se produjeron coches en unas cantidades parecidas. Aunque hay un repunte en el año 73, donde se fabricaron unos cuantos coches más., algo que también sucede en los años 76 y 78 pero en menor medida.

Ahora trabajamos con el consumo de los vehículos. Generamos una tabla de frecuencias con intervalos de amplitud 22. Esto nos muestra que la mayoría de automóviles tienen un consumo entre 22 y 44 mpg, unos pocos menos poseen un consumo inferior a 22 mpg y el resto por encima de 44 mpg, aunque son muy pocos.

En cuanto al porcentaje de vehículos con un consumo inferior a 22 mpg, el resultado es del 46, 23%.

A continuación, hemos generado el gráfico de caja y bigotes (box-and-whisker) de la variable mpg. Este gráfico nos muestra que el rango intercuartílico se encuentra entre 17 y 30 mpg aproximadamente, que la mediana está próxima a 25 mpg y que los valores extremos por encima están más alejados que los valores extremos inferiores. Además, hemos realizado el gráfico que representa la función de distribución, se trata de un gráfico de distribución acumulada. Este gráfico nos muestra que los datos se encuentran entre 10 y 45 mpg y que hay una mayor concentración entre 15 y 35 mpg porque la pendiente en ese intervalo es mayor.

Por último, hemos realizado el polígono de frecuencias acumuladas de la variable "horsepower" (potencia – caballos). Este gráfico nos muestra que más del 60% de los coches tienen menos de 100CV, mientras que por encima de 150CV se encuentran el 15% de los vehículos aproximadamente. Apoyándonos en este polígono de frecuencias relativas acumuladas, calculamos que la cantidad de vehículos que tienen 100CV o menos es de 242.

4. Conclusiones

Después de hacer todos estos cálculos y gráficos, nos hemos dado cuenta de lo útiles que son para entender mejor los datos y sacar conclusiones más claras. Representaciones como histogramas, diagramas de barras o polígonos de frecuencia nos han ayudado a identificar patrones, ver cómo se distribuyen los valores y detectar posibles valores atípicos.

También hemos aprendido que no se trata solo de hacer cálculos, sino de saber qué estamos analizando en cada momento y elegir la herramienta adecuada para interpretar los resultados correctamente. En definitiva, trabajar con datos estadísticos va más allá de los números: es encontrar sentido a la información.