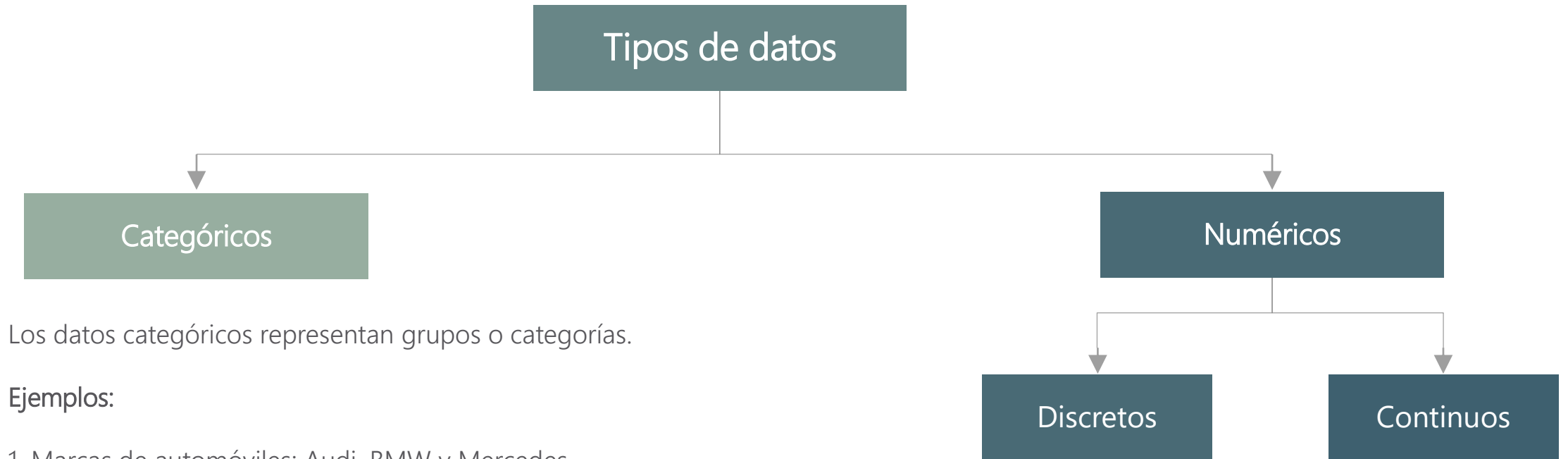


# **NOTAS DEL CURSO: ANALÍTICA DESCRIPTIVA**

# Tipos de datos



Los datos categóricos representan grupos o categorías.

## Ejemplos:

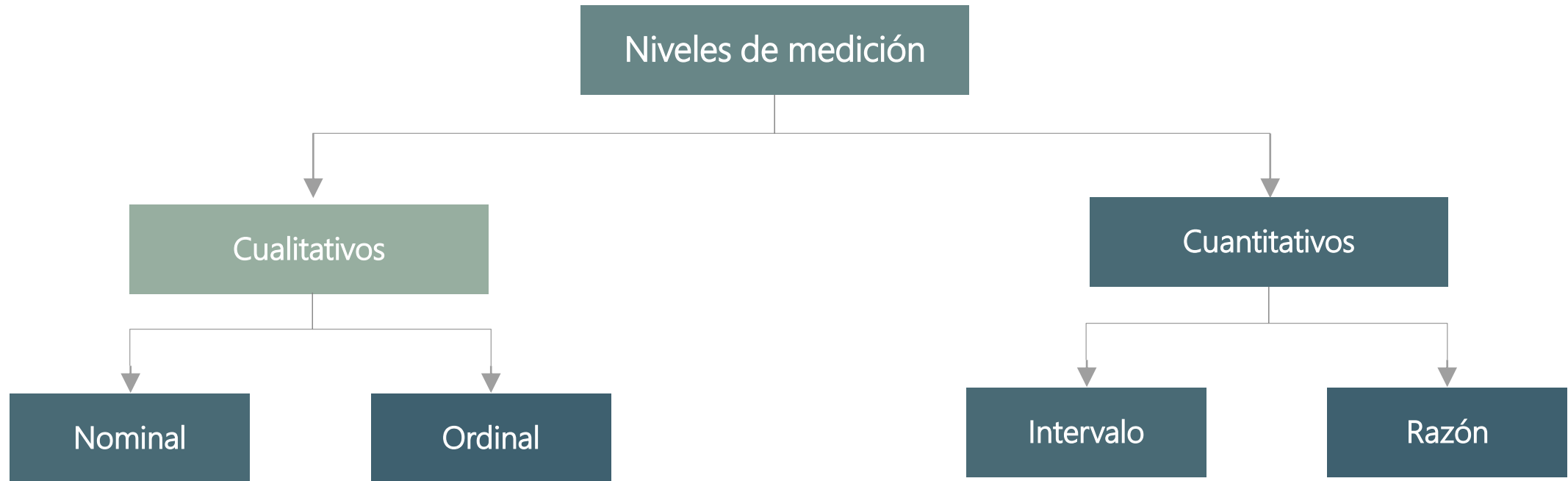
1. Marcas de automóviles: Audi, BMW y Mercedes.
2. Respuestas a preguntas de si o no: si y no.

Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.

## Ejemplos:

Discreto: # de niños que quieres tener, calificación SAT.  
Continuos: peso, altura.

# Niveles de medición



Hay dos niveles cualitativos: nominal y ordinal. El nivel nominal representa categorías que no se pueden poner en ningún orden, mientras que el nivel ordinal representa categorías que se pueden ordenar.

## Ejemplos:

Nominal: cuatro estaciones (invierno, primavera, verano, otoño)

Ordinal: calificar su comida (asquerosa, poco apetitosa, neutra, sabrosa y deliciosa).

Existen dos niveles cuantitativos: intervalo y razón. Ambos representan "números", sin embargo, las proporciones tienen un verdadero cero, mientras que los intervalos no.

## Ejemplos:

Intervalo: grados Celsius y Fahrenheit

Relación: grados Kelvin, longitud

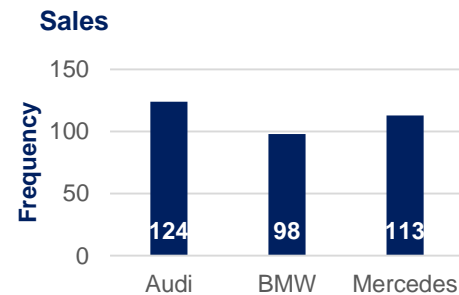
# Gráficos y tablas que representan variables categóricas

Tablas de  
distribución de  
frecuencias

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

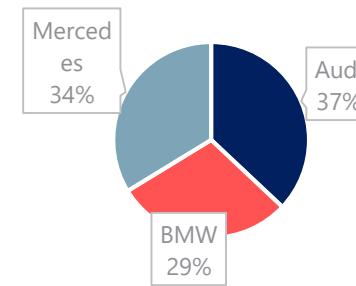
Las tablas de distribución de frecuencias muestran la categoría y su correspondiente frecuencia absoluta.

Gráficos de  
barras



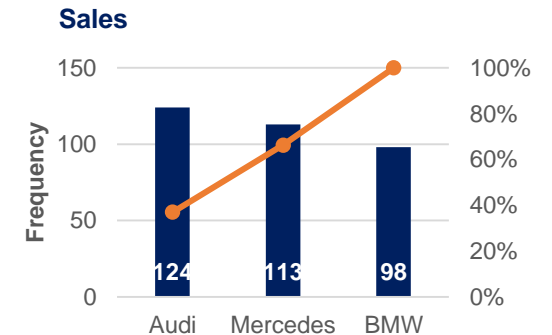
Los gráficos de barras son muy comunes. Cada barra representa una categoría. En el eje y tenemos la frecuencia absoluta.

Gráficos circulares



Los gráficos circulares se utilizan cuando queremos ver la parte que corresponde a un ítem dentro de un total. La cuota de mercado casi siempre se representa con un gráfico circular.

Diagramas de  
Pareto



El diagrama de Pareto es un tipo especial de gráfico de barras en el que las categorías se muestran en orden descendente de frecuencia, y una curva separada muestra la frecuencia acumulativa.

# Gráficos y tablas que representan variables categóricas. Fórmulas de Excel

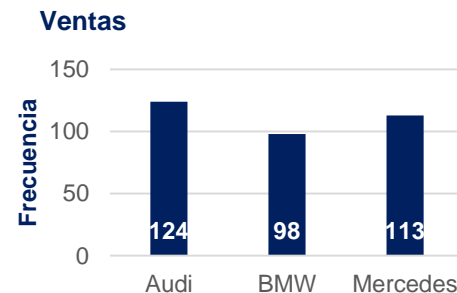
Tablas de  
distribución de  
frecuencias


	Frecuencia
Audi	124
BMW	98
Mercedes	113
Total	335

En Excel, podemos codificar las frecuencias o contarlas con una función de conteo. Esto vendrá más tarde.

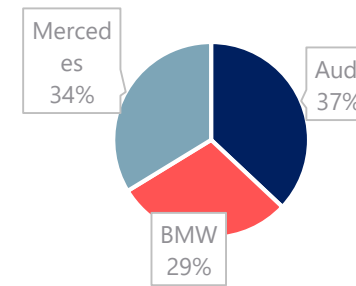
Fórmula total: =SUM() 


Gráficos de  
barras



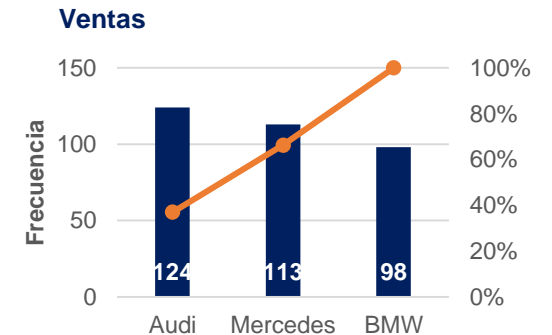
Los gráficos de barras también se denominan gráficos de columnas agrupadas en Excel. Seleccione sus datos, Insertar -> Gráfico -> Columnas agrupadas o barras. 

Gráficos circulares



Los gráficos circulares son creados de la siguiente forma: Selecciona tus datos, Insertar -> Gráfico -> Circular 

Diagramas de  
Pareto



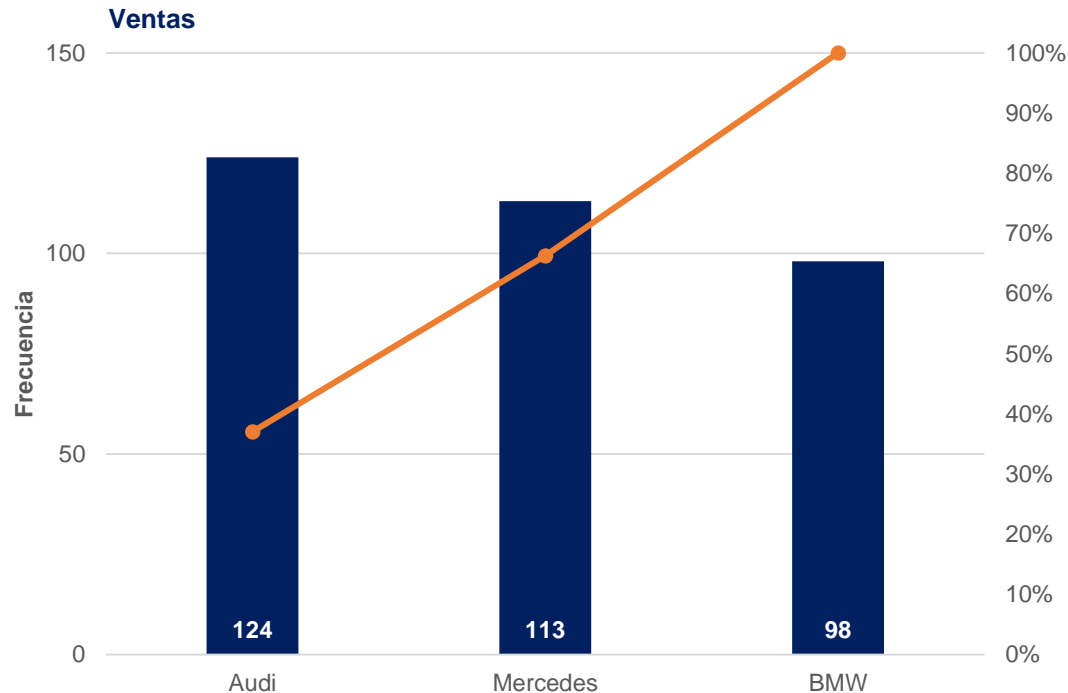
Próxima diapositiva.

# Diagramas de Pareto en Excel



## Creación de diagramas de Pareto en Excel:

1. Ordena los datos en la tabla de distribución de frecuencias en orden descendente.
2. Crea un gráfico de barras.
3. Añade una columna en tu tabla de distribución de frecuencias que mida la frecuencia acumulativa.
4. Selecciona el área del gráfico en Excel y **haz clic con el botón derecho del ratón**.
5. Selecciona **Seleccionar datos**.
6. Haz clic en **Agregar**
7. El nombre de la serie no importa. Puedes poner 'Linea'.
8. Para los **Valores de la serie**, selecciona las celdas que se refieren a la frecuencia acumulativa.
9. Haz clic en **Aceptar**. *Deberías ver dos barras una al lado de la otra.*
10. Selecciona el área del gráfico y **haz clic con el botón derecho del ratón**.
11. Selecciona **Cambiar tipo de gráfico de series**.
12. Selecciona **Combinado**.
13. Seleccione el tipo de representación de la lista desplegable. Sus categorías iniciales deben ser "Columnas agrupadas". Cambie la segunda serie, a la que llamó 'Linea', por 'Lineas'.
14. Hecho.



# Variables numéricas. Tabla de distribución de frecuencias e histograma

Inicio del intervalo	Cierre del intervalo	Frecuencia	Frecuencia relativa
1	21	2	0,10
21	41	4	0,20
41	61	3	0,15
61	81	6	0,30
81	101	5	0,25

Las tablas de distribución de frecuencias para las variables numéricas son diferentes a las de las categóricas. Por lo general, se dividen en intervalos de igual (o desigual) duración. Las tablas muestran el intervalo, la frecuencia absoluta y a veces es útil incluir también las frecuencias relativas (y acumulativas).

La amplitud del intervalo es calculada usando la siguiente fórmula:

$$\text{amplitud del intervalo} = \frac{\text{número más grande} - \text{número más pequeño}}{\text{número de intervalos deseados}}$$

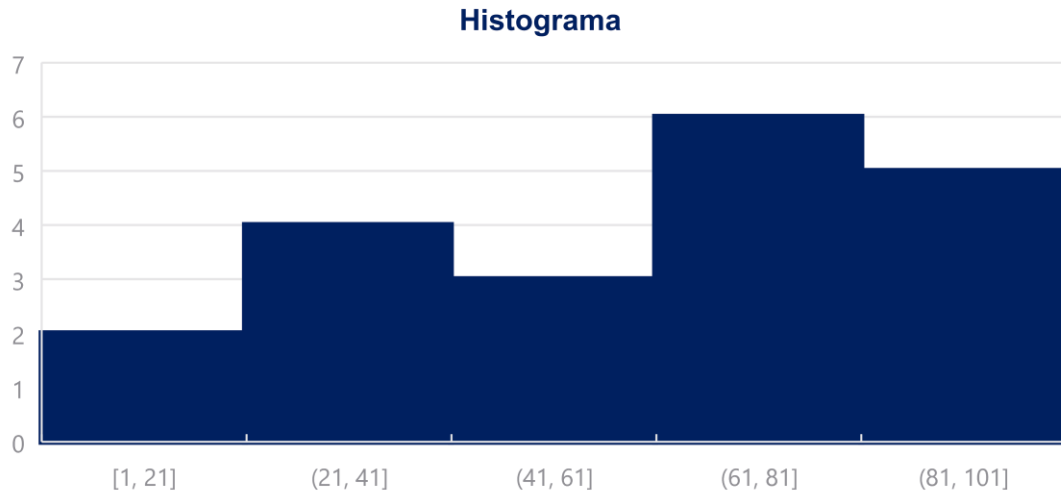


Creación de la tabla de distribución de frecuencias en Excel:

1. Decide el número de intervalos que deseas utilizar.
2. Encuentra la amplitud del intervalo (usando la fórmula anterior).
3. Comienza tu primer intervalo con el valor más bajo de tu conjunto de datos.
4. Termina el primer intervalo con el valor más bajo + la amplitud del intervalo ( = celda de inicio del intervalo + amplitud del intervalo ).
5. Comienza tu segundo intervalo donde termina el primero (esta también es una fórmula - simplemente haz que la celda de inicio del intervalo 2 = cierre del intervalo 1).
6. Continúa de este modo hasta que hayas creado el número de intervalos deseado.
7. Cuenta las frecuencias absolutas utilizando la siguiente fórmula CONTAR.SI:  
=CONTAR.SI(rango\_del\_conjunto\_de\_datos,">="&inicio del intervalo) – CONTAR.SI(rango\_del\_conjunto\_de\_datos,">"&cierre del intervalo).
8. Para calcular las frecuencias relativas, utiliza la siguiente fórmula: = frecuencia / número total de observaciones.
9. Para calcular las frecuencias acumulativas:
  - i. La primera frecuencia acumulativa es igual a la frecuencia relativa.
  - ii. Cada frecuencia acumulativa consecutiva = frecuencia acumulativa previa + la frecuencia relativa respectiva

Ten en cuenta que todas las fórmulas se pueden encontrar en los archivos de Excel de las lecciones y en las soluciones de los ejercicios que se proporcionan con cada lección.

# Variables numéricas. Tabla de distribución de frecuencias e histogramas

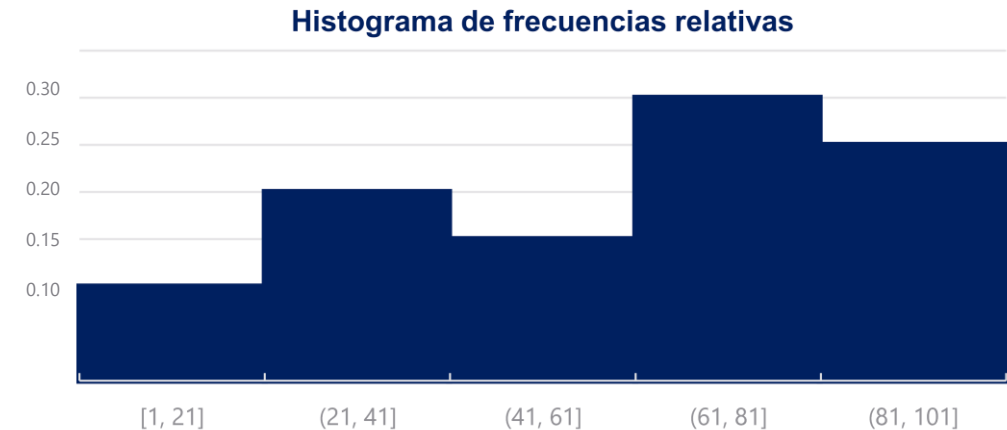


Los histogramas son una de las formas más comunes de representar datos numéricos. Cada barra tiene un ancho igual a la amplitud del intervalo. Las barras se tocan, ya que hay una continuación entre los intervalos: donde termina una -> comienza la otra.



## Creando un histograma en Excel:

1. Selecciona tus datos
2. Insertar -> Gráfico -> Histograma
3. Para cambiar el número de bins (intervalos o rangos):
  1. Selecciona el eje x.
  2. Haz clic en **Dar formato a eje -> Opciones del eje**
  3. Puedes seleccionar la amplitud del bin (ancho del rango o intervalo), número de rangos, etc.





# Gráficos y tablas para las relaciones entre variables. Tablas cruzadas

Tipo de inversión / inversor	Inversor A	Inversor B	Inversor C	Total
Acciones	96	185	39	320
Bonos	181	3	29	213
Bienes raíces	88	152	142	382
<b>Total</b>	<b>365</b>	<b>340</b>	<b>210</b>	<b>915</b>

Tipo de inversión / inversor	Inversor A	Inversor B	Inversor C	Total
Acciones	0,10	0,20	0,04	0,35
Bonos	0,20	0,00	0,03	0,23
Bienes raíces	0,10	0,17	0,16	0,42
<b>Total</b>	<b>0,40</b>	<b>0,37</b>	<b>0,23</b>	<b>1,00</b>

Una forma común de representar los datos de una tabla cruzada es utilizando un gráfico de barras agrupadas.

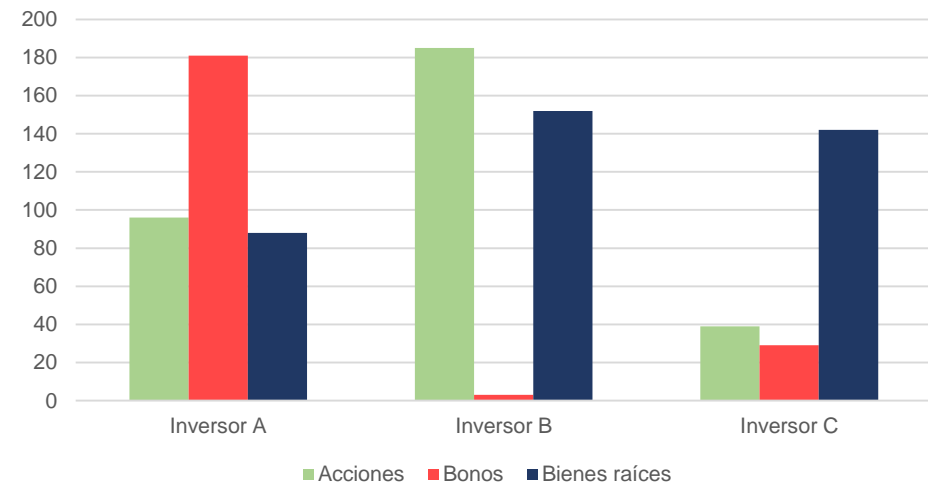
 Creando un gráfico de barras agrupadas en Excel:

1. Selecciona tus datos
2. Insertar-> Gráfico -> Barras agrupadas

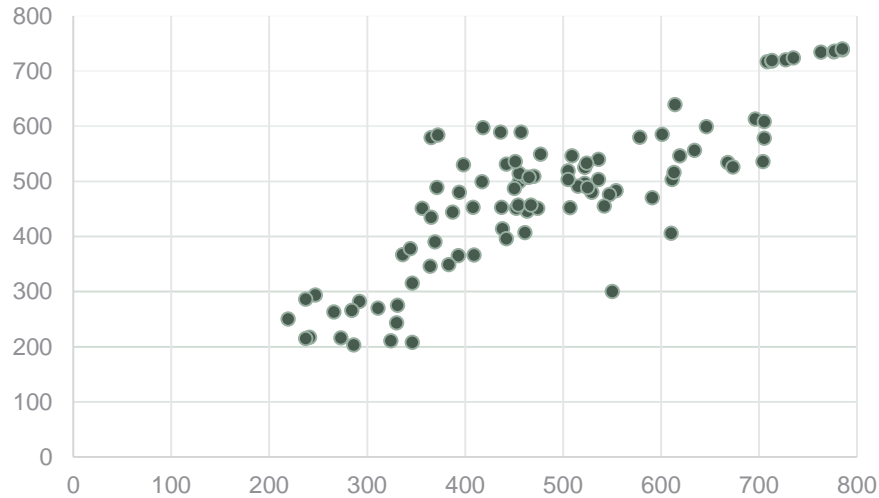
Al seleccionar más de una serie ("grupos de datos") se le pedirá automáticamente a Excel que cree un gráfico de barras (columnas) paralela.

Las tablas cruzadas (o tablas de contingencia) se utilizan para representar variables categóricas. Un conjunto de categorías etiqueta las filas y otro etiqueta las columnas. A continuación, rellenamos la tabla con los datos correspondientes. Es una buena idea calcular los totales. A veces, estas tablas se construyen con las frecuencias relativas que se muestran en la siguiente tabla.

Gráfico de barras agrupadas



# Gráficos y tablas para las relaciones entre variables. Diagramas de dispersión



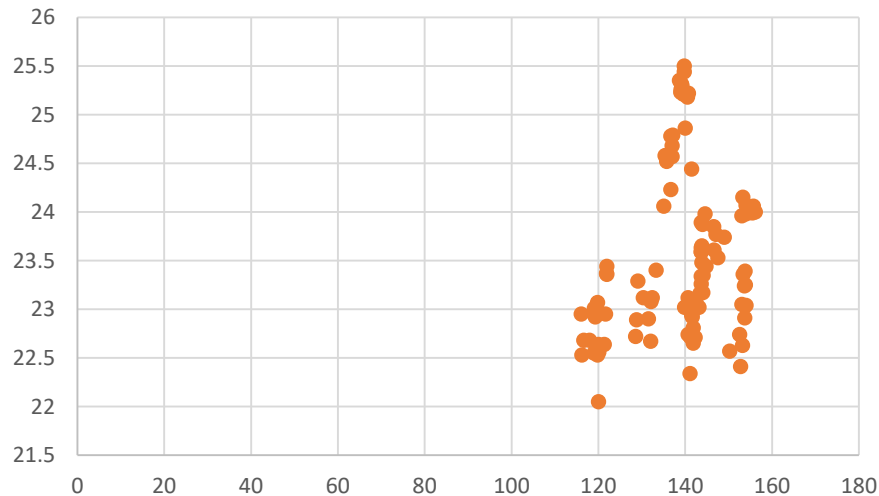
Cuando queremos representar dos variables numéricas en el mismo gráfico, normalmente usamos un gráfico de dispersión. Las gráficas de dispersión será especialmente útiles después, cuando hablemos de análisis de regresión, ya que nos ayudan a detectar patrones (linealidad, homocedasticidad).

Las gráficas de dispersión generalmente representan una gran cantidad de datos. Típicamente, no estamos interesados en observaciones individuales, sino más bien en la estructura del conjunto de datos.



Creando un diagrama de dispersion en Excel:

1. Selecciona los dos conjuntos de datos que quieres graficar.
2. Insertar -> Gráficos -> X Y (Dispersión)



Un gráfico de dispersión que se ve de la siguiente manera (hacia abajo) representa datos que **no tienen un patrón**. Las 'formas' completamente verticales no muestran asociación.

Por el contrario, la gráfica anterior muestra un patrón lineal, lo que significa que las observaciones se mueven juntas.

# Media, mediana, moda

## Media

La media es la medida de tendencia central más extendida. Es el promedio simple del conjunto de datos.

**Nota:** fácilmente afectada por los datos atípicos

La fórmula para calcular la media es:

$$\frac{\sum_{i=1}^N x_i}{N} \quad \text{o}$$

$$\frac{x_1 + x_2 + x_3 + \dots + x_{N-1} + x_N}{N}$$

 En Excel, la media se calcula con:

=PROMEDIO()

## Mediana

La mediana es el punto medio del conjunto de datos ordenado. No es tan popular como la media, pero se utiliza a menudo en la academia y en la ciencia de datos. Eso es porque no se ve afectada por los valores atípicos..

En un conjunto ordenado de datos, la mediana es el número en la posición:

$$\frac{n+1}{2}.$$

Si esta posición no es un número entero, la mediana es el promedio simple de los dos números en las posiciones más cercanas al valor calculado.


 En Excel, la mediana se calcula con:

=MEDIAN()

## Moda

La moda es el valor que ocurre con más frecuencia. Un conjunto de datos puede tener 0 modas, 1 moda o múltiples modas.

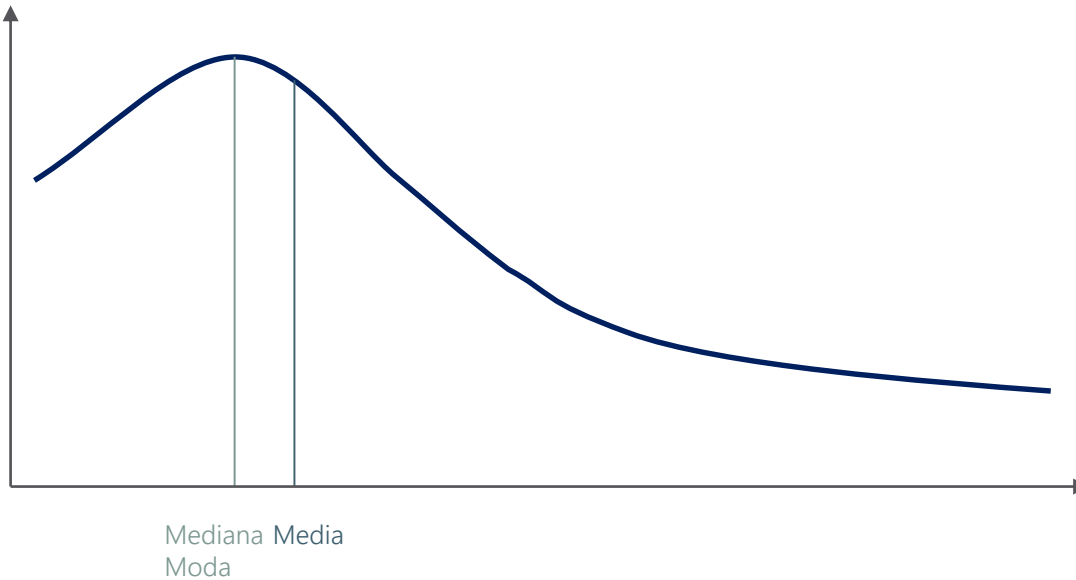
La moda se calcula simplemente encontrando el valor con la frecuencia más alta.

 En Excel, la moda se calcula con:

=MODA.UNO() -> da una moda

=MODA.VARIOS() -> da diversas modas. Se utiliza cuando tenemos más de 1 moda.

# Sesgo



El sesgo es una medida de asimetría que indica si las observaciones en un conjunto de datos se concentran en un lado.

El sesgo a la derecha (positivo) se parece a la del gráfico. Significa que los valores atípicos están a la derecha (cola larga a la derecha).

El sesgo hacia la izquierda (negativo) significa que los valores atípicos están a la izquierda.

Por lo general, se utiliza software para calcular la asimetría.

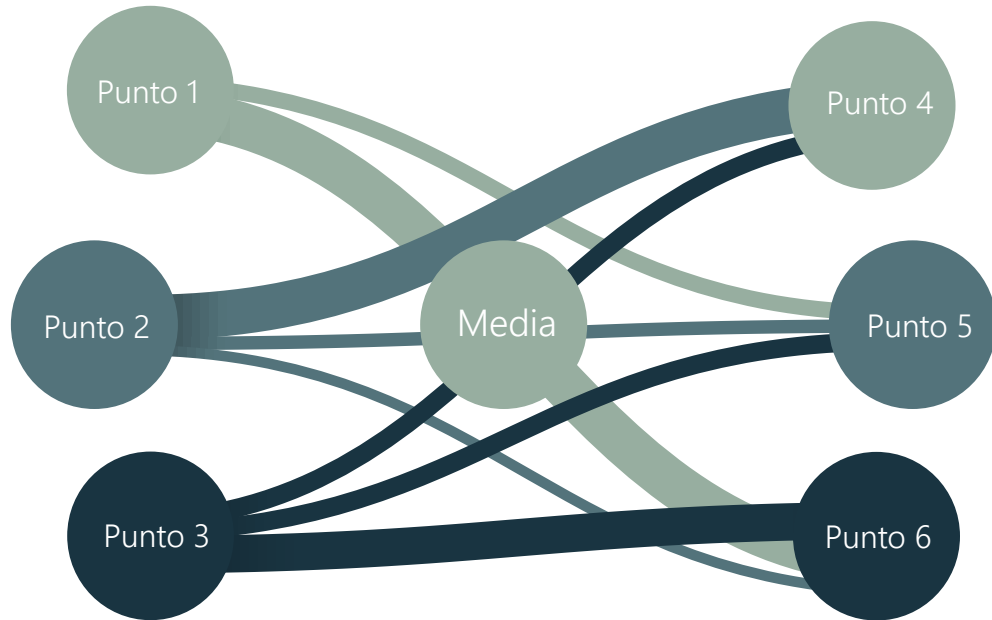


Calculando el sesgo en Excel:

=COEFICIENTE.ASIMETRIA()

Formula para calcular el sesgo: 
$$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}^3}$$

# Varianza y desviación estándar



## Calculando la varianza en Excel:

Varianza Poblacional: =VAR.P()

Desviación estándar muestral: =DESVEST()

Desviación estándar poblacional: =DESVEST.P()

La varianza y la desviación estándar miden la dispersión de un conjunto de puntos de datos en torno a su valor medio.

Existen diferentes fórmulas para las varianzas y la desviaciones estándar muestrales y poblacionales. Esto se debe a que las fórmulas de la muestra son los estimadores imparciales de las fórmulas de la población. Más sobre las matemáticas que hay detrás. [Más sobre las matemáticas que hay detrás.](#)

Fórmula de varianza muestral:

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

Fórmula de varianza poblacional:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Fórmula de desviación estándar muestral:

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

Fórmula de desviación estándar poblacional:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

# Covarianza y correlación

## Covarianza

La covarianza es una medida de la variabilidad conjunta de dos variables.

- Una covarianza positiva significa que las dos variables se mueven juntas.
- Una covarianza de 0 significa que las dos variables son independientes.
- Una covarianza negativa significa que las dos variables se mueven en direcciones opuestas.

La covarianza puede asumir valores desde  $-\infty$  hasta  $+\infty$ . Esto es un problema, ya que es muy difícil poner tales cifras en perspectiva.

Fórmula de covarianza muestral:  $s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

Fórmula de covarianza poblacional:  $\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x) * (y_i - \mu_y)}{N}$

 En Excel, la covarianza se calcula con:

Covarianza muestral: =COVAR()

Covarianza poblacional: =COVARIANCE.P()

## Correlación

La correlación es una medida de la variabilidad conjunta de dos variables. A diferencia de la covarianza, la correlación podría considerarse una medida estandarizada. Asume valores entre -1 y 1, por lo que es fácil para nosotros interpretar el resultado.

- Una correlación de 1, conocida como correlación positiva perfecta, significa que una variable está perfectamente explicada por la otra.
- Una correlación de 0 significa que las variables son independientes.
- Una correlación de -1, conocida como una correlación perfecta negativa, significa que una variable está explicando a la otra perfectamente, pero se mueven en direcciones opuestas.

Fórmula de correlación muestral:  $r = \frac{s_{xy}}{s_x s_y}$

Fórmula de correlación poblacional:  $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

 En Excel, la correlación se calcula con:

=COEF.DE.CORREL()