*Clase 09. Estadística descriptiva: cómo conocer a los datos*

Estadística descriptiva. Generalidades

La estadística descriptiva es una rama de la estadística que nos ayuda a entender mejor los datos, a partir de su **descripción** y **análisis**. Como parte del proceso de Data Science, la estadística descriptiva es uno de los pilares del **Análisis Exploratorio de Datos** o **EDA**, que veremos más adelante en este curso.

## Tipos de variables y tipos de datos

En las clases anteriores pudimos ver cómo funcionan los tipos de variables en Python. En estadística, **los tipos de variables son análogos en su concepto a los tipos de datos**. Los tipos de datos existen en programación, entre otras razones, porque **hay operaciones que podemos hacer con algunos de ellos pero con otros no**. Podemos partir de una noción básica en estadística sobre los tipos de variables a partir de esta analogía. En general hablaremos sobre una gran división en los tipos de variables, que es la que será más práctica para Data Science: variables cuantitativas y cualitativas.

## Variables cuantitativas

Las variables cuantitativas son variables para las que tiene sentido realizar la suma, la resta o el promedio, de acuerdo al caso. En general son numéricas y pertenecientes al conjunto de los números reales o a alguno de sus subconjuntos Por ejemplo, las alturas de las personas pueden estar medidas en centímetros, y las edades podrían estar medidas en números enteros, lo cual en realidad no sucede de esa forma (nuestra edad aumenta día a día, segundo a segundo) pero sí puede servir a fines prácticos. Tener una edad de 20,4 años puede ser muy importante para un estudio médico, pero muy impráctico para edades de consumidores de un producto.

De acuerdo a la noción que dimos anteriormente, un chequeo rápido para verificar si la variable es numérica es hacerse esta pregunta: **“Si sumo, resto o promedio los valores, ¿tiene sentido el resultado?”**. Si tiene sentido para alguno de estos casos, estamos ante una variable numérica. Es importante tener en cuenta que esto siempre **depende del problema** que estamos enfrentando. Por ejemplo, es válido decir que en promedio una familia tiene 2,5 hijos. A los fines estadísticos es muy útil saberlo. Pero por ejemplo, decir que en promedio los autos de marca Ford tienen 4,5 puertas no es útil para nada. Para estos ejemplos, “cantidad de hijos por familia” es una variable numérica, mientras que “cantidad de puertas de los autos”, si bien es un número (3 puertas, 5 puertas, etc.) no es variable cuantitativa en su naturaleza. Las variables numéricas pueden implementarse en Python con los tipos de datos int, float o complex.

### Series de tiempo

Las variables de tipo fecha constituyen un mundo aparte, tanto para la visión estadística como para el enfoque de tipos de datos de programación. Tienen algunas cualidades distintivas como la secuencialidad, la relación entre datos con valores sucesivos y las características particulares del devenir temporal, tales como años bisiestos, estaciones del año, periodicidad en escalas que no encajan de forma perfecta (como las semanas en el mes), etc. Si bien estas cuestiones pueden pensarse con simple sentido común, a la hora de realizar análisis y obtener conclusiones, especialmente con grandes volúmenes de datos, este tipo de datos puede traer inconvenientes si lo tratamos de forma manual. Afortunadamente, los lenguajes de programación tienen en cuenta estas situaciones y las resuelven con funciones específicas, y de una manera muy práctica. Recordemos que vimos una pequeña introducción a las Series de Tiempo en la Clase 04 con Pandas.

En términos estadísticos, una serie de tiempo puede ser considerada como una variable cuantitativa, ya que si bien el promedio no es una operación que tenga mucho sentido, sí se pueden establecer diferencias entre fechas, que dan como resultado duraciones. Por ejemplo, es práctica común en el desarrollo de sistemas que la edad de una persona se calcule como la diferencia entre la fecha de nacimiento (cargada en la base de datos) y la fecha de hoy (calculada con una función del lenguaje de programación).

## 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. Encontramos aquí por ejemplo algunos datos comunes que se utilizan en data science para caracterizar a las personas, tales como el sexo/género, la ocupación/profesión y el lugar de procedencia. En general las variables categóricas son datos de texto libre, por lo que en principio son fáciles de identificar. Como regla rápida, **no está mal recurrir a la pregunta anterior: “Si sumo, resto o promedio los valores, ¿tiene sentido el resultado?”**. Cuando la respuesta es “**no” para todos los casos, entonces la variable es categórica**.

## Casos particulares

En general los lenguajes de programación ofrecen herramientas para detectar los tipos de variables automáticamente, pero no está de más **verificar siempre a conciencia** los tipos de variables, en especial para detectar los casos que se muestran a continuación

### Variables que son categóricas pero que están escritas como número

Como mencionamos anteriormente, es importante **tener en cuenta la naturaleza de los datos** para verificar si tiene sentido que la variable que representan sea numérica o categórica. En el ejemplo de las puertas de los autos, tenemos una variable que puede estar escrita con números (para el ejemplo dado: variable “Cantidad de puertas”, con valores posibles 3, 4 y 5), pero cuya naturaleza no es numérica. Las variables categóricas, si son números como en este caso, no tienen sentido al sumarse, restarse o promediarse, pero **sí tienen sentido si se cuentan**. Por ejemplo, podría ser de mucho interés saber cuántos autos tienen 3 puertas frente a cuántos tienen 4 ó 5.

### Variables lógicas

Las variables lógicas son **el análogo de los tipos de datos Booleanos o lógicos** (bool) de Python. Este tipo de dato tiene como valores posibles verdadero (True) y falso (False). Esta variable tiene su tipo de dato específico en Python porque puede utilizarse para hacer operaciones particulares. No obstante, a los efectos del análisis estadístico, **sigue siendo una variable categórica**.

### Variables de alta cardinalidad y metadatos

En data science, no es raro que aparezcan variables que en cuanto a su tipo pueden ser numéricas, o bien de texto, pero que **no tienen sentido ni para sumarse, restarse o promediarse ni para contarse**. Ejemplos típicos de esta situación son los números de documento o cualquier tipo de identificación personal o fiscal, los números de serie de producto, o bien los identificadores o códigos de cliente. En general estos valores, **más que categorizar o cuantificar** una porción de todo el conocimiento que tenemos acerca de un elemento, **los identifican** de alguna manera. Los identificadores **ayudan a localizar a un elemento**, lo cual es indispensable para su almacenamiento y manipulación en bases de datos, pero **no tienen utilidad práctica a fines estadísticos**. Toda variable que cumpla con esta función es un **metadato**, es decir, un dato que describe o brinda información acerca del dato en cuestión. En este sentido, el metadato **no es cuantitativo ni categórico** y lo más conveniente es **separar este tipo de variable del análisis estadístico**.

Una forma rápida de identificar estos metadatos es a partir de su **cardinalidad**. En el campo de las Bases de Datos, la cardinalidad es una medida de **cuántos datos distintos existen para una variable**. En el caso presentado de las puertas de los autos, la cardinalidad sería 3: autos de 3 puertas, de 4 puertas y de 5 puertas, esto es, tres valores distintos. En el caso de los valores lógicos o booleanos la cardinalidad es de 2, y en el caso de los géneros puede ser variable. Por ejemplo, un usuario de Facebook puede elegir en un rango de más de 60 géneros distintos en la aplicación. Los lenguajes de programación calculan **la cardinalidad de la variable en relación a la cantidad de elementos**, y esta regla puede ser muy útil si no sabemos el significado concreto de la variable con la que estamos trabajando. **Si la cardinalidad de una variable es cercana a la cantidad de elementos, entonces es posible que la variable sea un metadato**, por ejemplo seguramente habrá tantos números de documento como personas. Por otra parte, 60 géneros parece un número grande para una cantidad de categorías cuando se lo analiza de forma aislada, pero en comparación con los 2700 millones de usuarios de Facebook, son realmente muy pocos. En sí, el género es una característica que puede inferirse a partir del simple sentido común, pero en los casos en los que el significado de la variable no sea tan claro, **la regla de la cardinalidad sirve muy bien de guía**.

# Conceptos básicos

Toda vez que tomamos un conjunto de datos, asumimos que representan una parcialidad de la realidad. Por ejemplo, al hacer una encuesta telefónica solamente tenemos los datos de los encuestados, no de todas las personas. Al hacer un relevamiento de indicadores demográficos (altura, peso, edad), solamente consideramos los datos de las personas relevadas. Aún si queremos ver los datos de nuestra empresa, y tenemos acceso a la base de datos completa, si hacemos data science queremos saber qué es lo que puede pasar con los datos que todavía no tenemos: por ejemplo, qué clientes pueden comprar más, o cuáles son más propensos a cometer fraude.

En este sentido, siempre hay datos a los que no tenemos acceso, porque no podemos relevarlos o porque todavía no sucedieron. No obstante, son precisamente estos datos los que serán objeto de nuestras predicciones y estimaciones de la mano de la estadística y del data science. Es por esto que es necesario distinguir entre estos conjuntos de datos.

En términos estadísticos, a cada elemento individual que pueda describirse con un conjunto de variables representadas con datos lo llamaremos **unidad de observación** o simplemente observación. El universo de observaciones que consta de todas las unidades de observación posibles, incluidas aquellas a las que no tenemos acceso, o que todavía no ocurrieron, se denomina **población**. Finalmente, el conjunto de observaciones que contiene las unidades de observación a las que tenemos acceso se denomina **muestra**. A partir de la muestra aplicaremos herramientas estadísticas para obtener conclusiones acerca de la población.

# Medidas de resumen

A menudo se presentan situaciones donde obtenemos muestras en forma de conjuntos de datos, y luego de acomodarlos para poder introducirlos adecuadamente en nuestro código (tema que veremos en profundidad en la clase correspondiente a Data Wrangling), nos toca echar un vistazo a las variables, para poder interpretar las características del conjunto de datos y así obtener conclusiones preliminares sobre la población a partir de la muestra. En el proceso de data science, estas conclusiones pueden servir para la toma de decisiones en cuanto a qué algoritmos aplicar, o qué consideraciones tomar con respecto a los datos.

Dos ejemplos habituales de este análisis previo pueden verse en los datos de carácter económico. En marketing es conocida la “regla de Pareto”, que dice que el 80% de la ganancia está dada por el 20% de los clientes. Si bien esto no siempre es exactamente así, es muy común que suceda, y en este sentido los análisis de clientes muchas veces exigen realizar un recorte (dejar clientes fuera del análisis) para considerar solamente los más importantes, y este recorte se realiza en base a métodos de estadística descriptiva. Por otro lado, siempre que se habla de salarios o distribuciones de la riqueza, es conocido que en general grandes sumas de dinero se concentran en pocas personas. De esta forma, calcular un simple promedio para entender cuál es el salario más representativo para la población es una mala elección, ya que sería el equivalente a medir el promedio de las alturas de una salita de niños de jardín,... ¡junto con algunos jugadores de básquet! Las diferencias en dinero son en general muy grandes (como las diferencias de altura en el caso anterior), y es necesario analizarlas con herramientas más adecuadas que el promedio. Veremos estas cuestiones a continuación.

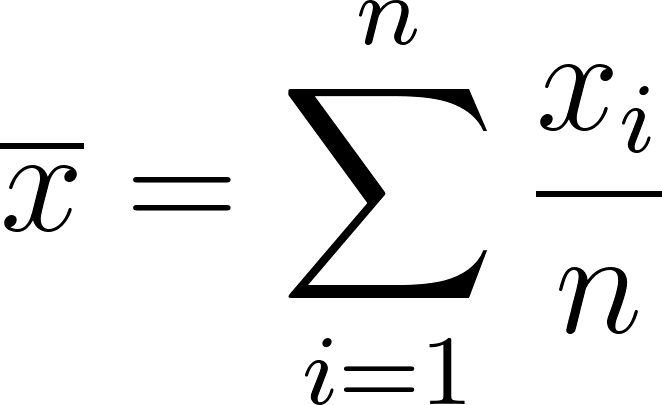


Para ordenar las medidas siguientes, vamos a hacer una distinción entre variables cualitativas y cuantitativas. Tengamos siempre presente en este sentido que la correcta caracterización de la variable es fundamental para saber qué medida aplicar, y de esta forma entender mejor a los datos.

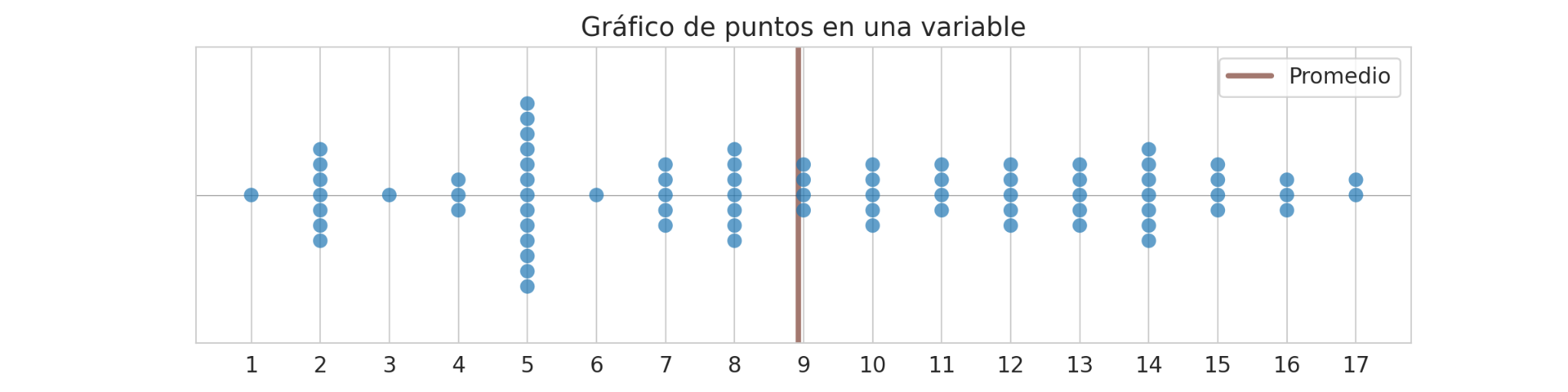
## Variables cuantitativas

### Media o promedio

El promedio o media aritmética se define como la suma de todos los valores dividida entre la cantidad de datos. En formula:

[](https://www.codecogs.com/eqnedit.php?latex=%20%5Coverline%7Bx%7D%20%3D%20%5Csum_%7Bi%3D1%7D%5En%20%5Cfrac%7Bx_i%7D%7Bn%7D%20#0)

El número n representa la cantidad de valores. El número i representa un índice que va desde 1 hasta n. El promedio se representa con [](https://www.codecogs.com/eqnedit.php?latex=%20%5Coverline%7Bx%7D%20#0). La media es la medida más conocida, y pretende mostrar la posición más central de todos los datos, describiendo su ubicación. Por ejemplo, es útil saber que el promedio de las alturas de los niños de la salita de jardín es de 100 centímetros, y que por su parte el promedio de las alturas de los jugadores de basquetbol es de 195 centímetros. Los dos grupos están bien definidos, ya que sus promedios son muy diferentes. Con esto, podemos decir que en general las posiciones de los datos para los dos grupos son distintas entre sí.



### Mediana y cuartiles

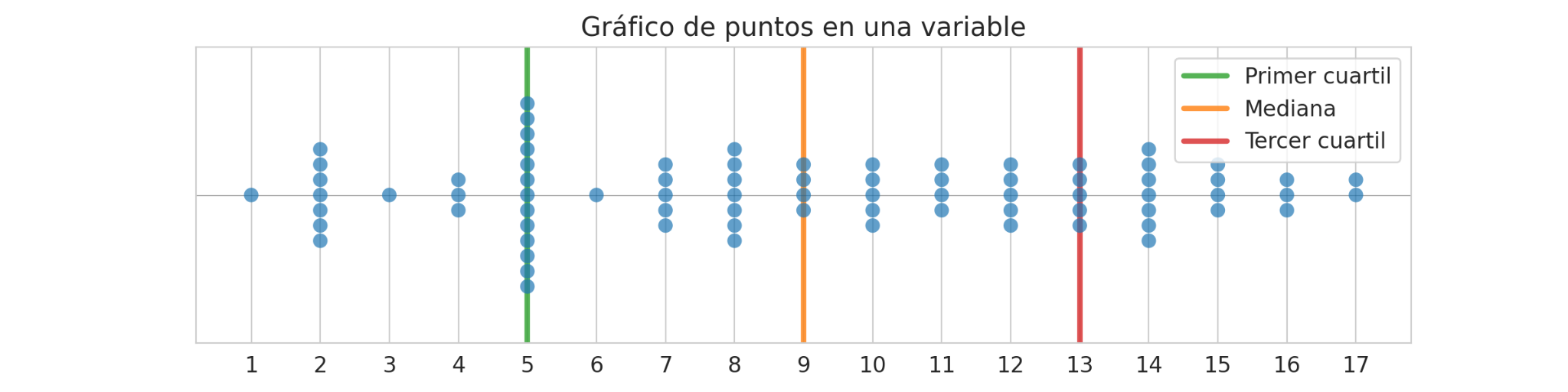
Muchas veces el promedio no es una medida suficiente para poder describir los datos. Puede haber valores muy alejados del “centro” de los datos, o pueden estar todos los datos muy dispersos, o bien el “centro” puede estar “corrido” hacia algún lado, como pasa muy comúnmente con los sueldos: muchos trabajadores 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. 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.
   1. Si la cantidad de datos es impar, busquemos el valor que está exactamente en la mitad de los datos. Esa es la mediana.
   2. 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

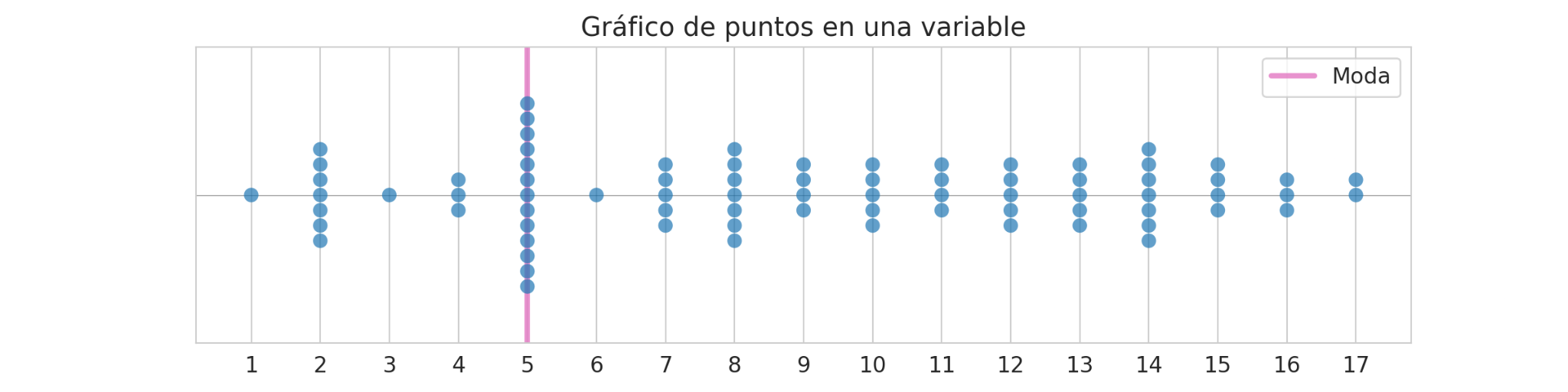
Pensemos por un segundo en lo que significa contar con la mediana de un conjunto de datos. Si tuvimos que ubicarnos en la mitad de los datos ordenados para llegar a la mediana, entonces podemos decir con total seguridad que la mitad de los datos son menores o iguales a la mediana, y que la otra mitad 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 mediana es más comprensible cuando hablamos de salarios, porque podemos decir a partir de su uso que un porcentaje determinado de la población gana más (o menos) que determinado valor. Si este valor es igual, por ejemplo, a la canasta básica de consumo, entonces el porcentaje de la población que gane menos de este valor será considerada “por debajo de la línea de pobreza”, como un ejemplo de cómo se mide este indicador económico. Los valores porcentuales derivados de la mediana que se ubican entre 0 y 100 se denominan en forma genérica percentiles. En este sentido podemos decir, por ejemplo, que el primer cuartil corresponde al percentil 25, y la mediana corresponde al percentil 50.



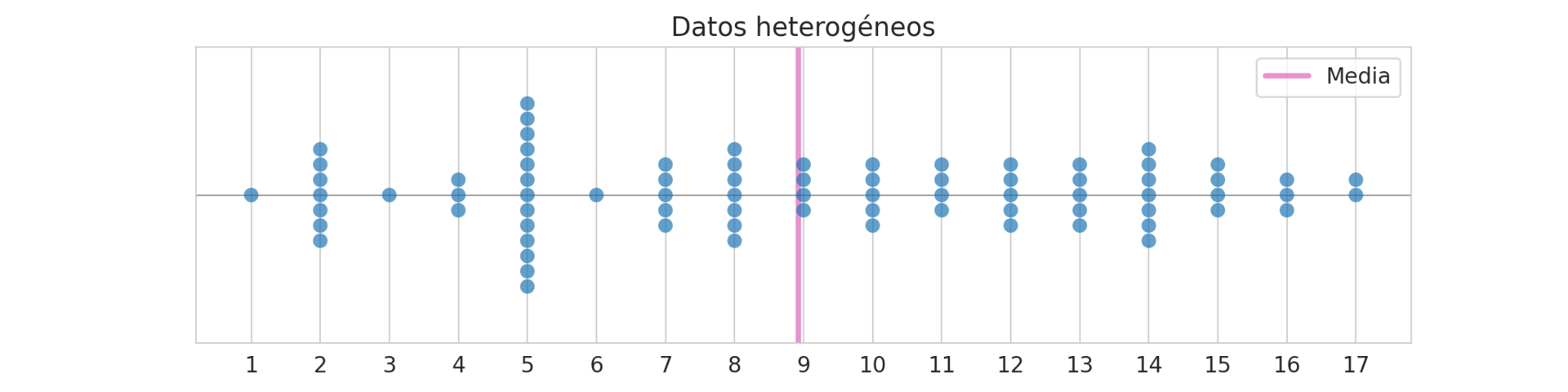
### Moda

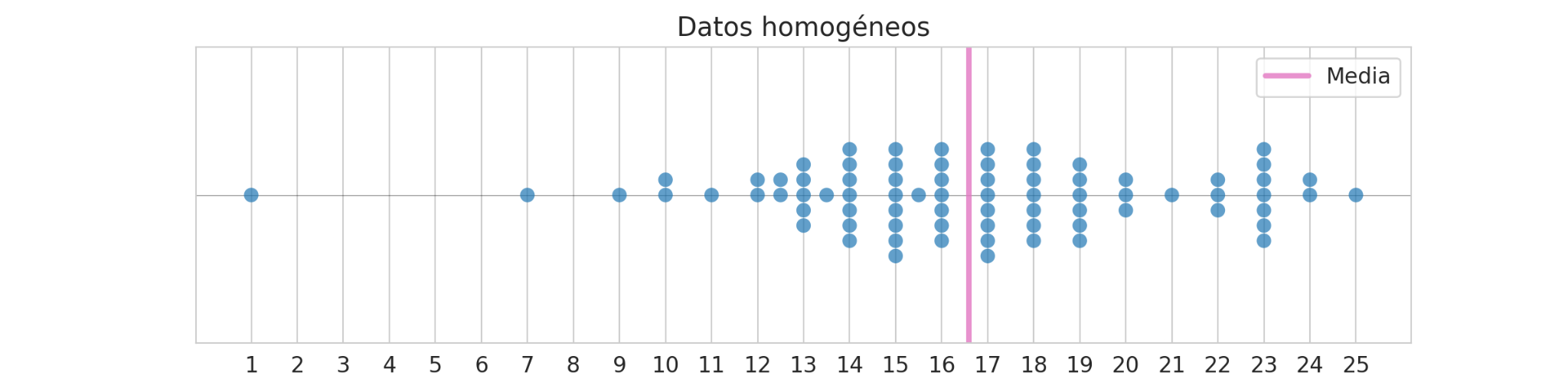
La moda es el valor que aparece más frecuentemente en un conjunto de datos. Se obtiene a partir de un simple conteo de los datos, calculando cuál valor aparece más veces. Es importante tener en cuenta que esta medida puede servir tanto para variables cuantitativas como cuantitativas. En el caso en el que se calcule la moda para variables cuantitativas, es más práctico trabajar con valores enteros, ya que en este caso habrá más valores que posiblemente se repiten, y en ese caso tiene más sentido saber cuál es el más frecuente.



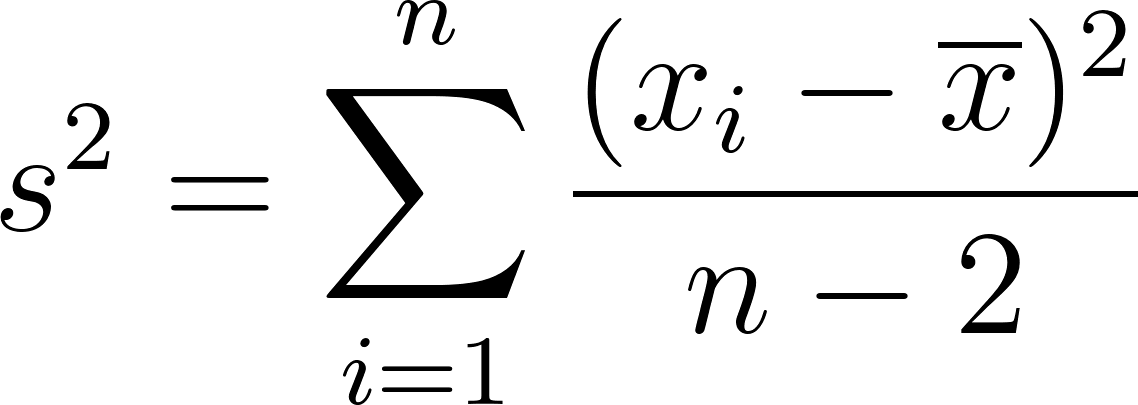
### Varianza

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. Por otra parte, una vez que se conoce la ubicación de una variable, es también conveniente saber cómo es su forma. 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. Mostramos un par de ejemplos en las dos figuras siguientes.





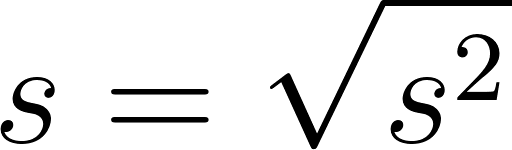
Una medida estadística que funciona muy bien a este efecto es la varianza. Consiste en medir las distancias de todos los datos hasta la media, elevar cada distancia al cuadrado, realizar la suma y dividir por la cantidad de datos menos dos unidades. En fórmula:

[](https://www.codecogs.com/eqnedit.php?latex=%20s%5E2%20%3D%20%5Csum_%7Bi%3D1%7D%5En%20%5Cfrac%7B(x_i%20-%20%5Coverline%7Bx%7D)%5E2%7D%7Bn-2%7D%20#0)

El número n representa la cantidad de valores. El número i representa un índice que va desde 1 hasta n. La varianza se representa con s2. No vamos a tener que utilizar esta fórmula “a mano” para poder entender cómo funciona la varianza, pero sí es bueno pensar este número de la siguiente manera: es 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.

### 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:

[](https://www.codecogs.com/eqnedit.php?latex=%20s%20%3D%20%5Csqrt%7Bs%5E2%7D%20#0)

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.

## 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.

# Distribuciones de variables

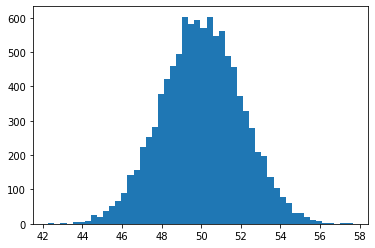
## Introducción

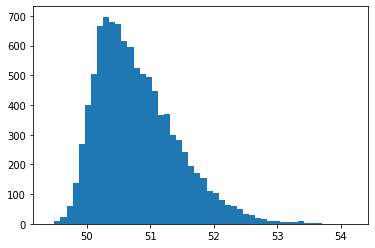
Entrando un poco más a fondo en el lenguaje estadístico, veremos que si las variables cuantitativas tienen una posición y una forma, entonces cada una de ellas puede compararse y/o diferenciarse de otras variables con distintas posiciones y/o formas. Decimos entonces que cada variable tiene una **distribución** diferente.

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.

Tal como mencionamos cuando hablábamos de la moda, no es práctico medir la frecuencia de aparición de los datos cuando los valores no son enteros. No obstante esto, es posible construir pequeños intervalos que separen los números en rangos consecutivos. De esta forma podemos considerar cada intervalo como un “valor” con entidad propia, y de esta forma podemos tener una idea más fiel acerca de la forma de los datos. De esta manera graficamos la distribución de los datos en un gráfico denominado histograma. Los histogramas son muy convenientes para visualizar rápidamente la posición y dispersión de los datos. Veremos esta herramienta gráfica en profundidad en la clase 7.

De acuerdo a las formas de las distribuciones podemos realizar suposiciones que nos ayudan a entender mejor los datos. Una característica a tener en cuenta al analizar distribuciones es si son simétricas o asimétricas. El histograma de la primera figura a continuación muestra una distribución simétrica mientras que la distribución de la segunda figura muestra una distribución asimétrica a la izquierda, como podría suceder con los salarios de una población: muchas personas con bajos salarios y pocas personas con salarios altos.





Existen muchas distribuciones con nombre propio, que tienen aplicaciones particulares en las múltiples áreas de la estadística aplicada y el Data Science. Dos distribuciones muy importantes son la distribución uniforme y la distribución normal. Veremos estas distribuciones a continuación.

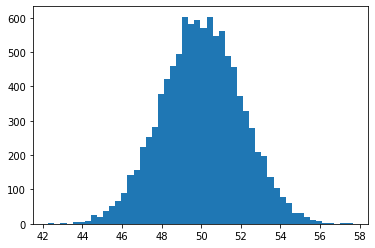
## La distribución uniforme

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. Es útil tener en cuenta que si generamos números al azar, sería conveniente que siguieran esta distribución, porque ello nos garantizaría que cada número tenga la misma probabilidad de aparición. Las funciones de generación de números aleatorios (random) que tienen las calculadoras y las planillas de cálculo siguen esta distribución. En la figura podemos observar un histograma de datos con la forma 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 cuya varianza se aleja no más de una determinada distancia del promedio. A continuación se muestra un histograma con datos agrupados de acuerdo a una distribución normal.



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.

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.

