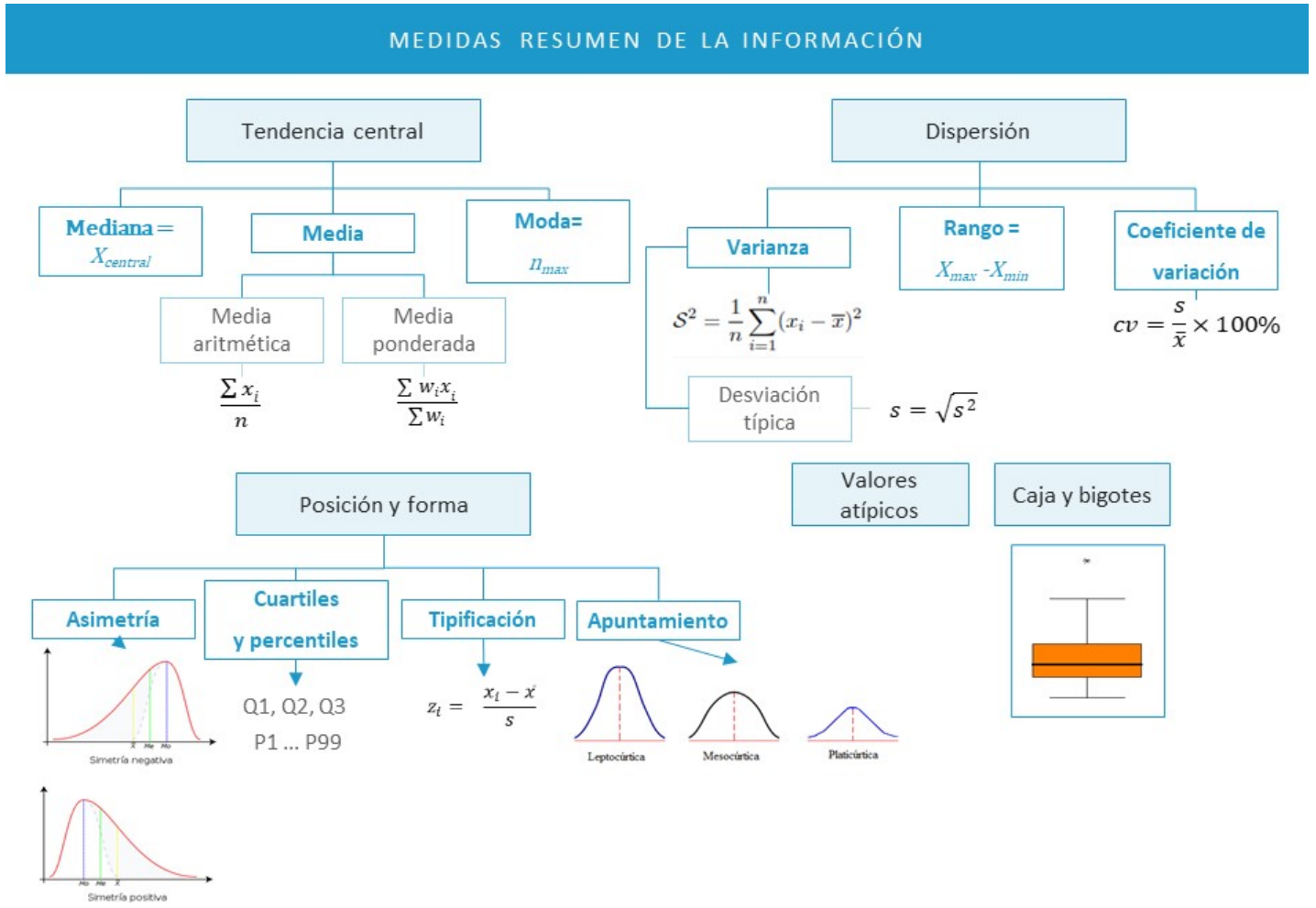
Análisis e Interpretación de Datos

Tema 3. Medidas que resumen la información

Índice

Esquema



3.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **páginas 39-68** del siguiente libro:

Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones.

Versión electrónica: https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/

Este tema versa sobre cómo resumir la información de una distribución estadística en números. Estos números son medidas que resumen la información y características de la muestra y/o la población. Para hacerse una idea global es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado el tema.

También será clave que practiques con las actividades propuestas en el tema, los cuales están diseñados para que apuntales las ideas más importantes sobre medidas resumen y algunos conceptos asociados también muy importantes, como la gráfica de caja y bigotes que es fundamental para hacerse una idea resumida de la distribución así como de la posible presencia de valores atípicos.

3.2. Medidas de tendencia central

Las primeras medidas que vamos a estudiar son las que giran alrededor de la idea de centro de la distribución de los datos. Son por tanto valores que se encuentran en el medio o la mitad de un conjunto o distribución de datos. Estas medidas o **estadísticos** también persiguen identificar valores que sean algo así como representantes de todos los datos.

La primera medida de tendencia central y más sencilla ya es conocida por nosotros aunque sea de forma informal, la **media aritmética** es la medida que consiste en la suma de todos los valores divida por el número de estos valores.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Cuando el tipo de distribución de datos contenga de un dato para algún valor será necesario completar un poco la fórmula anterior para sumar tantas veces el valor repetido (x_i) como frecuencia presente este (n_i). De modo que nos queda esta nueva versión así:

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k}$$

En algunas ocasiones —con mucha menos frecuencia que la media aritmética— surge una media que no está basada en una concepción frecuentista, sino que es ponderada estando cada valor multiplicado por un peso. Esta es la **media ponderada**:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

donde los w_i son los pesos o ponderaciones de cada x_i.

En contadas ocasiones se presenta un tercer tipo: la **media armónica** (que curiosamente se denomina H a pesar de que se escribe sin «h») que se emplea por ejemplo con las velocidades. Ningún valor puede ser cero para poder calcularla, pues como sabemos no existen números reales que se obtengan al dividir por cero.

$$H = \frac{n}{\sum \frac{n_i}{x_i}}$$

Ejemplo 1: Una familia se va de puente de Madrid a Barcelona a una velocidad de 100km/h, y tres días después regresa el domingo por la noche a 120 km/h. ¿Qué velocidad media ha tenido la familia en los dos trayectos?

$$H = \frac{2}{\frac{1}{100} + \frac{1}{120}} = 109{,}1\ km/h$$

Si quieres profundizar sobre las clases de medias conviene que sepas que existen aún dos más: la **geométrica** y la **cuadrática**, las cuales puedes ver en el capítulo indicado en «Cómo estudiar este tema».

L a **limitación de la media aritmética** (de ahora en adelante «media» a secas) consiste en lo mucho que le afectan las observaciones que presentan valores atípicos. Es por ello que la información que condensa no es suficiente para explicar cómo se distribuyen las observaciones.

Una medida de tendencia central que es más robusta que la media frente a valores extremos es la **mediana,** observación que ocupa el lugar central en un conjunto de datos. Al ocupar esta posición se nos presentan **dos casos**:

1. Cuando hay un número impar de observaciones la mediana ocupa justo el valor central.

2. Cuando el número de observaciones es par no existe posición central por lo que la mediana será el promedio entre las dos observaciones centrales.

En los casos en los que **tengamos grandes cantidades de datos**, la **mediana** nos proporciona información mucho fiable sobre la tendencia general de los datos que la media. Si la cantidad (que no el porcentaje) **de outliers en nuestra muestra es alta**, es interesante separar dichos datos y realizar sobre ellos **un estudio aparte**. De esta forma, podremos constatar de forma mejor su naturaleza y **entender qué está ocurriendo**.

Ejemplo 2: Si tenemos los siguientes datos correspondientes a los puntos anotados por Gasol durante sus 13 años en la NBA:

Tabla 1: Puntos de Gasol en la NBA.

Lo primero que haríamos sería ordenarlos de menor a mayor (o de mayor a menor es indiferente).

Tabla 2: Puntos de Gasol ordenados.

Y de esta manera la mediana sería 1246 puntos, que es el valor que ocupa la posición siete que es la central al dejar tantos a la izquierda como a la derecha (6 menores y 6 mayores).

Para estudiar el caso par imaginemos que solo contamos con las temporadas completas por lo que llegamos hasta el 2012-13. De esta manera tendríamos dos puntuaciones anuales centrales.

Tabla 3: Puntos de Gasol hasta el 2011/12 (número de temporadas par).

Y ahora la mediana sería el promedio de estas dos puntuaciones centrales:

$$Me = \frac{1246 + 1381}{2} = 1313{,}5$$

Un último caso de mediana surge cuando la distribución de datos se muestra en intervalos. En este caso es preciso emplear la siguiente fórmula de interpolación de la mediana:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times a_i$$

$$rango = x_{max} - x_{min}$$

Para aplicar esta fórmula hay que manejar el concepto de **intervalo mediano** que es aquel que contiene a la mediana. El límite inferior de este intervalo es L_{i-1}, su amplitud a_i, su frecuencia n_i, y por último N_{i-1} es la frecuencia acumulada hasta el intervalo anterior. Sabiendo esto ya puedes calcular la fórmula anterior. Como curiosidad la mediana para estas distribuciones con los **datos agrupados** (el otro nombre que reciben las distribuciones en intervalos) coincide con el punto cuya área acumulada es la mitad de la historia del histograma.

Una ventaja de la mediana frente a la media aritmética reside en que se emplea para variables cualitativas también, mientras que la media no puede ser empleada para estas variables sino tan solo para las cuantitativas.

Una última medida considerada tradicionalmente como de tendencia central es la **moda,** que no es más que el valor más frecuente del conjunto de datos. Si el conjunto contiene dos datos con la misma frecuencia diremos que la distribución es **bimodal** por tener dos modas. Y por otro lado nada impide que tenga tres modas, o cuatro, o las que sean, aunque no es habitual puede suceder.

3.3. Medidas de tendencia central robustas

Tal y como hemos comentado, la media es una medida que **se ve muy afectada por los valores extremos**. Por tanto, no proporciona un valor de tendencia central fiable cuando la aplicamos en conjuntos de datos que poseen datos outliers. Para solucionar este problema, podemos, como ya hemos comentado, usar la **mediana**. Dicha medida se centra en encontrar el valor central del conjunto de datos y, por tanto, **no tiene en cuenta los valores extremos**. Esto hace que, de manera natural, sea una medida muy robusta a la que los outliers no le afectan. Sin embargo, su significado, tal y como hemos estudiado es algo diferente del de la media aritmética.

Para solucionar esto, existen varias **versiones de la media aritmética que tratan de evitar que los outliers influyan en el resultado final obtenido**. Estudiaremos, en concreto, cómo funcionan la **media recortada** y la **media winsorizada**:

Media recortada: la media recortada realiza la media aritmética a un **subconjunto central del conjunto de datos**. De esta manera, los valores outliers quedan a los extremos y no influyen en el resultado final obtenido. Por lo general, hablamos de «**media recortada al *y* %**» donde *y* indica el porcentaje de datos que debemos dejar de lado por cada extremo. Por ejemplo, si tenemos 10 datos y calculamos una media recortada al 40 %, debemos obviar 4 datos a la izquierda y 4 a la derecha calculándose la media únicamente sobre los dos valores centrales. Otros datos a tener en cuenta es que una media recortada al 0 % es equivalente a calcular una media aritmética y que cuando hablamos de medias recortadas al 25 % el cálculo se denomina «**centrimedia**».

Veamos un ejemplo. Imaginad que tenemos los siguientes datos y nos piden calcular una **media recortada al 10 %**:

El 10 % de 10 valores corresponde a 1 valor. Por tanto, **eliminamos un valor por la**

derecha y otro por la izquierda y realizamos la media aritmética sobre el siguiente subconjunto de datos:

Lo que nos da un valor de 6. Si directamente queremos recortar un número fijo de elementos en vez de trabajar con porcentajes podemos hablar de niveles. El ejemplo visto es de nivel 1 porque hemos recortado un valor a cada lado.

- **Media winsorizada**: la media winsorizada funciona de manera similar a la media recortada. La principal diferencia radica en que en la media winsorizada, en vez de eliminar los valores**, los sustituye por el menor y mayor valor que queda en el conjunto tras el proceso de eliminación**. La media winsorizada de nivel 2 del ejemplo visto consistiría en realizar una media aritmética sobre el siguente conjunto de datos:

Como podemos ver, hemos eliminado los dos valores más extremos del conjunto y los hemos sustituido por los valores extremos del conjunto de valores restante. El resultado en este caso nos da un valor de 5,9.

Tal y como puede observarse, la idea principal de estas medias consiste en **eliminar los valores extremos y realizar los cálculos sobre valores situados en torno al centro de la distribución**. De esta manera los valores outliers no afectan al resultado final y podemos calcular un valor de tendencia central centrado únicamente en los valores más típicos encontrados en el modelo de datos.

3.4. Medidas de dispersión

Las medidas de dispersión nos indican cuánto se desvían los datos, aspecto que es fundamental para conocer cómo se distribuye el conjunto de datos.

La medida de dispersión más básica es el **rango**, que no es más que la diferencia entre la observación mínima y la máxima.

$$rango = x_{max} - x_{min}$$

El rango es útil hasta cierto punto pero son necesarias otras medidas que incluyan información sobre el resto de valores y no solo del máximo y el mínimo. Bajo esta filosofía surge la **varianza** que no es más que un promedio de las desviaciones de los datos a su media. Estas desviaciones son elevadas al cuadrado para que no le afecte el sentido de estas (si es negativo o positivo).

$$Varianza = s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Cuando nos refiramos a la varianza poblacional la designaremos con ***σ2***.

Como manejando la varianza hemos elevado las desviaciones al cuadrado perdemos un tanto la referencia de magnitud respecto a los datos. Por este motivo nosotros manejamos su raíz, que es la conocida como **desviación típica** o desviación estándar.

$$Desviación\ típica = s = +\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Donde he colocado el más delante de la raíz (lo cual estrictamente en notación del lenguaje matemático no es del todo correcto) para indicar que siempre es positivo su valor, cosa que es lógica puesto que mide el tamaño de una desviación o distancia que nunca puede ser negativa.

Conviene hacer una puntualización en cuanto al denominador empleado en esta fórmula. La tradición anglosajona de la estadística acostumbra a dividir por *n-1* en lugar de entre *n*. La razón tras ello será explicada cuando veamos los estimadores de la varianza, pues veremos que al dividir entre *n-1* este estadístico muestral cumplirá ciertos requisitos que nos interesarán. Se comenta esto para evitar futuras confusiones especialmente si se recurre a fuentes estadísticas americanas por ejemplo.

Ejemplo 3: Volviendo al ejemplo de Gasol anterior, ¿cuál sería la desviación típica de sus puntuaciones anuales? Para calcularla lo primero que hacemos es dibujar la tabla de frecuencias adecuada.

Y entonces la fórmula para el cálculo es mejor realizarla a través de esta versión:

$$s^2 = \frac{\sum x_i^{\,2}}{n} - \bar{x}^2 = \frac{21872551}{13} - 1374{,}42^2 = 213132 \rightarrow s = \sqrt{213132} = 461{,}66$$

Un aspecto que no cubre la desviación típica es la comparación entre poblaciones, ya que esta refleja una magnitud que depende de la escala que tenga la población. Por ello, se maneja un estadístico llamado **coeficiente de variación**, el cual nos permite comparar la variación entre diferentes poblaciones ya que al dividir la dispersión por la media logramos que la medida carezca de unidades.

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Ejemplo 4: Tenemos una población de mujeres en la que hemos medido dos variables: Peso y estatura. La estatura media resulta de 1,68m mientras que el peso medio es de 57Kg. No podemos comparar ambas magnitudes porque se encuentran en unidades diferentes, pero si también disponemos de las desviaciones típicas sestatura=7,5cm y speso=12kg entonces podemos calcular los coeficientes de variación empleando la fórmula anterior:

$$CVestatura = \frac{7,5}{1,68} \times 100\% = 4,46\%;\ CVpeso = \frac{12}{57} \times 100\% = 21,05\%$$

Ahora ya podríamos comparar las variables observando que la estatura tiene una variación casi cinco veces menor que el peso.

3.5. Medidas de dispersión robustas

De forma análoga a como calculábamos la media winsorizada en el apartado 2.3, es posible calcular la **varianza winsorizada** siguiendo la siguiente fórmula:

$$s_W^2 = \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{x}_\alpha^W)^2$$

Donde W_i hace referencia al conjunto de datos winsorizado según el mismo proceso visto en el apartado 2.3 con la media y

$$\overline{x}_\alpha^W$$

hace referencia a la media winsorizada.

Siguiendo el mismo patrón que para definir la cuasivarianza, podemos definir la **cuasivarianza winsorizada** de la siguiente manera:

$$S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{x}_\alpha^W)^2$$

$$z_i = \frac{x_i - \bar{x}}{s}$$

Donde la única diferencia con la fórmula anterior es que **se divide entre *n-1* en vez de entre *n***. Es importante tener en cuenta que es posible, tal y como sucedía con la desviación típica y la varianza, que puede calcularse la **cuasidesviación típica winsorizada** como la raíz de la cuasivarianza winsorizada.

3.6. Medidas de posición y forma

Las medidas de posición son necesarias para poder saber si un valor está alejado o no de su media, lo cual nos da idea de lo extremo que es comparado con la «mayoría» de los datos de su conjunto. Para esto una herramienta útil es la tipificación a través de la siguiente fórmula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Los valores zi son llamados **puntuaciones tipificadas** y nos indican el número de desviaciones típicas que se aleja un valor de su media. Además, otra propiedad interesante que tienen es que nos permiten comparar diferentes variables que pueden ser de la misma o de diferentes poblaciones también, pues al tipificar un valor desaparecen sus unidades. Un elemento común del coeficiente de variación y de las puntuaciones típicas es precisamente esta propiedad. Sin embargo no conviene confundirlos, pues mientras que el coeficiente de variación lo es de toda la muestra o población la puntuación tipificada se calcula para cada puntuación (Rius et al., 1998).

Ejemplo 5: Regresando al ejemplo de Gasol, si quisiéramos saber cuándo fue comparativamente mejor su puntuación, si en 2002/2003 o en 2012/2013, tendríamos que disponer de las medias de puntuación de la NBA para ambas temporadas y sus desviaciones típicas. Si la media de puntos para los pívots fue de 707 puntos con una desviación típica de 451 puntos para el 2002/03 y de 729 puntos con una desviación típica de 411 puntos para el 2012/13. ¿Cuándo obtuvo Gasol mejores resultados en comparación con el resto de jugadores de la NBA en su misma posición? Para saberlo calcularíamos las puntuaciones tipificadas de Gasol para estas dos temporadas.

$$z_{2002/03} = \frac{1555-707}{451} = 1{,}88;\ z_{2012/13} = \frac{673-729}{411} = -0{,}14$$

Y de esta manera sabemos que fue en el 2002/03 cuando fue comparativamente mejor, alejándose casi dos desviaciones típicas frentes a sus compañero pívot. En la temporada 2012/13 fue sin embargo algo peor que la media de sus compañeros de esa misma temporada.

Ideas clave

A partir de las puntuaciones típicas podemos saber precisamente si una puntuación es frecuente o no lo es dentro de su población. Para ello, basta con contemplar si z está comprendida entre -2 y 2. Cuando salen fuera de este rango se considera infrecuente. Esta afirmación se basa en la distribución normal, la cual veremos más adelante cuando abordemos las principales distribuciones de probabilidad.

Otra medida fundamental para tratar la posición de los datos es el **cuartil** que, como su nombre indica, proviene de dividir en cuartos iguales el conjunto de datos. De este modo tenemos tres cuartiles: Q1, Q2 y Q3. Así el primer cuartil es el valor que hace que el 25% de los datos sean inferiores a él, el cuartil segundo deja el 50% de las observaciones y por tanto coincide con la mediana. El cuartil tercero por último será aquel que deja el 75% de las observaciones a su izquierda.

En realidad los cuartiles son a su vez casos específicos de **percentiles** y estos a su vez de **cuantiles** que son el nombre global que reciben estas medidas de posición. Los percentiles son aquellos valores que dejan un tanto por ciento de los datos a su izquierda. Así diremos que por ejemplo el percentil 40 es aquel que deja el 40% de los datos a su izquierda y el resto a su derecha (un 60%). Todavía hay un tipo de cuartil más que son los **deciles** que como su nombre indica habría nueve (el décimo no cuenta de modo análogo a los cuartiles que son tres).

Según esto los cuartiles 1º, 2º y 3º equivaldrán a los percentiles 25, 50 y 75 respectivamente. El cuartil segundo a su vez también equivaldría al decil quinto y a la mediana. Matemáticamente para que te familiarices con la notación quedaría:

$$P_{50} = D_5 = Q_2 = Me$$

Ejemplo 6: De nuevo nos basaremos en los datos de las puntuaciones de Gasol, ya que las hemos empleado anteriormente para ilustrar la mediana, que de hecho coincide con el cuartil segundo, como ya hemos comentado. Se ha generado esta tabla traspuesta (no es la habitual en columnas) con las frecuencias acumuladas, que son las que marcan las posiciones donde estarán los percentiles. Los primeros cálculos por tanto son los que identifican la posición de cada percentil del siguiente modo:

Para el P_{25} = Q1 tenemos que el valor que ocupa la posición (N+1)/4=14/4= 3,5 y por tanto sería el promedio entre 1129 y 997, que es 1063.

Para el P_{50} = Q2= Me tenemos que el valor que ocupa la posición (N+1)/2=14/2= 7, luego es el que dato que ocupa el lugar 7 que es el 1246.

Para el P_{75} = Q3 tenemos que el valor que ocupa la posición 3(N+1)/4=42/4= 10,5; por lo que resulta el promedio o semisuma entre 1528 y 1541 que resulta 1534,5.

Tabla 5: «Los cuartiles de Gasol».

Una confusión habitual con los cuantiles es confundir la posición con el cuantil, algo para la que hay que estar atentos cuando los calculemos pues de otro modo acabamos dando por bueno un resultado que no lo es.

La fórmula vista anteriormente para el cálculo de la mediana con datos agrupados es extensible al resto de cuartiles de este modo:

$$Q_k = L_{i-1} + \frac{\frac{kN}{4} - N_{i-1}}{n_i} \times a_i$$

Ten en cuenta que estamos designando el inicio de un intervalo como L_{i-1} pero hay quien lo considera L_i. Esto te debe dar igual, pues escribiéndolo de un modo u otro siempre significa el inicio del intervalo mediano o del intervalo que contenga al cuartil correspondiente.

Pasaremos ahora al estudio de la forma de la distribución de datos, que engloba la simetría y el apuntamiento. En cuanto a la primera la distribución puede ser simétrica o bien presentar **asimetría positiva o negativa** (también designada simetría positiva o negativa), dependiendo hacia el lado que tenga la cola.

Gráfica 1: Tipos de asimetría y su relación con las medidas de tendencia central.

También podemos apreciar la relación entre la simetría o asimetría y la posición de la media, mediana y moda. En el caso de la simetría perfecta, que no aparece en la imagen coinciden media, mediana y moda. Para medir la simetría los programas estadísticos en ocasiones también confeccionan un coeficiente de simetría (*As* de Pearson) que resulta positivo cuando la asimetría es positiva resultando que la media

sea mayor que la moda, nulo cuando es simétrica (Moda=Media) y negativo cuando la asimetría es negativa resultando que la moda sea mayor que media.

$$As = \frac{\bar{x} - Mo}{s}$$

En cuanto al **apuntamiento**, conviene que se conozcan los tres tipos de distribuciones según su forma sea más achatada (**platicúrtica**) o si es más puntiaguda (**leptocúrtica**). Siendo **mesocúrtica** en los casos intermedios.

Gráfica 2: Tipos de distribución según su apuntamiento.

3.7 Gráficos de caja

Con objeto de resumir la información del conjunto de datos haciendo énfasis en la distribución general de estos, se desarrolló el **diagrama de caja y bigotes.** (***boxplot*** en inglés). Es probable que hayas visto alguna vez uno de ellos pero que no sepas con exactitud cómo está construido y por tanto como interpretarlo. Para confeccionarlo tenemos que contar con cinco medidas de las que ya hemos visto. Se dice que con estas **cinco medidas resumen** podemos condensar de manera rápida cualquier distribución estadística, reflejando algunas de las propiedades y facetas que no quedaban cubiertas por las gráficas tradicionales:

Mínimo, Q1, M, Q3, Máximo

A partir de los cuartiles primero y tercero se suele confeccionar otra medida que es muy útil: el **rango intercuartílico.**

Rango IQ= Q3 – Q1

Este estadístico es muy útil pues nos marca el intervalo que ocupa el 50% central del conjunto de datos.

La manera de construir el diagrama de caja y bigotes es la siguiente:

1. Con los cuartiles 1 y 3 marcamos los límites de la caja.
2. La mediana establece la línea que parte la caja en dos.
3. Los bigotes tienen como principio y final el mínimo y el máximo, salvo que haya valores atípicos, en cuyo caso estos alcanzarán el tope de 1,5 veces el Rango Intercuartílico.

Ejemplo 7: Aquí podemos ver cómo es un diagrama de caja y bigotes correspondiente al número de horas semanales de estudio que dedican los estudiantes de cierta asignatura.

Gráfica 3: Diagrama de caja y bigotes de la variable «número de horas semanales de estudio».

La mediana está situada dentro de la caja. En este caso la he puesto título y su valor para ilustrarlo pero lo normal es que no figure explícitamente sino que se muestre simplemente a través de la línea. Los límites de la caja, es decir, los cuartiles, son aproximadamente 4 (dicho «a ojo» por estar algo más cerca del 5) y 7. De modo que la mitad de los datos están contenidos entre estos dos valores.

Los bigotes se confeccionan trazando una línea o bigote que une el dato mínimo con el máximo, salvo en el caso de que haya valores que excedan 1,5 veces el rango intercuartílico. En el caso del gráfico existirían dos datos atípicos, los cuales se remarcan con circulitos. Otras maneras de marcar estos datos extremos o atípicos es empleando asteriscos.

También existen programa estadísticos que establecen diferentes categorías de valores atípicos (por ejemplo, el programa del que proviene el gráfico anterior es el llamado SPSS, el cual emplea circulitos para los datos muy extremos y asteriscos para los **atípicos «normales»**). Precisamente, sobre este programa se desarrolló lo que es hoy el código R, y por esta razón histórica hemos querido incluirlo, al menos, en un ejemplo a lo largo de este curso.

Veamos la **realización de un diagrama de cajas** similar al anterior, pero utilizando el **código R**.

Ejemplo: supongamos tenemos la lista de notas de un grupo de 10 alumnos, que introducimos en la variable NotasAlumnos. Queremos ilustrar estas notas en un diagrama de cajas, usando los recursos de funciones ya instaladas en R para la realización de este gráfico. El código y los resultados quedarían como se muestra a continuación.

RichText template tag **rawhtml** is not configured

Nótese con este ejemplo el proceder recomendado ante un *software* con el cual nos estamos iniciando en la programación.

Primero, probaremos con un caso modelo, computable «a mano» y que sirva para comprobar que las funciones que utilizamos devuelven los valores correctos (es decir, un elemento más en el proceso de validación de código).

En este caso la función que nos interesa parece ser *boxplot*. No obstante, como vemos en la figura, el gráfico presumiblemente contiene los valores numéricos correctos. Se trata ahora de **mejorarlo,** usando los avances propios de la función que permiten los etiquetados de ejes u otras funcionalidades que hagan más legible nuestra figura. Para esto, debemos seguir el tercer paso recomendado a los principiantes, **visitar la función en la propia ayuda** del programa y seguir las instrucciones para mejorar la salida y/o presentación de resultados.

En este caso, por ejemplo, bastaría con añadir etiquetados de ejes:

Ideas clave

RichText template tag **rawhtml** is not configured

3.8 Datos atípicos y análisis exploratorio de datos

Los llamados **valores atípicos o extremos** (*outliers*) son aquellos valores que distan de la mayoría de los datos. El diagrama de caja y bigotes de hecho los marca con puntitos los simplemente atípicos (ver gráfico anterior) y con estrellitas los extremos.

Para establecer la diferencia entre unos y otros asume como límite la diferencia 3 veces el Rango IQ. De este modo entre 1,5 veces el Rango IQ y 3 veces el Rango IQ los datos se marcarán como simples atípicos, y a partir de 3 veces el Rango IQ en adelante serán marcados como valores extremos.

En realidad esto es una división (entre «simples atípicos» y «extremos») que hace el paquete SPSS pero lo normal es considerarlos todos en el mismo saco que es el de datos atípicos o extremos según los llamemos.

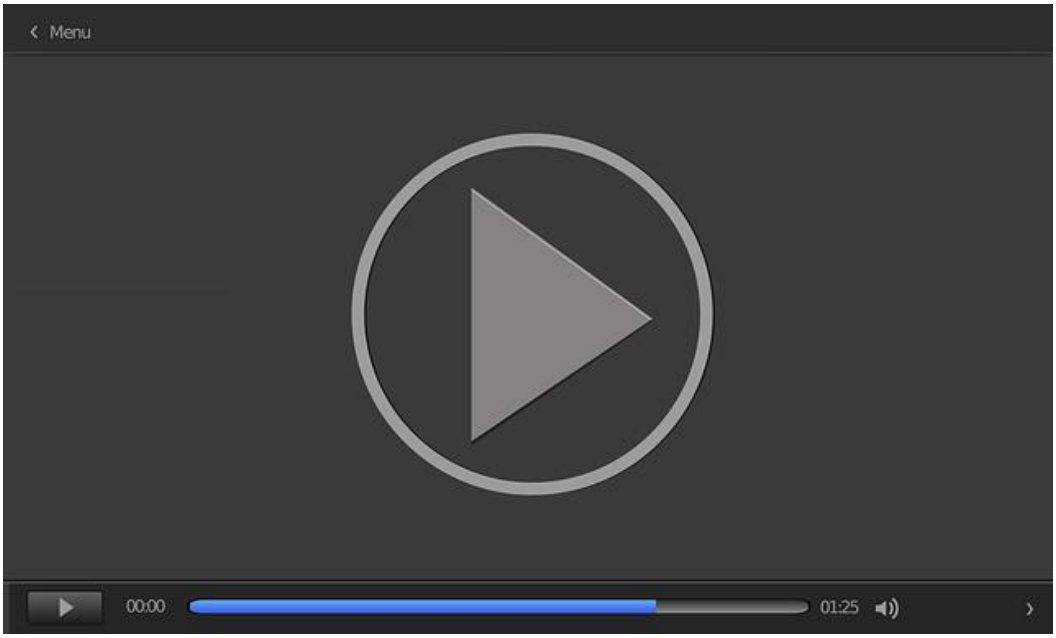
Aspectos que conviene saber de los valores atípicos:

- Afectan drásticamente a la media, pero por el contrario la mediana apenas se ve afectada por ellos.
- También se ve muy afectada la dispersión media que mide la desviación típica.
- Puede alterar la forma de la distribución de datos, especialmente en el histograma.

Por ello, cuando realizamos lo que se llama **análisis exploratorio de datos** cobra especial importancia la identificación y tratamiento de los *outliers* o valores atípicos. El análisis exploratorio de datos es el estudio de las características principales de un conjunto de datos sirviéndose de las medidas estadísticas (tendencia central, dispersión, posición, etc.) y de gráficas que ayuden a identificarlas.

Análisis exploratorio de datos con R

En este vídeo vamos a trabajar con R en el análisis de datos.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=dffa76a4-137e-4893-b2bb-acbd00b21cc9

Practica con R los conceptos estudiados

Asumiendo que tienes instalado R y RStudio, abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####################################################################
requiredPackages <- c("arsenal", "car","chemometrics","corrplot",
"gapminder","dplyr","DescTools", "foreign", "e1071", "expss", "GGally",
"ggplot2", "haven", "knitr","plotly", "psych","remotes",
"summarytools","ggridges","table1", "tableone", "tidyverse", "SmartEDA")

sesion1 <- function(pkg){
```

```
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}

sesion1(requiredPackages)
#######################################################
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club.
Factores que determinan el Default en los créditos. Modelo de riesgo
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")

Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-
default")
# Exploración inicial
#######################################################
# Sample mean
mean(Datalc$dti_n)
mean(Datalc$dti_n,trim=0.05)
winsor.mean(Datalc$dti_n, trim = 0.05)
mean(filter(Datalc, Datalc$Default == "Default")$dti_n)
mean(filter(Datalc, Datalc$Default == "Non-default")$dti_n)

# Dispersión
var(Datalc$annual_inc)
sd(Datalc$annual_inc)
var(Datalc$dti_n)
sd(Datalc$dti_n)
sd_trim(Datalc$dti_n,trim=0.05)
winsor.sd(Datalc$dti_n, trim = 0.05)

# Coefficient of variation
sd(Datalc$annual_inc)/mean(Datalc$annual_inc)
sd(Datalc$dti_n)/mean(Datalc$dti_n)

# Resumen de todas las variables
summary(Datalc$dti_n)
summary(Datalc)

# descriptivo de variables cuantitativas by default
by(select(Datalc, dti_n),factor(Datalc$Default),summary)
(by(select(Datalc, annual_inc, loan_amnt, int_rate, fico_n,
dti_n),factor(Datalc$Default),summary))

# descriptivo de variables cuantitativas by default (mean, sd)
(cuanti_summary<-Datalc %>% tab_cols(total(label = "Total"), Default) %>%
```

```
tab_cells(annual_inc, dti_n, loan_amnt, int_rate, fico_n)
%>%tab_stat_fun(Mean = w_mean, "Std. dev." = w_sd, "Valid N" = w_n, method =
list) %>%
tab_pivot%>% tab_caption("Resumen de variables cuantitativas"))
```

Prueba a ejecutar el *script* siguiendo estas indicaciones:

- Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».
- Observa la «Consola» y «Environment» cuando ejecutas cada línea.
- Comprende por qué las líneas que empiezan con # no se ejecutan.
- Encuentra la definición de cada función de R.
- Comparte aquellas funciones que no conozcas en el foro de la asignatura.
- Repasa con R todos los conceptos vistos hasta ahora.

Practica la creación de gráficos con R

Asumiendo que tienes instalado R y RStudio, abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
################################################################
requiredPackages <- c("arsenal", "car","chemometrics","corrplot",
"gapminder","dplyr","DescTools", "foreign", "e1071", "expss", "GGally",
"ggplot2", "haven", "knitr","plotly", "psych","remotes",
"summarytools","ggridges","table1", "tableone", "tidyverse", "SmartEDA")

sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}
```

```
sesion1(requiredPackages)
#########################################################
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club.
Factores que determinan el Default en los créditos. Modelo de riesgo
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")

Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-
default")
# Exploración inicial
#########################################################
#####Tablas de frecuencia de variables categóricas

# Frecuencias tablas cualitativas
table(Datalc$purpose, Datalc$Default)

#Totales filas y columnas
margin.table(table(Datalc$purpose, Datalc$Default), margin = 2)
margin.table(table(Datalc$purpose, Datalc$Default), margin = 1)
addmargins(table(Datalc$purpose, Datalc$Default))

## relativo total
prop.table(table(Datalc$purpose, Datalc$Default))

## relativo fila
prop.table(table(Datalc$purpose, Datalc$Default), margin=1)

## relativo columna
prop.table(table(Datalc$purpose, Datalc$Default), margin=2)

#Resumen de todas las variables cualitaitvas
(scu=by(select(Datalc,
emp_length,home_ownership_n,purpose),factor(Datalc$Default),summary))

(cuali_summary<-Datalc %>% tab_cols(Default) %>%
tab_cells(emp_length,home_ownership_n,purpose, total()) %>% tab_stat_rpct()
%>%
tab_pivot%>% tab_caption("Resumen de variables cualitativas"))
###############
#algunos gráficos
ggplot(Datalc, aes(x = term)) +
geom_bar(position = "dodge") + #position = "dodge", to have a side-by-side
(i.e. not stacked) barchart
theme_bw()

ggplot(Datalc, aes(x = term, fill = Default)) +
```

```
geom_bar(position = "dodge") + #position = "dodge", to have a side-by-side
(i.e. not stacked) barchart
theme_bw()

Datalc %>%
count(term = factor(term), Default = factor(Default)) %>%
mutate(pct = prop.table(n)) %>%
ggplot(aes(x = term, y = pct, fill = Default, label = scales::percent(pct))) +
geom_col(position = 'dodge') +
geom_text(position = position_dodge(width = .9), # move to center of bars
vjust = -0.5, # nudge above top of bar
size = 3) +
scale_y_continuous(labels = scales::percent)

ggplot(Datalc, aes(x= Default, group=term)) +
geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
geom_text(aes( label = scales::percent(..prop..),
y= ..prop.. ), stat= "count", vjust = -.5) +
labs(y = "Percent", fill="Default") +
facet_grid(~term) +
scale_y_continuous(labels = scales::percent)

######Medidas de localización###
# Resumen de todas las variables
summary(Datalc$dti_n)

# Sample median
median(Datalc$dti_n)

# First quartile
(Q1=quantile(Datalc$dti_n,probs = 0.25))

# Second quartile (= Median)
quantile(Datalc$dti_n,probs = 0.50)

# Third quartile
(Q3=quantile(Datalc$dti_n,probs = 0.75))

# quantiles
quantile(Datalc$dti_n,probs=c(0.25,0.5,0.75))

# Interquartile range
(IQR = Q3 - Q1)

# Lower and upper limits
Q1-1.5*IQR
Q3+1.5*IQR
```

```
# Boxplots

boxplot(Datalc$int_rate)
Q1 = quantile(Datalc$int_rate,probs = 0.25)
Q2 = quantile(Datalc$int_rate,probs = 0.50)
Q3 = quantile(Datalc$int_rate,probs = 0.75)
Q1-1.5*(Q3-Q1)
Q3+1.5*(Q3-Q1)

#Quitar Outliers

Q1 <- quantile(Datalc$int_rate, .25)
Q3 <- quantile(Datalc$int_rate, .75)
IQR <- IQR(Datalc$int_rate)
no_outliers_int_rate <- subset(Datalc, Datalc$int_rate> (Q1 - 1.5*IQR) &
Datalc$int_rate< (Q3 + 1.5*IQR))
dim(Datalc)
dim(no_outliers_int_rate)
boxplot(Datalc$int_rate)
boxplot(no_outliers_int_rate$int_rate)
###############
# Box-plot discriminando por variable categórica agregando la media
Datalc%>%
ggplot(aes(Default,int_rate, fill=Default)) +
geom_boxplot() +
stat_summary(fun.y="mean")+
theme(legend.position = "none")

#####Medidas de forma
skewness(Datalc$dti_n)
skewness(Datalc$annual_inc)
skewness(Datalc$int_rate)
kurtosis(Datalc$annual_inc, type = 1)
#https://search.r-project.org/CRAN/refmans/datawizard/html/skewness.html
####Tipificación de variables
Datalc$int_rate_z <- (Datalc$int_rate - mean(Datalc$int_rate)) /
sd(Datalc$int_rate)
df_z=Datalc %>% mutate(across(where(is.numeric), scale))
```

Prueba a ejecutar el *script* siguiendo estas indicaciones:

- Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».

- Observa la «Consola» y «Environment» cuando ejecutas cada línea.

Ideas clave

- Comprende qué hacen las funciones de R.
- Repasa con R todos los conceptos vistos hasta ahora.

Practica la exploración de datos y las técnicas de imputación con ayuda de R

Asumiendo que tienes instalado R y RStudio, abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#########################################################
requiredPackages <- c("arsenal", "car","chemometrics","corrplot",
"gapminder","dplyr","DescTools", "foreign", "e1071", "expss", "GGally",
"ggplot2", "haven", "knitr","plotly", "psych","remotes",
"summarytools","ggridges","table1", "tableone", "tidyverse", "SmartEDA",
"scales", "caret", "imputeMissings", "mice")
sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}
sesion1(requiredPackages)
#########################################################
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club.
Factores que determinan el Default en los créditos. Modelo de riesgo
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")

Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-
default")
# Exploración inicial
#########################################################
####identificación y tratamiento de outliers################
####Con boxplot identificar y eliminar atípicos############
Q1 <- quantile(Datalc$int_rate, .25)
Q3 <- quantile(Datalc$int_rate, .75)
Q2 <- quantile(Datalc$int_rate, .5)
Q1 - 1.5*IQR
Q3 + 1.5*IQR
IQR <- IQR(Datalc$int_rate)
boxplot(Datalc$int_rate)$stats
no_outliers_int_rate <- subset(Datalc, Datalc$int_rate> (Q1 - 1.5*IQR) &
Datalc$int_rate< (Q3 + 1.5*IQR))

#otra manera
boxplot(Datalc$int_rate)
```

```
out<-boxplot(Datalc$int_rate)$out
no_outliers_int_rate_1<-Datalc[-which(Datalc$int_rate %in% out),]

####identificar y truncar atípicos al P5 y P95############
Datalc$int_rate_p5_95<- squish(Datalc$int_rate, quantile(Datalc$int_rate,
c(.05, .95)))
boxplot(Datalc$int_rate,Datalc$int_rate_p5_95)

#######limitando a un num sd#########################
mean=mean(Datalc$int_rate)
std=sd(Datalc$int_rate)
Datalc$int_rate_3sd<-squish(Datalc$int_rate,c(mean-(3*std),mean+(3*std)))

#################################################################
####identificar e imputar NANs###################################
#################################################################
anyNA(Datalc)
sapply(Datalc, function(x) anyNA(x))
sapply(Datalc, function(x) sum(is.na(x)))
####
#################################################################
#Adicional, no está en el temario
###
#mtcars, iris, ..https://stat.ethz.ch/R-manual/R-
devel/library/datasets/html/00Index.html
data=airquality
anyNA(data)
sapply(data, function(x) anyNA(x))
sapply(data, function(x) sum(is.na(x)))
str(data)

#Imputar con median/mode(imputeMissings). Método rápido pero trata a cada
variable independientemente
DataIpmuted <- impute(data, method = "median/mode")
sapply(DataIpmuted , function(x) sum(is.na(x)))

#Imputar con KNN####(caret)
impknn <- preProcess(data, method = "knnImpute", k = 5)
data_knn <- predict(impknn, data)
sapply(data_knn, function(x) sum(is.na(x)))

#Imputar con Bagging (Bootstrap aggregating)
impbag <- preProcess(data, method = "bagImpute")
data_bag <- predict(impbag, data)
sapply(data_bag, function(x) sum(is.na(x)))

# Imputar con MICE (mice)...cart,rf,mean,https://cran.r-
project.org/web/packages/mice/mice.pdf
```

```
#Multivariate Imputation via Chained Equations
mice <- mice(data, method="cart")
data_mice <- complete(mice)#Creates imputed data
sapply(data_mice, function(x) sum(is.na(x)))
```

Recuerda ejecutar línea a línea el *script* anterior. Para ello, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter». Analiza lo que muestra la «Consola» y el «Environment» conforme vayas ejecutando cada línea.

3.9. Referencias bibliográficas

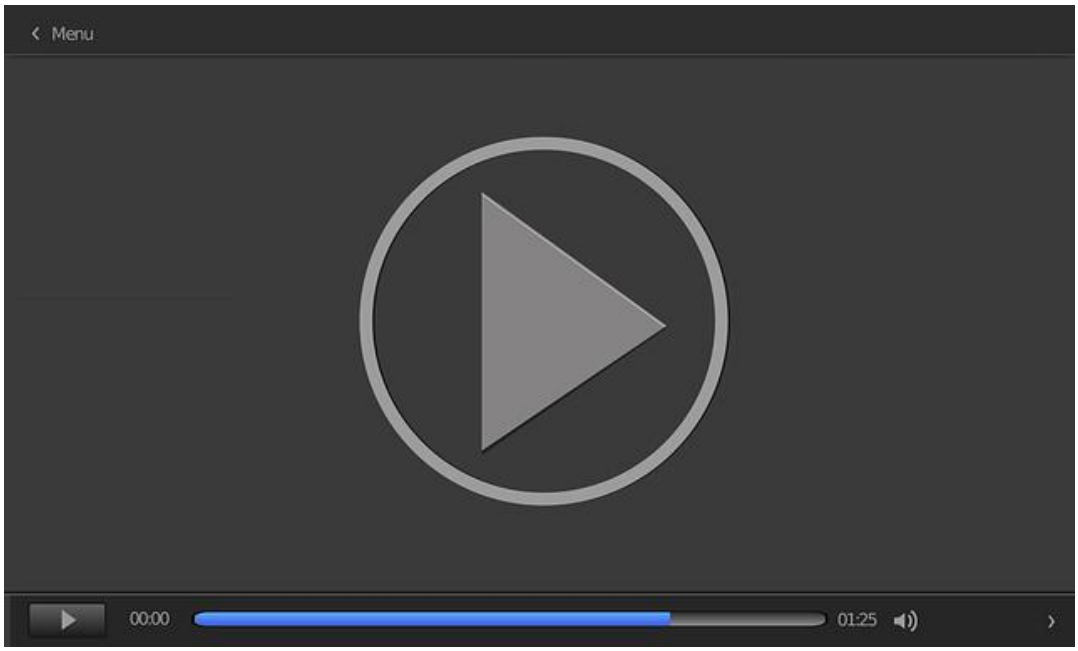
Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones.

Versión electrónica:https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/

Triola, M. F. (2009). *Estadística* (10ª ed.). México D.F.: Pearson Educación.

Medidas de Tendencia Central con Excel

En este vídeo os ilustro sobre las operaciones y pasos necesarios para estudiar las medidas de tendencia central con Excel, haciendo énfasis en la media aritmética y la desviación típica y en cómo se complementan.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=d2d7f1d6-e6a8-4a25-a59d-abdc00f2aaa1

Medidas estadísticas

Triola, M. F. (2009). *Estadística* (10ª ed., pp. 74-136). México D.F.: Pearson Educación.

Es recomendable que le eches un vistazo al libro *Estadística* de M.F. Triola, cuyo primer tema contiene prácticamente todo lo que se trata en este tema.

Estadísticas aplicadas al deporte

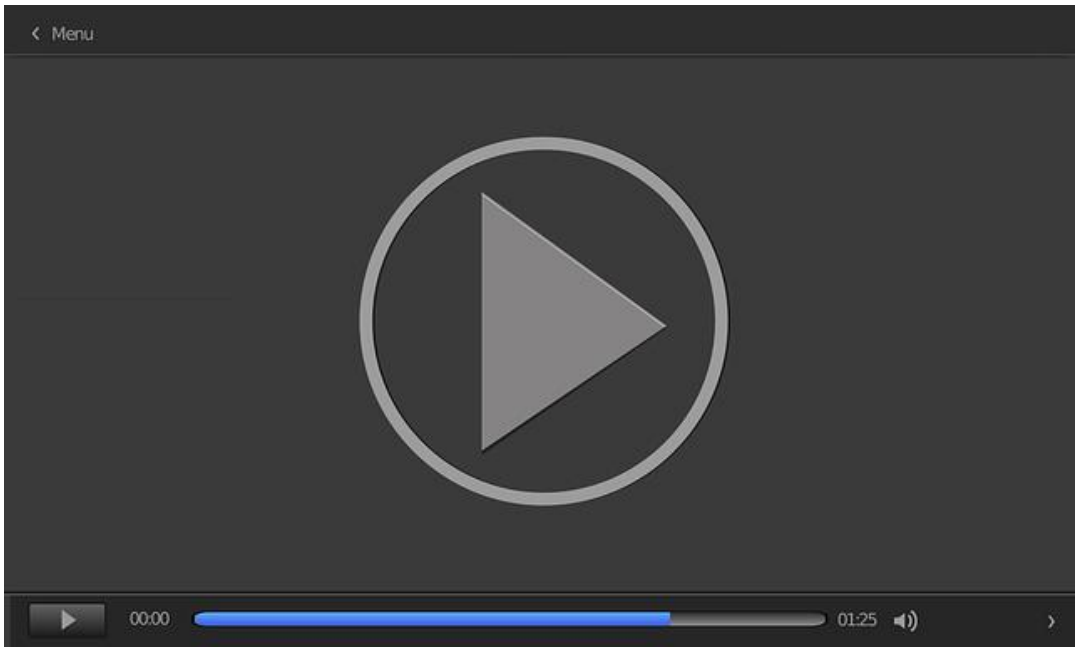
Como curiosidad, se recomienda el artículo *Estadísticas aplicadas al deporte*, de Paul Shirley publicado en El País.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web: http://deportes.elpais.com/deportes/2014/02/09/actualidad/1391968023_590198.html

---

Construir un diagrama de caja y bigotes en Excel

Interesante vídeo de Mr. Reive (en inglés, aunque podrás encontrar otros similares en español) sobre cómo construir en Excel un diagrama de caja y bigotes realizando las modificaciones oportunas, ya que por defecto no las facilita.



Accede al vídeo:

https://www.youtube.com/embed/ZFbPnwKwVWk

Estadística y probabilidad

También se recomienda que visites esta web donde están alojados mini-vídeos sobre las medidas estadísticas y otros temas de estadística y probabilidad. Se trata de un proyecto abierto realizado por la Universidad Carlos III de Madrid. También tiene la posibilidad de consulta de los materiales en formato pdf.

Accede a la página desde el aula virtual o a través de la siguiente dirección web: http://163.117.132.198/minivideos/

Bibliografía

Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones.

Versión electrónica: https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/

Triola, M. F. (2009). *Estadística* (10ª ed.). México D.F.: Pearson Educación.

Test

1. ¿Cuántos cuartiles hay en una distribución de datos?

   A. Paradójicamente hay dos, puesto que son tres pero como la mediana es el segundo se quedan en dos.

   B. 4.

   C. 3.

   D. Depende si el conjunto de datos presenta frecuencias repetidas.

2. ¿Qué cuantiles equivalen a la mediana?

   A. El quinto decil.

   B. El segundo percentil.

   C. El segundo cuartil.

   D. Las respuestas A y C son correctas.

3. La mediana…

   A. Es el valor central pero solo si el conjunto de datos es par.

   B. Es el valor central pero solo si el conjunto de datos es impar.

   C. Es el valor central siempre.

   D. Depende si el conjunto de datos presenta frecuencias repetidas.

4. La media…

   A. Se ve afectada drásticamente por los valores extremos.

   B. Es una medida con una representatividad mayor que la mediana.

   C. Es más útil que la mediana para las variables cualitativas.

   D. No es útil ni calculable para las variables cualitativas.

   E. Las respuestas A y D son correctas.

5. La medida estadística que menos se ve afectada por los valores atípicos es:

A. La desviación estadística.

B. La mediana.

C. La media aritmética.

D. La media armónica.

6. En la fórmula de la mediana para datos agrupados: ¿Qué representan las letras y símbolos?

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times a_i$$

A. L_{i-1} es el límite inferior del intervalo mediano.

B. $N/2$ corresponde con la posición que debería ocupar la mediana dentro del conjunto de datos.

C. a_i es la altura de cada intervalo.

D. L_{i-1} es el límite inferior del intervalo anterior al mediano.

E. Las repuestas A y B son correctas.

7. La varianza…:

A. Es parecida a la desviación típica.

B. Aporta la misma información sobre la dispersión que la desviación típica.

C. junto con la desviación típica y la desviación estándar conforman las medidas de dispersión más conocidas.

D. Es el cuadrado de la desviación típica.

E. Las repuestas B y D son correctas.

8. El diagrama de cajas se construye con:

A. Cuatro valores: La mediana, el cuartil 1, el cuartil 3 y la desviación típica.

B. Cuatro valores: La mediana, el cuartil 1, el cuartil 3 y la varianza.

C. Cinco valores: La mediana, el cuartil 1, el cuartil 3, el mínimo y el máximo.

D. Los cinco valores de C. más los valores atípicos sin los cuales no se puede construir.

9. Una medida estadística que nos permite comparar entre diferentes poblaciones es:

A. El coeficiente de variación.

B. La desviación estándar.

C. La puntuación tipificada.

D. Las respuestas A y C son correctas.

10. En cuanto a la asimetría…

A. Es positiva cuando la cola está a la derecha y la Moda es mayor que la media.

B. Es negativa cuando la cola está a la izquierda y la Moda es mayor que la media.

C. Es negativa cuando la cola está a la derecha y la Moda es menor que la media.

D. Es positiva cuando la cola está a la derecha y la Moda es menor que la media.

E. Las respuestas B y D son correctas.