Temas del encuentro

Tema 1 / Población y muestra. Parámetros y estimadores.

Tema 2 / Estadísticos descriptivos para el análisis exploratorio de datos.

Tema 3 / Variables aleatorias y distribuciones de probabilidad.

1 ¿Qué es la estadística?

La estadística es la ciencia que analiza los datos para obtener, a partir de ellos, inferencias. La estadística es a la ciencia como la gramática a la lengua: aporta valor a la hora de comunicar correctamente los resultados. Un análisis estadístico mal hecho no va a aportar conclusiones correctas y puede hacernos perder tiempo y dinero. Si creen que esta es una afirmación exagerada, pensemos en la cantidad de encuestas mal hechas que llevaron al fracaso de productos (¡¿qué pasó con la Fanta manzana?!) o a candidatos electorales festejando antes de tiempo.

La estadística se divide en 2 grandes ramas: **la estadística descriptiva** útil para realizar **análisis exploratorios** y **la estadística inferencial** fundamental **para hacer análisis predictivos**.

La estadística descriptiva (o deductiva) se enfoca en descubrir patrones presentes en los datos. Como parte del proceso de Data Science, **la estadística descriptiva es uno de los pilares del Análisis Exploratorio de Datos** o **EDA**. Los métodos de estadística descriptiva nos permiten:

- Determinar la tendencia central de una variable
- Determinar la variabilidad de una variable
- Determinar cómo es la distribución de una variable

Por su parte, **la estadística inferencial o inductiva nos sirve para confirmar hipótesis** (es A similar a B) **o para predecir características de una población, basados en los datos**

obtenidos de una muestra de esa población. Dentro de las técnicas aplicadas en la estadística inferencial podemos nombrar:

- Analizar si las diferencias entre dos grupos ocurren al azar o de forma sistemática indicando una diferencia real (comparación de medias).
- Comparar la varianza entre dos o más grupos de variables para saber si son similares (ANOVA).
- Analizar la correlación para saber si dos variables tienden a cambiar sistemáticamente.
- Realizar análisis de regresión para predecir un valor a partir de otro.

Algunos autores separan de este último grupo al **análisis prescriptivo**, que busca optimizar los recursos y aumentar la eficiencia operativa. Este análisis se aleja de la estadística tradicional, dado que utiliza diferentes técnicas de simulación y optimización para señalar el camino que conviene seguir.

¿Te animás a decir qué tipo de análisis se requiere para..

- calcular la media de un KPI?
- identificar la mejor ruta?
- estimar el número de productos que se venderán?

2 Tipos de variables

Independientemente si deseamos hacer un análisis descriptivo, predictivo o prescriptivo, siempre trabajaremos sobre **variables** (en inglés, *features*). La RAE nos dice que una variable es **una magnitud que puede tener un valor cualquiera de los comprendidos en un conjunto**. Cuando hablamos de variable estadística estamos hablando de una cualidad que puede adoptar forma numérica o categórica. A veces, la categoría es binaria (ej. sí/no).

Las variables que analizaremos se desprenden de nuestro problema de estudio. Identificar correctamente las variables de importancia requiere experiencia y conocimiento en el tema. Si queremos estudiar el fenómeno de las lluvias torrenciales en nuestra región, probablemente sea importante incluir dentro de nuestras variables de estudio la temperatura, la humedad relativa, la frecuencia de los vientos, etc, pero poco nos importe el precio del dólar. Desarrollar criterio científico es una clave importante en nuestro desarrollo profesional.

Identificadas las variables, podemos reconocer que las hay de distinto tipo:

- **Variable cuantitativa**: Son variables que se expresan numéricamente.
    - **Variable continua**: Toman un valor infinito de valores entre un intervalo de datos.
    - **Variable discreta**: Toman un valor finito de valores entre un intervalo de datos.
- **Variable cualitativa**: Son variables que se expresan en palabras.
    - **Variable nominal**: Expresa un nombre.
    - **Variable ordinal**: Expresa diferentes niveles y orden.

A continuación, veremos más sobre las variables cuantitativas y las variables cualitativas.

2.1 / Variables cuantitativas

Las variables cuantitativas son variables para las que tiene sentido realizar la suma, la resta o el promedio, de acuerdo al caso. Por tanto, son numéricas y pertenecen al conjunto de los números reales o a alguno de sus subconjuntos. Una variable continua puede tomar cualquier valor dentro de un intervalo, mientras que una variable discreta solo toma ciertos valores (por ejemplo, los números enteros).

Las alturas de las personas es una variable continua. Hay personas que miden 1.64 m; 1.72 m; 1.53 m. No hay valores prohibidos, solo poco probables. Lo mismo sucede con la temperatura, puede haber días de 23 ºC, otros de 32.5 ºC, incluso de -2 ºC. La temperatura es una variable continua.

El número de personas es una variable discreta. Puede haber 1, 4 o 200 personas, pero no puede haber 1.7 personas. ¿Se lo imaginan? Todas las variables que se obtienen por conteo son variables discretas: número de vasos, de sillas, de autos.

2.2 / Variables categóricas o cualitativas

Las variables categóricas o cualitativas, por su parte, como su nombre lo indica, **sirven efectivamente para categorizar elementos**. Esto significa que podríamos armar subconjuntos o subgrupos de elementos de acuerdo a dicha variable. Las variables categóricas no las podemos sumar, ni restar, ni calcular el promedio porque son palabras. Ejemplo de variables categóricas son: el sexo/género, la ocupación/profesión y el lugar de procedencia de las personas.

En general las variables categóricas son datos de texto libre, por lo que en principio son fáciles de identificar. Las variables nominales son aquellas que reciben un nombre, como el nombre de una calle, de un negocio, de un color. Por otra parte, las variables categóricas reciben una categoría: alto/mediano/bajo, buenísimo/bueno/malo/malísimo, aceptable/inaceptable. Las variables categóricas podemos ordenarlas por aquello que representan.

3 Algunos conceptos importantes

A continuación, presentaremos algunos términos que nos serán útiles a lo largo del curso. La rigurosidad con la que aquí se abordan es introductoria para simplificar su estudio. Si se desea profundizar en este contenido pueden consultarse las lecturas complementarias que se recomiendan al final de este documento.

3.1 / Población, muestra e inferencia estadística

Una **muestra estadística es un subconjunto de datos perteneciente a una población de datos**. Estadísticamente hablando, una muestra debe estar constituida por un cierto número de observaciones que representen adecuadamente el total de los datos (la población).

La población de nuestro estudio está relacionada con el alcance de nuestro trabajo. Aquí la población puede entenderse como los sujetos de estudio, la región o regiones de análisis o cualquier otro parámetro que limite nuestro trabajo, siendo la muestra una fracción de esa totalidad sobre la cual estudiaremos nuestra variable.

Normalmente, la totalidad de la población es inaccesible (es muy grande, es costoso, no podemos acceder, etc.) y, por ello, usamos la estadística para inferir qué sucede en el todo conociendo solo una parte. Llamamos **inferencia estadística** a la capacidad de predicción sobre la población usando los datos de una muestra.

La inferencia estadística funciona cuando la muestra es representativa de la población. Es decir, las variables que nos interesa estudiar se comportan de igual manera en la muestra como en la población (la distribución muestral es igual a la distribución poblacional). De esta forma, podemos suponer que todo lo que aprendimos de nuestra variable en la muestra también se aplica en el resto de la población.

¿Cuándo una muestra es representativa? Como regla general, cuanto más grande mejor! Y es muy importante que la toma de la muestra sea al azar!

3.2 / Distribución de probabilidad y distribución de frecuencia

La **distribución de probabilidad** de una variable es una función que asigna a cada valor la probabilidad de que dicho valor ocurra. La distribución de probabilidad está definida sobre el conjunto de **toda la población** y los valores pertenecen al rango de los valores posibles que toma la variable.

La **función de distribución de probabilidad** de una variable tiene una relación estrecha con la función de **distribución de frecuencia**. De hecho, una distribución de probabilidades puede comprenderse como una frecuencia teórica, ya que describe cómo se espera que varíen los resultados en la población.

De nuestros datos muestrales construiremos la función de distribución de frecuencia, y por inferencia, asociaremos a nuestra población, una función de probabilidad.

MUESTRA > Función de distribución de frecuencia > Función de probabilidad > POBLACIÓN

Una forma de pensar la distribución como concepto es la siguiente: pongamos sobre una recta todos los valores posibles de la variable, a la que llamaremos con la letra x. Ahora grafiquemos

un punto por cada valor que ocurre en nuestros datos, apilándolos cuando hay más de un valor que se repite. De esta forma, la altura de cada pila de puntos representa la cantidad de observaciones para cada valor.

3.3 / Ejemplos de distribuciones

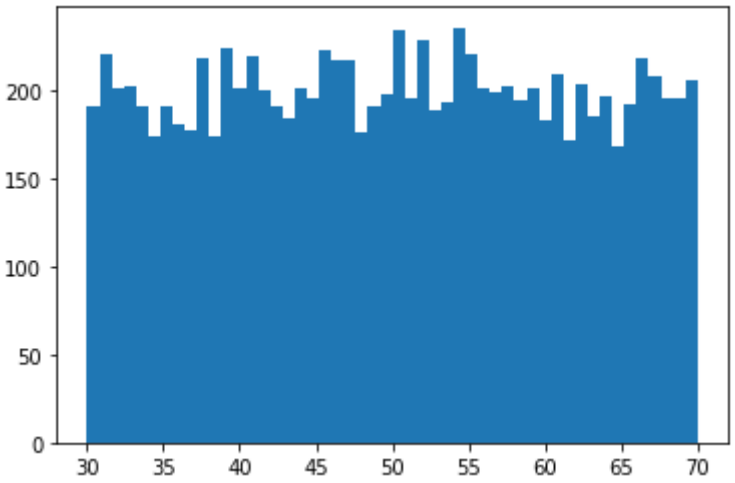
Si bien cada variable tiene una distribución diferente, en la naturaleza se observa que ciertas formas se repiten. Teniendo esto presente, se han construido funciones de distribución de probabilidad teóricas que, como se mencionó anteriormente, nos sirven para estimar cómo se comportará nuestra población.

Existen muchas distribuciones con nombre propio, sin embargo, son pocas las que se usan. Y si bien pueden tener forma parecida, **se utilizan distintas funciones de probabilidad para las variables continuas y las discretas.**

Dos distribuciones muy importantes para las variables continuas son la distribución uniforme y la distribución normal. Veremos estas distribuciones a continuación:

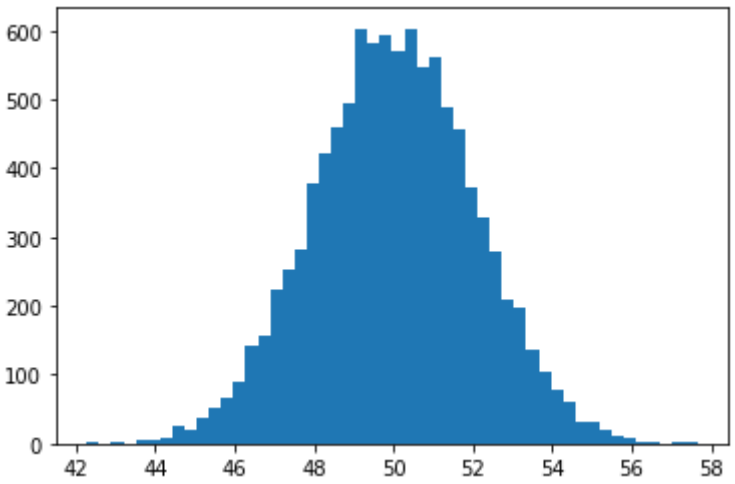
La distribución uniforme

El planteo de las distribuciones parte de una idea de una forma “perfecta” teórica, a la cual se ajustan los datos en mayor o menor grado. Si todos los valores posibles aparecen aproximadamente la misma cantidad de veces, hablaremos de una distribución uniforme.



La distribución normal

Muchos procesos y variables del mundo real siguen una distribución con una forma particular denominada distribución normal. Esta distribución está formada por puntos que se agrupan de manera simétrica en torno a un valor promedio, y cuyos valores se alejan de forma decreciente del promedio.



Estos valores y características están profundamente estudiados y desarrollados en el campo de la estadística. La importancia de la distribución normal radica en su aparición en múltiples campos del mundo real. Si logramos verificar que la distribución de los datos es

aproximadamente normal, entonces podemos echar mano de sus propiedades y ahorrar muchas suposiciones acerca del comportamiento de los datos.

4 Estadísticos descriptivos

Los estadísticos descriptivos son métodos que nos permiten conocer mejor nuestra muestra, y por tanto, la población. La estadística formal llama de distinta forma y utiliza distintos símbolos para representar los estadísticos descriptivos de una muestra y de la población. Para facilitar la lectura, no utilizaremos esa diferenciación aquí.

4.1 / Parámetros de posición

Los parámetros de posición se dividen en **parámetros de tendencia central** o **posición relativa.**

La media

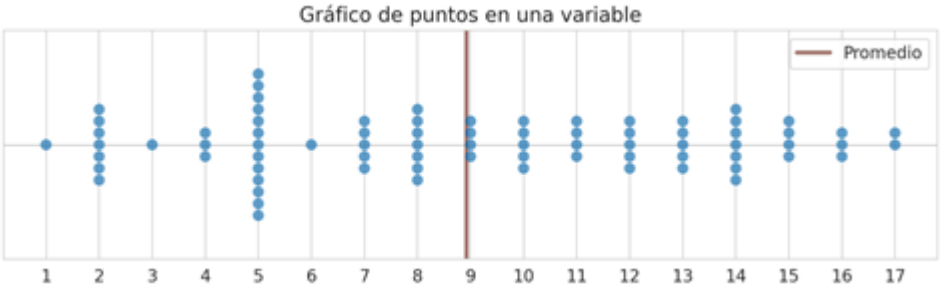
El promedio o media aritmética es la medida más conocida de tendencia central, y pretende mostrar la posición central de todos los datos. Este se define como la suma de todos los valores dividida entre la cantidad de datos.

En fórmula:

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

donde x es una variable que toma valores dentro del conjunto de números reales.

El símbolo Σ representa la sumatoria. El número *n* representa la cantidad de valores que toma la variable *x*. El número *i* representa un índice que va desde 1 hasta *n*. El promedio se representa con $\bar{x}$. El término x_i representa el valor que toma x en el lugar i. Por tanto, de corrido se lee: el promedio es igual a la sumatoria desde i = 1 hasta *n* de la los valores de x_i dividido *n*.



La mediana y los cuartiles

Muchas veces el promedio no es una medida suficiente para poder describir los datos. Puede haber valores muy alejados del "centro" de los datos, o pueden estar todos los datos muy dispersos, o bien el "centro" puede estar "corrido" hacia algún lado, como pasa muy comúnmente con los sueldos: muchos trabajadores cobran salarios mínimos o cerca del mínimo, y muy pocas personas cobran valores exorbitantes. En estos casos, el promedio no es un número claro para describir al conjunto de datos. Una medida para poder resolver esta situación es **la mediana**, que es una medida de posición relativa al conjunto de los datos.

La mediana se calcula de la siguiente manera:

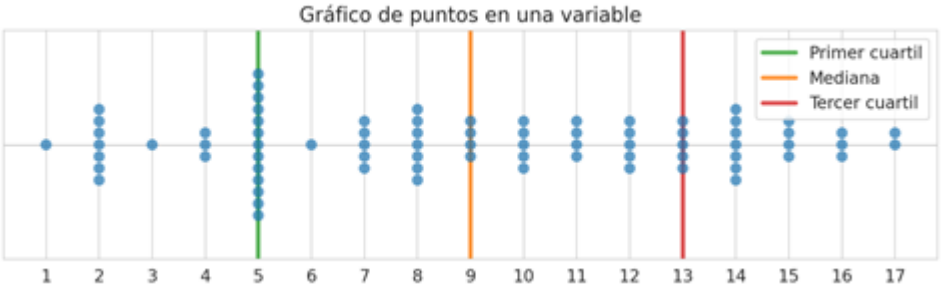
1. Ordenar los datos de menor a mayor. Si hay valores repetidos, simplemente ponerlos la cantidad de veces que aparezcan. Quedará la cantidad de datos originales, pero esta vez ordenados.
2. Con los datos ordenados, contemos ahora la cantidad de datos.
    - Si la cantidad de datos es impar, busquemos el valor que está exactamente en la mitad de los datos. Esa es la mediana.

- Si la cantidad de datos es par, habrá dos valores en el centro de los datos. Calculemos el promedio de estos dos valores. Esta será la mediana.

Pensemos por un segundo en lo que significa contar con la mediana de un conjunto de datos.

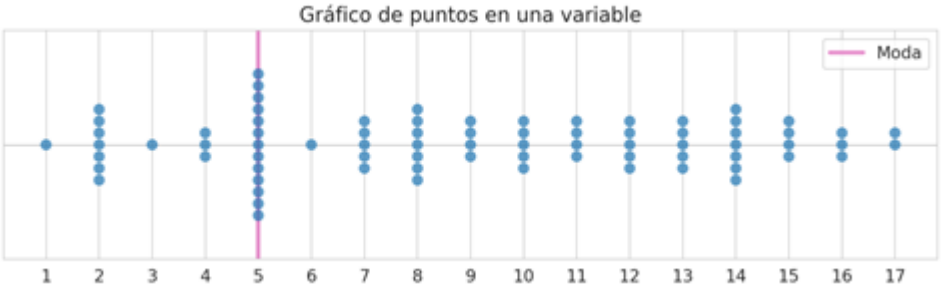
Si tuvimos que ubicarnos en la mitad de los datos ordenados para llegar a la mediana, entonces podemos decir con total seguridad que la mitad de los datos son menores o iguales a la mediana, y que la otra mitad de los datos son mayores o iguales a la mediana. De esta forma, decimos que **el 50% de los datos tiene valores menores a la mediana, y análogamente el 50% restante de los datos tiene valores mayores a la mediana**.

Con esta idea en mente, podemos extender el concepto de mediana a los valores que llegan al 25% y al 75% de los datos. A estos valores les llamaremos **primero y tercer cuartiles**, porque representan una cuarta parte y tres cuartas partes de los datos, respectivamente. En este sentido, la mediana es equivalente a las dos cuartas partes de los datos, con lo cual también la llamaremos segundo cuartil.



La moda

La moda es el valor que aparece más frecuentemente en un conjunto de datos. Se obtiene a partir de un simple conteo de los datos, calculando cuál valor aparece más veces. Es importante tener en cuenta que esta medida puede servir tanto para variables cuantitativas como cuantitativas.



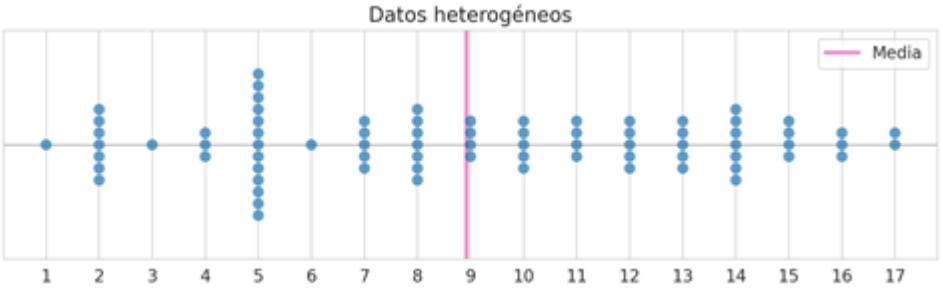
4.2 / Estadísticos descriptivos de dispersión

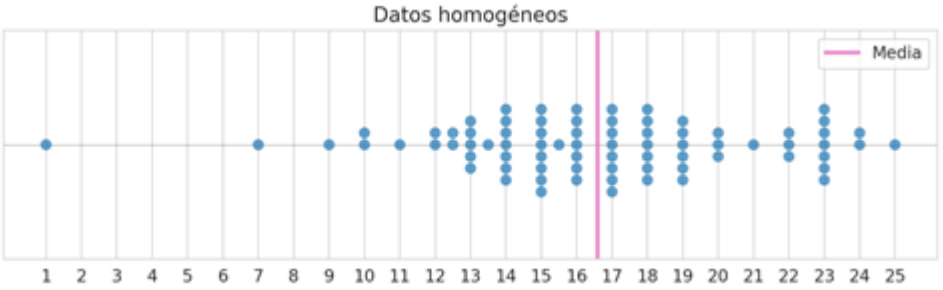
Las medidas anteriores sirven para "ubicar" los datos. Conociendo los valores de media, mediana y moda, podemos obtener un buen resumen acerca de la posición de los datos dentro del conjunto de los números reales.

Por otra parte, una vez que se conoce la ubicación de una variable, es también conveniente saber cómo cambian los valores. ¿Son todos los valores muy parecidos? ¿Cuánto cambian?

Consideraremos en este sentido a dos tipos de formas: datos homogéneos y heterogéneos. Si los datos son más homogéneos, significa que están agrupados en general más cerca de la media. En cambio, si los datos son heterogéneos, significa que en general están lejos de la media, decimos que los datos están más dispersos.

Mostramos un par de ejemplos en las dos figuras siguientes:





La varianza

Una medida estadística que funciona muy bien a este efecto es la **varianza**.

Su cálculo requiere medir las distancias de todos los datos hasta la media, elevar cada distancia al cuadrado, realizar la suma y dividir por la cantidad de datos menos dos unidades. En fórmula:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-2}$$

La varianza se representa con s^2. El número *n* representa la cantidad de valores. El número *i* representa un índice que va desde 1 hasta *n*.

Es bueno pensar este estadístico como una especie de promedio de las distancias de los datos con respecto a la media de la muestra. Es decir, cuanto más alejados estén los datos de la media, las distancias serán mayores y, por lo tanto, la varianza será más grande. Por otra parte, si los datos están muy cerca de la media, sus distancias serán menores y el valor de la varianza será más pequeño.

El desvío estándar

La varianza tiene un problema: dado que su fórmula involucra una operación de elevar al cuadrado, su resultado estará expresado en unidades al cuadrado. Por ejemplo, si estamos midiendo alturas en centímetros, la varianza estará dada en centímetros al cuadrado, lo cual no tiene mucho sentido si queremos interpretar el valor con respecto a los datos.

Para esto, simplemente se aplica la raíz cuadrada sobre la varianza y se obtiene un valor en las unidades de la variable. A este nuevo valor se le denomina **desvío estándar** y se simboliza con la letra s. En fórmula:

$$s = \sqrt{s^2}$$

Para el caso anterior de la medición de alturas en centímetros, el desvío estándar estará también expresado en centímetros, y dará una idea de cuánto están alejados los datos, en promedio, de la media de la muestra.

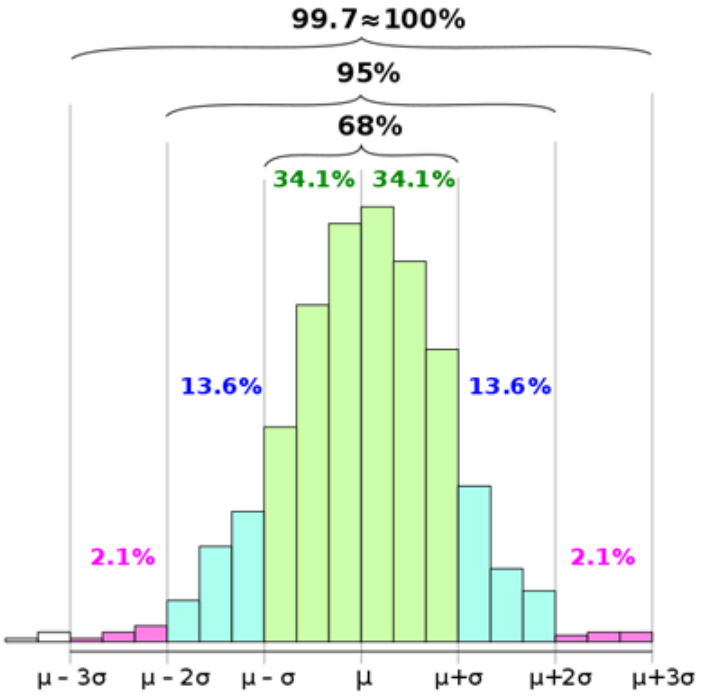
4.3 / Parámetros de forma

De acuerdo a las formas de las distribuciones podemos realizar suposiciones que nos ayudan a entender mejor nuestros datos. Una característica a tener en cuenta al analizar distribuciones es si son simétricas o asimétricas.

Como idea general sobre la distribución normal tengamos en cuenta las siguientes propiedades muy útiles que se cumplen cuando los datos presentan dicha distribución.

- Los datos normales son simétricos con respecto al promedio
- La media, mediana y moda tienen aproximadamente el mismo valor.
- Se cumple generalmente la llamada “regla empírica” a partir de la cual:
    - El 68 % de los datos está alejado a una distancia de aproximadamente 1 desvío estándar del promedio
    - El 95 % de los datos está alejado a una distancia de aproximadamente 2 desvíos estándar del promedio
    - El 99,7 % de los datos está alejado a una distancia de aproximadamente 3 desvíos estándar del promedio.

Con esto, cualquier dato que esté más allá de 3 veces el valor del desvío estándar alejado del promedio puede considerarse un valor extremo o atípico. Veremos los valores extremos más adelante.



4.4 / Variables cualitativas

En el caso de las variables cualitativas, tal como mencionamos anteriormente, los cálculos que tienen sentido son aquellos relacionados al conteo de las observaciones y su categorización. De acuerdo a lo visto, entonces, nos interesan entre otras las medidas que se enuncian a continuación.

- El conteo del total de datos (valor de n).
- El conteo de los datos por categoría, esto es, para cada valor posible de la variable, la cantidad de observaciones.
- El valor más frecuente, es decir el que tiene más observaciones. Esto es el cálculo de la moda, tal como vimos anteriormente.

5 Análisis exploratorio en R

El análisis exploratorio de datos (EDA) fue promovido por el estadístico John Tukey en su libro "Exploratory Data Analysis" (1977). El propósito general del EDA es ayudar a formular y definir las hipótesis que conduzcan al análisis inferencial e incluso a una posterior recolección de datos. Esta exploración implica una mezcla de métodos analíticos y visuales de análisis. Otros métodos estadísticos se utilizan a veces para complementar EDA, pero su principal objetivo es facilitar la comprensión antes de sumergirse en el modelado estadístico formal.

En general, las técnicas de análisis exploratorio de datos se utilizan en las primeras fases del análisis estadístico y sirven para:

1. Evaluar la calidad, completitud y consistencia de los datos
2. Investigar la distribución de las variables de interés
3. Resumir información mediante diferentes estadísticos y gráficos
4. Evaluar la necesidad de realizar transformación de las variables de interés
5. Detectar valores atípicos (outlier)
6. Explorar formas de categorizar variables (puntos de corte)

0. Cargar los datos, ordenarlos y dividirlos

Para mostrar como realizar un EDA trabajaremos con el fichero **"mydata.csv"**. Utilizaremos la función ***read.csv()*** para abrir el archivo y luego, usaremos las funciones de ordenamiento más comunes incorporadas en el paquete base en R que son ***sort()*** y ***order()***.

```
> data <- read.csv("mydata.csv")
```

La función ***sort()*** ordena en forma ascendente o descendente los valores almacenados en las estructuras de datos. Devuelve siempre un vector aunque los datos provengan de otras estructuras y opera sobre datos numéricos, caracteres, lógicos y complejos.

Podemos ordenar la variable EDAD en forma ascendente:

```
> sort(data$EDAD)

[1]  20 20 20 20 21 21 21 22 22 22 22 22 23 24 24 25 25 26 26 26
[21] 26 26 26 26 26 27 27 29 29 29 29 29 30 30 30 30 31 31 32 33
[41] 33 33 33 34 34 34 34 35 35 35
```

o en forma descendente con el parámetro **decreasing = TRUE** :

```
> sort(data$EDAD, decreasing = TRUE)

[1]  35 35 35 34 34 34 34 33 33 33 33 32 31 31 30 30 30 30 29 29
[21] 29 29 29 27 27 26 26 26 26 26 26 26 26 25 25 24 24 23 22 22
[41] 22 22 22 21 21 21 20 20 20 20
```

Nótese que por más que los datos estén contenidos en variables (columnas) de un dataframe el resultado respeta un formato vectorial. Esto impide que se pueda aplicar a todo un conjunto de datos por lo que ***sort()*** tiene muchas limitaciones.

La otra función mencionada, ***order()*,** en lugar de trabajar sobre los valores lo hace sobre los índices de los datos y esto le permite poder operar sobre todos los datos, por ejemplo, sobre un *data frame* completo.Además, ***order()*** puede anidar ordenamientos, donde primero se cumple uno y a iguales valores de este, ordena el siguiente, así sucesivamente.

Observemos cómo trabaja con el mismo dataframe:

```
> data[order(data$EDAD),c("EDAD","SEXO")]
```

Utilizaremos el campo SEXO para dividir nuestro conjunto de datos. Para esto, utilizaremos los operadores lógicos y relacionales que vimos en las clases anteriores.

```
> data_F <- data[data$SEXO == "femenino",]
> data_M <- data[data$SEXO == "masculino",]
```

1. Parámetros de posición

Veamos ahora cómo obtener los parámetros de posición. Para la media utilizamos la función ***mean()*** y para la mediana la función ***median()***

```
> mean(data_F$EDAD)
[1] 28

> mean(data_F$EDAD)
[1] 27
```

Para los cuantiles usamos la función ***quantile()***

```
> quantile(data_F$EDAD)
  0%  25%  50%  75% 100%
  20   26   27   31   34
```

Si solo nos interesa uno solo podemos especificarlo en el parámetro ***probs***:

```
> quantile(data_F$EDAD, probs = 0.50) # Q2 (mediana)
 50%
 27
```

Si queremos obtener los deciles, podemos construir un vector de 0 a 1 cada 0.1:

```
> quantile(data_F$EDAD, probs = seq(0, 1, by = 0.1)) # deciles

  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
20.0 22.8 25.8 26.0 26.0 27.0 29.4 30.0 31.4 33.0 34.0
```

2. Parámetros de dispersión

Con la función ***range()*** vemos el rango de sus valores (el valor mínimo y el máximo). Mientras que la varianza se calcula con la función ***var()*** y el desvío estándar con la función ***sd()***.

```
> range(data_F$EDAD)
[1] 20 34

> var(data_F$EDAD)
[1] 15.25

> sd(data_F$EDAD)
[1] 3.905125
```

Otras 2 funciones relacionadas son: ***min()*** que nos devuelve el valor mínimo y ***max()*** que nos devuelve el valor máximo.

```
> min(data_F$EDAD)
[1] 20

> max(data_F$EDAD)
[1] 34
```

3. Funciones de resumen

Para facilitar el trabajo, R cuenta con funciones que agrupan medidas de resumen: ***summary()*** y ***fivenum()***. Donde la primera es igual a la segunda pero con el valor de la media.

```
> summary(data_F$EDAD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     20      26      27      28      31      34

> fivenum(data_F$EDAD)
[1] 20 26 27 31 34
```

La función ***summary()*** también la podemos aplicar directamente sobre todo el dataframe:

```
>  summary(data_F)
      EDAD          PESO              SEXO        TALLa
 Min.   :20   Min.   :36.32   femenino :25   Min.   :1.180
 1st Qu.:26   1st Qu.:51.02   masculino: 0   1st Qu.:1.488
 Median :27   Median :57.40                  Median :1.650
 Mean   :28   Mean   :55.61                  Mean   :1.645
 3rd Qu.:31   3rd Qu.:60.84                  3rd Qu.:1.810
 Max.   :34   Max.   :66.21                  Max.   :2.000
                                             NA's   :1
```

Bibliografía recomendada

Introducción a la Probabilidad y a la Estadística
http://bibliotecadigital.econ.uba.ar/download/libros/Bacchini_Introduccion-a-la-probabilidad-y-a-la-estadistica-2018.pdf

Fundamentos Básicos de Estadística
http://www.dspace.uce.edu.ec/bitstream/25000/13720/3/Fundamentos%20B%C3%A1sicos%20de%20Estad%C3%ADstica-Libro.pdf