Práctica 3

Estadística descriptiva unidimensional: medidas de síntesis

En esta práctica se describe cómo calcular las medidas de síntesis, que permiten resumir distintos aspectos de un conjunto de datos: cuál es el centro de los datos, la variabilidad, la localización, y la dispersión.

Contenido de la práctica

3.1. Medidas de centralización: media, mediana y moda **23**
3.2. Medidas de localización: cuantiles **24**
3.3. Medidas de dispersión: rango, rango intercuartílico, varianza, desviación típica y coeficiente de variación **25**
3.4. Medidas de forma: coeficientes de asimetría y curtosis **26**
3.5. Diagrama de caja **26**
3.6. Ejercicios propuestos **28**

3.1. Medidas de centralización: media, mediana y moda

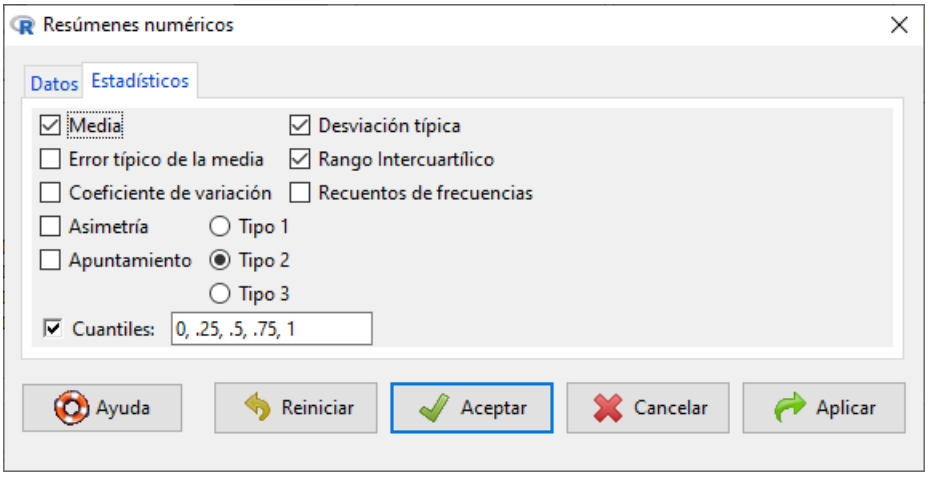
En la primera práctica se vio cómo, al realizar un resumen del conjunto de datos, R proporcionaba, entre otros valores, el valor de la media y la mediana para las variables numéricas. Recuerda que para realizar ese resumen se utiliza: *Estadísticos > Resúmenes > Conjunto de datos activo*

Ejercicio 17: Realiza un resumen de las variables contenidas en el archivo *Informacion.RData*, que contiene información relativa a varias entrevistas de trabajo. Observa los valores de la media y la mediana que se han calculado para las variables numéricas: ¿se parecen? ¿cuántos años de experiencia tienen, como máximo, la mitad de los entrevistados que menos tiempo llevan trabajando?

En ocasiones, el conjunto de datos puede tener demasiadas variables como para poder localizar fácilmente la media y la mediana de alguna de ellas dentro del resumen de todas las variables. Por ello, cuando necesitemos información específica sobre alguna de las variables numéricas utilizaremos:

Estadísticos > Resúmenes > Resúmenes numéricos

En la pestaña *Datos* se elige la variable o variables que quieras resumir. En la pestaña *Estadísticos*, puedes elegir qué medidas descriptivas quieres como resultado. Por defecto, se calcula la media de la variable, su desviación típica, el rango intercuartílico, los cuartiles (cuantiles 0.25, 0.5 y 0.75), el máximo (cuantil 1) y el mínimo (cuantil 0).



Tras pulsar *Aceptar* observa que, además de los estadísticos que hayas seleccionado, también aparece como resultado el número de datos disponible para la variable seleccionada.

Ejercicio 18: Calcula la media y la mediana de las variables *Peso* y *Altura* del archivo *Informacion.RData*.

Para calcular la moda (o las modas) utilizaremos la tabla de frecuencias y buscaremos en ella cuál es el valor (o valores) mas frecuentes:

- En las variables de tipo factor la tabla de frecuencias se calcula con: *Estadísticos* > *Resúmenes* > *Distribuciones de frecuencias...*
- En las variables de tipo numérico la tabla de frecuencias se calcula con: *Estadísticos* > *Resúmenes* > *Resúmenes numéricos...* (hay que activar la opción *Recuento de frecuencias* de la pestaña *Estadísticos*)

Ten en cuenta que en el caso de variables numéricas que tomen muchos valores distintos, éstos se agruparán en intervalos y lo que estarás calculando, en lugar de la moda, será el intervalo modal.

Ejercicio 19: Calcula la moda (o intervalo modal) de las variables *Peso* y *Lugar* del archivo *Informacion.RData*.

3.2. Medidas de localización: cuantiles

Para las variables numéricas utilizaremos los cuantiles para determinar la localización específica de una parte de los datos. A menudo, los cuantiles reciben otras denominaciones:

- Percentiles: son los cuantiles $0.01, 0.02, 0.03, \ldots 0.98, 0.99$.
- Deciles: son los cuantiles $0.10, 0.20, 0.30, \ldots, 0.90$

- Cuartiles: son los cuantiles $0.25, 0.50, 0.75$
- Mediana: es el cuantil 0.50

Para calcular los cuantiles de una variable numérica utilizaremos:

Estadísticos > *Resúmenes* > *Resúmenes numéricos*

En la pestaña *Datos* se elige la variable y en la pestaña *Estadísticos* puedes indicar todos aquellos cuantiles que quieras calcular (separados por comas).

Ejercicio 20: Con la variable *Peso* del archivo *Informacion.RData*, calcula:

- El percentil 80.
- El primer cuartil.
- El segundo decil.
- ¿Cuánto pesa, como mucho, el 60 % de los menos pesa?
- ¿Cuánto pesa, como mínimo, el 90 % de los que mas pesan?
- El 20 % de los que menos pesan, ¿cuántos kilos pesa a lo sumo?

3.3. Medidas de dispersión: rango, rango intercuartílico, varianza, desviación típica y coeficiente de variación

R Commander permite calcular fácilmente algunas de las medidas de dispersión más habituales, como son el rango intercuartílico, la desviación típica o el coeficientes de variación. Todas ellas puede calcularse utilizando el menú:

Estadísticos > *Resúmenes* > *Resúmenes numéricos*

En la pestaña *Datos* se elige la variable y en la pestaña *Estadísticos* puedes seleccionar los estadísticos mencionados para que los calcule R.

Observa que el resto de estadísticos utilizados para estudiar la variabilidad, como son el rango y la varianza, los puedes obtener utilizando R como calculadora.

Ejercicio 21: Compara la variabilidad de las variables *Altura* y *Peso* del archivo *Informacion.RData*. ¿Qué medida es la más indicada para hacer la comparación?

Ejercicio 22: A partir de la variable *Altura* del archivo *Informacion.RData*, calcula una nueva variable llamada *AlturaM* que contenga el valor de la altura en metros, en lugar de centímetros. Después, compara la media, la mediana, la desviación típica, la varianza, el coeficiente de variación, el rango y el rango intercuartílico de ambas variables.

3.4. Medidas de forma: coeficientes de asimetría y curtosis

Los coeficientes de asimetría y curtosis de una variable numérica pueden calcularse utilizando el menú:

Estadísticos > *Resúmenes* > *Resúmenes numéricos*

En la pestaña *Datos* se elige la variable y en la pestaña *Estadísticos* se selecciona *Asimetría* y *Apuntamiento*

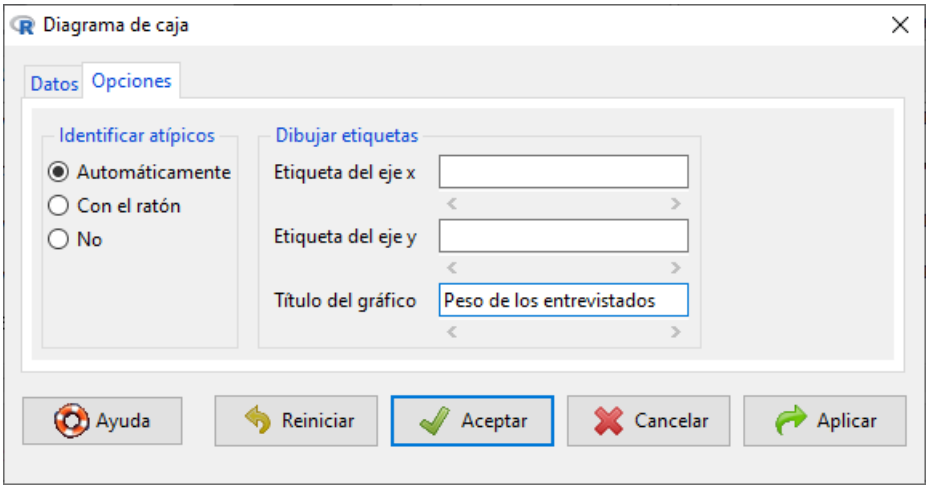
Ejercicio 23: Representa gráficamente las variables *Peso* y *Altura* del archivo *Informacion.RData*: ¿qué variable presenta mayor asimetría? ¿de qué tipo es en cada caso? Después, calcula el coeficiente de asimetria de ambas variables.

Ejercicio 24: Representa gráficamente las variables *Peso* y *Altura* del archivo *Informacion.RData*: ¿qué variable presenta mayor apuntamiento? Después, calcula el coeficiente de apuntamiento de ambas variables.

3.5. Diagrama de caja

Para obtener el diagrama de caja de una variable numérica, como la variable *Peso*:

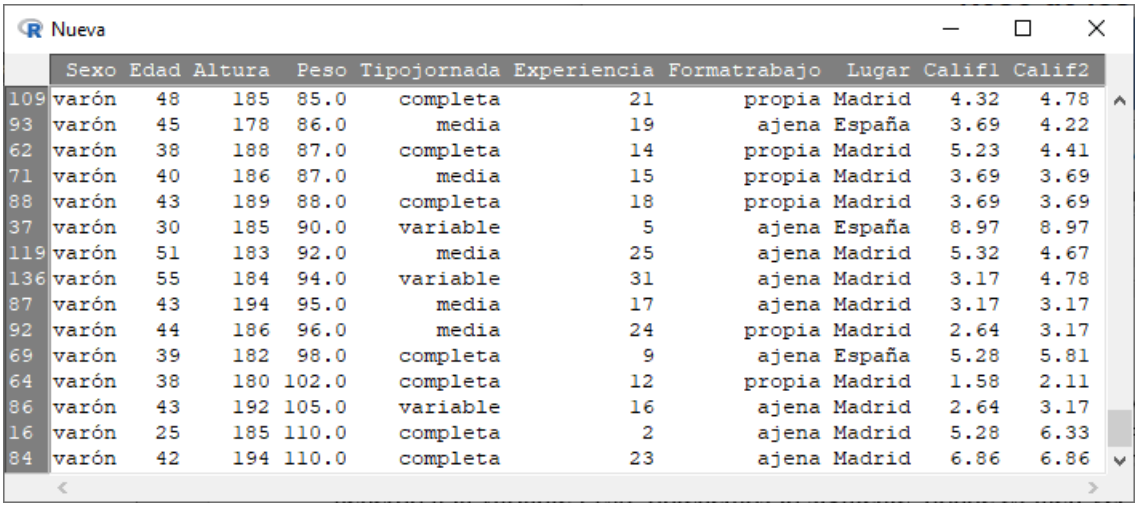
1. Selecciona en el menú principal *Gráficas* > *Diagrama de caja*.
2. En la pestaña *Datos* del cuadro de diálogo que aparece, elige la variable *Peso*.
3. En la pestaña *Opciones*, puedes elegir la identificación automática de los datos atípicos o manual (con el ratón). Se recomienda dejar la opción *Automáticamente*. También se pueden elegir las etiquetas en los ejes y un título para el gráfico.



Al pulsar *Aceptar* obtendremos el diagrama de caja. Observa la existencia de varios datos atípico: los casos 16, 64, 84 y 86. Las etiquetas de los casos atípicos aparecen en el gráfico y en el cuadro de resultados.

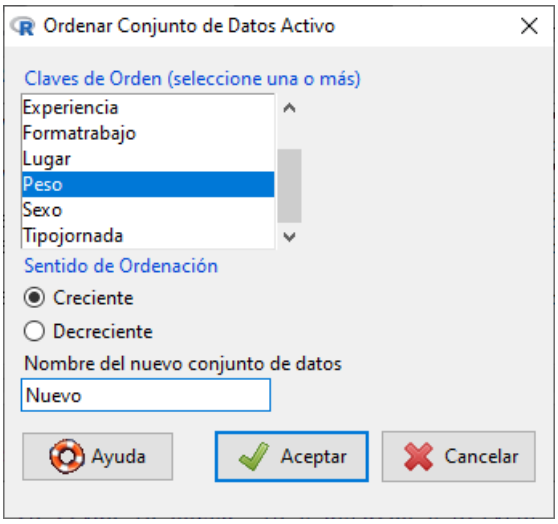
```
> Boxplot( ~ Peso, data=Información, id=list(method="y"), ylab="",
+   main="Peso de los entrevistados")
[1] "16" "64" "84" "86"
```

Para saber los valores correspondientes a esos casos atípicos, hay que acudir al conjunto de datos. A veces es útil recurrir a la opción *Datos* > *Conjunto de datos activo* > *Ordenar el conjunto de datos activo*. Si usamos esa opción para crear un conjunto de datos llamado *Nuevo*, ordenando con los datos de acuerdo a la variable *Peso*, obtenemos lo siguiente, donde es fácil ver que a los casos 64, 86, 16 y 84 les corresponden unos pesos de 102, 105, 110 y 110 kg, respectivamente.



	Sexo	Edad	Altura	Peso	Tipojornada	Experiencia	Formatrabajo	Lugar	Calif1	Calif2
109	varón	48	185	85.0	completa	21	propia	Madrid	4.32	4.78
93	varón	45	178	86.0	media	19	ajena	España	3.69	4.22
62	varón	38	188	87.0	completa	14	propia	Madrid	5.23	4.41
71	varón	40	186	87.0	media	15	propia	Madrid	3.69	3.69
88	varón	43	189	88.0	completa	18	propia	Madrid	3.69	3.69
37	varón	30	185	90.0	variable	5	ajena	España	8.97	8.97
119	varón	51	183	92.0	media	25	ajena	Madrid	5.32	4.67
136	varón	55	184	94.0	variable	31	ajena	Madrid	3.17	4.78
87	varón	43	194	95.0	media	17	ajena	Madrid	3.17	3.17
92	varón	44	186	96.0	media	24	propia	Madrid	2.64	3.17
69	varón	39	182	98.0	completa	9	ajena	España	5.28	5.81
64	varón	38	180	102.0	completa	12	propia	Madrid	1.58	2.11
86	varón	43	192	105.0	variable	16	ajena	Madrid	2.64	3.17
16	varón	25	185	110.0	completa	2	ajena	Madrid	5.28	6.33
84	varón	42	194	110.0	completa	23	ajena	Madrid	6.86	6.86

La forma de rellenar del cuadro de diálogo asociado a la opción *Datos* > *Conjunto de datos activo* > *Ordenar el conjunto de datos activo* se muestra a continuación.



3.6. Ejercicios propuestos

Ejercicio 25: En el fichero Informacion.RData, que recoge información sobre un conjunto de personas entrevistadas por un departamento de recursos humanos, se encuentran, entre otras, las variables *Sexo* (varon o mujer), *Lugar* (si el entrevistado prefiere trabajar en Madrid o no le importaría trabajar en otro lugar de España), y *Calif1* y *Calif2* (puntuaciones otorgadas a cada entrevistado por dos personas del departamento de recursos humanos).

1. Compara la media y la mediana de la variable *Calif1* e interpretalas. ¿Crees que existe algún tipo de asimetría en los datos?

2. Representa gráficamente la variable *Calif1* utilizando un diagrama de caja y comenta toda la información que observes en él.

3. Representa gráficamente la variable *Calif1* utilizando un histograma y calcula su coeficiente de asimetría.

4. ¿Que nota (*Calif1*) han sacado, como mínimo, el 10 % de los entrevistados que han sacado mayor nota?

5. Compara la variabilidad en las calificaciones de cada uno de los entrevistadores, ¿cuál de los dos es más homogéneo?

6. Calcula una nueva variable *CalifM* que sea la media aritmética de las calificaciones *Calif1* y *Calif2*.

7. Calcula, a partir de la variable *CalifM*, una nueva variable *CalifCuali* que contenga el valor cualitativo de las calificaciones (suspenso, aprobado, notable y sobresaliente).

8. Representa gráficamente, utilizando un diagrama de sectores, la variable *CalifCuali* (no olvides ordenar los valores de la variable, ya que es cualitativa ordinal). A la vista del diagrama de sectores, ¿puedes decir cuál es el valor del primer cuartil, la mediana y el tercer cuartil?.

9. Calcula el primer cuartil, la mediana y el tercer cuartil de la variable *CalifM* y observa si coinciden sus valores con tus respuestas al apartado anterior.

10. Obtén el diagrama de caja de la variable *CalifM*. ¿Observas algún valor atípico? En caso afirmativo, comenta qué casos son y qué calificaciones han obtenido por parte de cada uno de los entrevistadores.

Ejercicio 26: En un seguimiento realizado a un grupo de 284 clientes de una empresa, se ha anotado el número de veces que los clientes han realizado reclamaciones a la empresa en el último año, para resolver alguna incidencia. Así, un cliente puede que no haya tenido que reclamar nunca a lo largo del año, y otro puede que haya tenido que reclamar 4 veces. Los datos se encuentran en el archivo *Veces.RData*.

1. Calcula la mediana e interprétala.

2. Representa gráficamente el número de reclamaciones utilizando un diagrama de caja. Explica el porqué de la forma de dicho diagrama.

3. Utiliza un diagrama de barras para representar gráficamente el número de reclamaciones.

4. Si preguntamos a 213 de los 284 clientes (un 75 %), elegidos al azar, seguro que alguno de ellos ha reclamado, al menos, ___ veces.

Ejercicio 27: El archivo *Encuesta Continua Hogares INE 2020.xlsx* contiene algunos de los datos obtenidos por el Instituto Nacional de Estadística en la Encuesta Continua de Hogares del año 2020. Esta encuesta ofrece información anual sobre las características demográficas básicas de la población, de los hogares que componen y de las viviendas que habitan. Para los hogares aporta información sobre su tamaño y composición y para las viviendas sobre su régimen de tenencia, superficie útil, habitaciones, año de edificación y tipología del edificio. Las variables que se incluyen en el archivo, que hacen referencia a las viviendas y a los hogares analizados, son:

- *TamanoMunicipio*: tamaño del municipio en el que se ubica
- *Provincia*: provincia.
- *Comunidad*: comunidad autónoma.
- *RegimenVivienda*: régimen de tenencia.
- *Cocina*: disponibilidad de cocina.
- *Aseos*: número de cuartos de baño o aseos.
- *Comedores*: número de comedores.
- *Dormitorios*: número de dormitorios.
- *Trasteros*: número de trasteros.
- *OtrasHabitaciones*: número de otras habitaciones.
- *TotalHabitaciones*: número de habitaciones.
- *SuperficieVivienda*: superficie útil de la vivienda, en metros cuadrados.
- *TipoVivienda*: tipo de vivienda.
- *TotalPersonas*: número total de personas que habitan en la vivienda.
- *TipoHogar*: tipo de hogar.

Responde a las siguientes preguntas:

1. ¿De qué tipo son cada una de las variables? Haz un resumen de todas las variables y observa, globalmente, qué valores toman.
2. ¿Cuántas viviendas se han analizado?
3. ¿Cuántos valores distintos toma la variable *Aseos*?
4. ¿Qué porcentaje de viviendas no dispone de cocina? ¿Y de baño?
5. ¿Cuántas personas viven, de media, en las viviendas analizadas?
6. ¿Cuál es el tipo de hogar mas frecuente? ¿Y el tipo de vivienda más frecuente?
7. Representa los distintos tipos de regímenes de tenencia de viviendas, utilizando tablas y gráficos.
8. ¿Cuál es el tamaño medio de las viviendas? ¿Y el tamaño mediano?
9. Calcula todos los deciles de la variable *SuperficieVivienda*.

10. Representa la superficie de las distintas viviendas, utilizando tablas y gráficos. ¿Observas algo raro en los gráficos?

11. ¿Cuál es la superficie del 0.1 % de las viviendas que más ocupan?

12. Filtra el conjunto de datos (sin cambiarle el nombre) y guarda únicamente aquellos datos de viviendas con una superficie de menos de $500\,m^2$. Repite ahora la representación gráfica de la superficie de las viviendas.

13. Dibuja el diagrama de caja de la variable *SuperficieVivienda*. ¿Observas asimetría en la distribución de los datos? ¿De qué tipo? Después, calcula el coeficiente de asimetría.

14. Calcula una nueva variable llamada *MetrosCuadradosPorPersona*, que represente el número de metros cuadrados que le corresponde a cada una las personas que habitan la vivienda.

15. ¿En qué provincia están las 5 viviendas en las que tienen menos espacio para vivir por persona?

16. Crea una variable de tipo factor que recoja el número de dormitorios de cada vivienda y llámala *DormitoriosFactor*. En dicha variable, deberán aparecer agrupados en una misma categoría aquellas viviendas que tiene 5 o más dormitorios.

17. Representa gráficamente la variable *DormitoriosFactor*, utilizando un diagrama de barras y un diagrama de sectores.
    - ¿En cuál de los dos gráficos distingues mejor las diferencias entre las frecuencias de las distintas categorías?
    - ¿En cuál de los dos gráficos puedes determinas fácilmente el valor de los cuartiles?

18. ¿Cuántos dormitorios tiene, al menos, el 90 % de la población (que más dormitorios tiene)?

19. ¿Cuántos dormitorios tiene, como mucho, el 90 % de la población (que menos dormitorios tiene)?

20. Filtra el conjunto de datos y guarda únicamente aquellos datos de viviendas cuyo número de aseos esté entre 1 y 4 (llama *filtradoAseos* al nuevo conjunto de datos). Después:
    - Calcula una nueva variable llamada *PersonasPorAseo*, definida como *TotalPersonas/SuperficieVivienda*.
    - Representa gráficamente la variable.
    - ¿En qué vivienda se comparten más los aseos? ¿Y menos?

21. Crea, mediante filtrado,dos conjuntos de datos llamados *Madrid* y *Zaragoza*, cada uno con las viviendas ubicadas en las respectivas provincias. Después, compara la variabilidad de la superficie de las viviendas en ambas provincias.