

Práctica 9

Estadística descriptiva bidimensional: una variable cualitativa y una cuantitativa

En esta práctica aprenderemos a usar R Commander para estudiar la relación entre variables de distinta tipología, en particular, cuando una de ellas es cualitativa y la otra es cuantitativa.

Contenido de la práctica

9.1. Medidas de síntesis por grupos	62
9.2. Gráficas por grupos	62
9.2.1. Histograma	62
9.2.2. Diagrama de caja múltiple	63
9.2.3. Gráfico de medias y desviaciones	63
9.3. Ejercicios propuestos	64

Para ilustrar esta práctica vamos a utilizar los datos que aparecen en el fichero *Salarios.RData*. Este fichero ha sido generado a partir de los datos publicados por el INE en la Encuesta de Estructura Salarial correspondiente al año 2018. Cada una de las filas del conjunto de datos contiene la información de un trabajador, siendo las variables:

- *Edad*: edad del trabajador.
- *Estudios*: nivel máximo de estudios alcanzado.
- *Zona*: zona geográfica según la Nomenclatura de las Unidades Territoriales Estadísticas (NUTS).
- *Responsable*: responsabilidad en organización y/o supervisión.
- *Sexo*: sexo del trabajador.
- *Duracion.Contrato*: duración del contrato (indefinida o determinada).
- *Tipo.Jornada*: completa o parcial.
- *Salario*: salario bruto anual, en euros.

Ejercicio 58: Observa, mediante un histograma y un diagrama de caja, los valores que toma la variable *Salario*, ¿qué observas? Después, filtra el conjunto de datos para eliminar aquellos casos de trabajadores con un salario anual superior a 100000 €, ¿cuántos casos se han eliminado? En adelante, trabajaremos con el conjunto de datos filtrado.

9.1. Medidas de síntesis por grupos

En general, para estudiar dos variables, una cualitativa y otra cuantitativa, se forman grupos según los valores que toma la variable cualitativa y, después, se analiza la variable cuantitativa en cada grupo como si de un análisis univariante se tratara.

El análisis de una variable numérica se realiza mediante el cálculo de sus estadísticos más habituales. Para ello, se selecciona la opción:

Estadísticos > Resúmenes > Resúmenes numéricos...

En la pestaña *Datos* se selecciona la variable numérica y también, tras pulsar el botón *Resumir por grupos...*, la variable cualitativa a tener en cuenta. En la pestaña *Estadísticos* se marcan los estadísticos que se desean calcular.

Ejercicio 59: Estudia el salario de los trabajadores según su sexo y responde:

1. ¿Cuántas mujeres y hombres hay?
2. ¿Cuál es el salario medio de los hombres? ¿y de las mujeres?
3. ¿Para que sexo hay mas variabilidad de salarios? Interpreta los valores.
4. ¿Cómo es la asimetría en cada uno de los sexos? Interpreta los valores.
5. ¿Qué salario reciben, como mucho el 50% de las mujeres que menos dinero cobra?
6. ¿Qué salario reciben, como mínimo, el 20% de los hombres que más dinero cobran.
7. ¿Son *Salario* y *Sexo* variables independientes para estos trabajos?

9.2. Gráficas por grupos

9.2.1. Histograma

Para realizar un histograma por grupos se selecciona la opción:

Gráficas > Histograma...

En la pestaña *Datos* se selecciona la variable numérica y también, tras pulsar el botón *Gráfica por grupos...*, la variable cualitativa a tener en cuenta. En la pestaña *Opciones* se pueden especificar distintas opciones para el gráfico.

Observa que los histogramas que se dibujan utilizando la misma escala para los ejes de abscisas y ordenadas, de manera que son fácilmente comparables.

Ejercicio 60: Representa, mediante histogramas, la variable salario dependiendo de si el trabajador tiene responsabilidad o no en la empresa.

1. Observa las diferencias entre realizar el histograma con frecuencias absolutas y con porcentajes.
2. ¿En qué caso se observa mayor proporción de salarios altos?
3. ¿En qué caso se observa mayor número de trabajadores con salario superior a 20000 €?
4. ¿En qué caso se observa mayor proporción de trabajadores con salario superior a 20000 €?
5. ¿En que caso es mayor la asimetría? ¿y el apuntamiento?
6. Comprueba la respuesta a la pregunta anterior calculando los coeficientes de asimetría y apuntamiento.
7. ¿Son *Salario* y *Responsable* variables independientes para estos trabajadores?

9.2.2. Diagrama de caja múltiple

Para realizar un diagrama de caja múltiple se selecciona la opción:

Gráficas > Diagrama de caja...

En la pestaña *Datos* se selecciona la variable numérica y también, tras pulsar el botón *Gráfica por grupos...*, la variable cualitativa a tener en cuenta. En la pestaña *Opciones* se pueden especificar distintas opciones para el gráfico.

Ejercicio 61: Representa, mediante un diagrama de caja múltiple, el salario de los trabajadores dependiendo de su nivel de estudios.

1. ¿Para qué nivel de estudios es mayor el salario mediano?
2. ¿Cuál es el salario mediano de aquellos que han estudiado hasta bachillerato o grado medio?
3. ¿Para qué nivel de estudios hay menos casos atípicos de salario?
4. Localiza a la persona con mayor salario de cuyo nivel de estudios es primaria, ¿en qué zona vive? ¿qué edad tiene? ¿ocupa algún cargo de responsabilidad?
5. Teniendo en cuenta que *Estudios* es una variable ordinal, ¿qué relación observas entre las dos variables?

9.2.3. Gráfico de medias y desviaciones

Para realizar un diagrama de medias y desviaciones típicas se selecciona la opción:

Gráficas > Diagrama de las medias...

En la pestaña *Datos* se selecciona la variable cualitativa (primer recuadro) y la variable numérica (segundo recuadro). En la pestaña *Opciones* se marca *Desviaciones típicas* en el apartado *Barras de error*.

Ejercicio 62: Representa, mediante un gráfico de medias y desviaciones, el salario de los trabajadores dependiendo de su rango de edad.

1. ¿Para qué rango de edad es mayor el salario medio?
2. ¿Para qué rango de edad hay menos variabilidad?
3. ¿Cuál es el salario medio de los más jóvenes?

Este tipo de gráfico permite representar, además de la variable numérica, dos variables cualitativas al mismo tiempo. Para ello, basta seleccionar dos variables cualitativas en el apartado *Factores* de la pestaña *Datos*. Representa, mediante un gráfico de medias y desviaciones, el salario de los trabajadores dependiendo de su rango de edad y de si trabajan a jornada completa o parcial. Comenta el gráfico.

9.3. Ejercicios propuestos

Ejercicio 63: Responde a las siguientes preguntas haciendo uso del conjunto de datos utilizado a lo largo de la práctica:

1. ¿En qué zona geográfica es mayor/menor el salario medio? ¿y el mediano?
2. Compara la variabilidad de los salarios en cada zona, ¿dónde es mayor/menor?
3. ¿Cuál es el salario que recibe, como máximo, el 70 % de los trabajadores de la Comunidad de Madrid que menos cobra?
4. Representa gráficamente, mediante histogramas con frecuencias relativas, el salario que reciben los trabajadores según su zona geográfica. ¿En cuál observas mayor/menor asimetría? ¿En cuál observas mayor/menor apuntamiento?
5. Calcula los coeficientes de asimetría y apuntamiento para la variable *Salario* para cada *Zona* y compara los resultados con las respuestas dadas en el ejercicio anterior.

Ejercicio 64: Crea un gráfico que refleje la información descrita en las siguientes afirmaciones:

1. El salario mediano de las mujeres es menor que el de los hombres.
2. La distribución de salarios para los trabajadores con mayor nivel de estudios es más simétrica que para los trabajadores del resto de niveles de estudios.
3. En el conjunto de datos, hay menos datos relativos al salario de trabajadores con un contrato de duración determinada que de trabajadores con un contrato indefinido.
4. Tanto la media como la desviación típica es mayor para los salarios de trabajadores a jornada completa, en comparación con los trabajadores a jornada parcial.
5. La mitad de los trabajadores que más salario recibe cuyo nivel de estudios es de Licenciatura (o superior) recibe más salario que casi todos aquellos que no alcanzan el nivel de estudio de Educación Primaria.

Ejercicio 65: En el portal del Ministerio de Universidades se ofrece información estadística relativa al sistema universitario español. Descarga el archivo que se encuentra en el siguiente enlace, que contiene información sobre el número de alumnos matriculados en cada uno de los grados y másteres que se imparten en España durante los últimos seis cursos académicos :

<https://www.universidades.gob.es/stfls/universidades/Estadisticas/ficheros/MatriculadosEEU.xlsx>

Vamos a analizar la información relativa a los dos últimos cursos académicos (2019/2020 y 2020/2021) para los estudios de grado.

1. Antes de importar los datos en R Commander, modifica la hoja *Matriculados Grado* del fichero de Excel, de manera que la primera fila de la tabla contenga directamente el nombre de las siguientes variables:

- Comunidad autónoma
- Universidad
- Rama
- Titulación
- Matriculados 20_21
- Porcentaje Mujeres 20_21
- Matriculados 19_20
- Porcentaje Mujeres 19_20

El resto de información puedes borrarla del fichero de Excel ya que no la vamos a utilizar.

2. Importa los datos relativos a estudios de grado en R Commander (pon atención a los datos ausentes).

3. Realiza un breve análisis de los datos y responde a las siguientes preguntas:

- a) ¿Cuáles son las ramas del conocimiento?
- b) ¿Cuántos grados se imparten en Cataluña?
- c) ¿Qué porcentaje de los grados se imparten en la Comunidad de Madrid?
- d) ¿Cuántos grados hay en la Universidad de Zaragoza?
- e) ¿Cuál es la universidad con más grados universitarios? ¿Y la que tiene menos?
- f) En el curso 2020/2021, ¿cuántos estudiantes tiene el grado con más matriculados? ¿Qué grado es?
- g) ¿Qué universidades tienen grados con mayor número de matriculados en el curso 2020/2021?
- h) En el curso 2020/2021, ¿cuál es el número medio de estudiantes matriculados en un grado?
- i) ¿Es simétrica la distribución del número de matriculados en un grado durante el curso 2020/2021?
- j) Si consideramos el 25 % de los grados en los que la proporción de mujeres es menor durante el curso 2020/2021, ¿qué porcentaje de mujeres hay, como mucho, en dichos grados?

4. Representa gráficamente el número de grados según la rama de conocimiento. ¿Qué rama de conocimiento tiene menos titulaciones?
5. Representa gráficamente la proporción de mujeres matriculadas en los grados del curso 2019/2020 y en el 2020/2021. ¿Son simétricas las distribuciones?
6. Representa gráficamente la proporción de mujeres matriculadas en los grados del curso 2020/2021 para cada una de las ramas del conocimiento. ¿En qué rama se observa una distribución asimétrica negativa? ¿cómo se interpreta este hecho?
7. Representa gráficamente el número de matriculados en los grados del curso 2020/2021, ¿qué observas? Filtra los datos para mejorar la representación, considerando únicamente los grados con menos de 2500 matriculados.
8. Utilizando el conjunto de datos filtrado en el apartado anterior, estudia la relación entre las variables numéricas. Explica el porqué de la relación o ausencia de relación que se observa entre cada par de variables.

Filtra el conjunto de datos para mantener únicamente aquellos que corresponden a universidades de Aragón. En adelante trabajaremos con el fichero filtrado.

9. ¿Cuántos grados tiene cada una de las universidades aragonesas?
10. ¿Cuál es el grado con más alumnos matriculados?
11. ¿Cuántos grados de Ingeniería y Arquitectura hay en la Universidad de Zaragoza?
12. De los grados que se imparten en la Universidad San Jorge, ¿qué porcentaje son de Ciencias de la Salud?
13. ¿En qué universidad hay más grados de Ciencias Sociales y Jurídicas?
14. ¿En qué universidad hay mayor proporción de grados de Ciencias Sociales y Jurídicas?
15. Realiza un gráfico que permita responder a la pregunta anterior.
16. Calcula el número de mujeres en cada titulación en los cursos 2019/2020 y 2020/2021.
17. ¿Qué grado tiene mayor número de mujeres matriculadas? ¿y mayor proporción?
18. Si un grado tuvo 30 estudiantes matriculados en el curso 2019/2020, ¿cuántos se espera haya tenido en el curso 2020/2021? ¿Es fiable como predicción?