Analítica Descriptiva

¿Qué es analítica descriptiva?

- La ***analítica descriptiva*** es una etapa preliminar del procesamiento de datos que crea un resumen de los datos históricos para proporcionar información útil y preparar los datos para su posterior análisis.
- Para responder a la pregunta «¿Qué pasó en el negocio?» se emplea la analítica descriptiva.
- A través de ella, se analizan los datos y la información para describir la situación actual de los negocios de una manera que las tendencias, patrones y excepciones se hacen evidentes. Esto después toma la forma de informes , cuadros de mando, etc.

Analítica Descriptiva

https://www.youtube.com/watch?v=WERLckcjdfc&ab_channel=MildredQuispe

¿Por qué utilizar la analítica descriptiva?

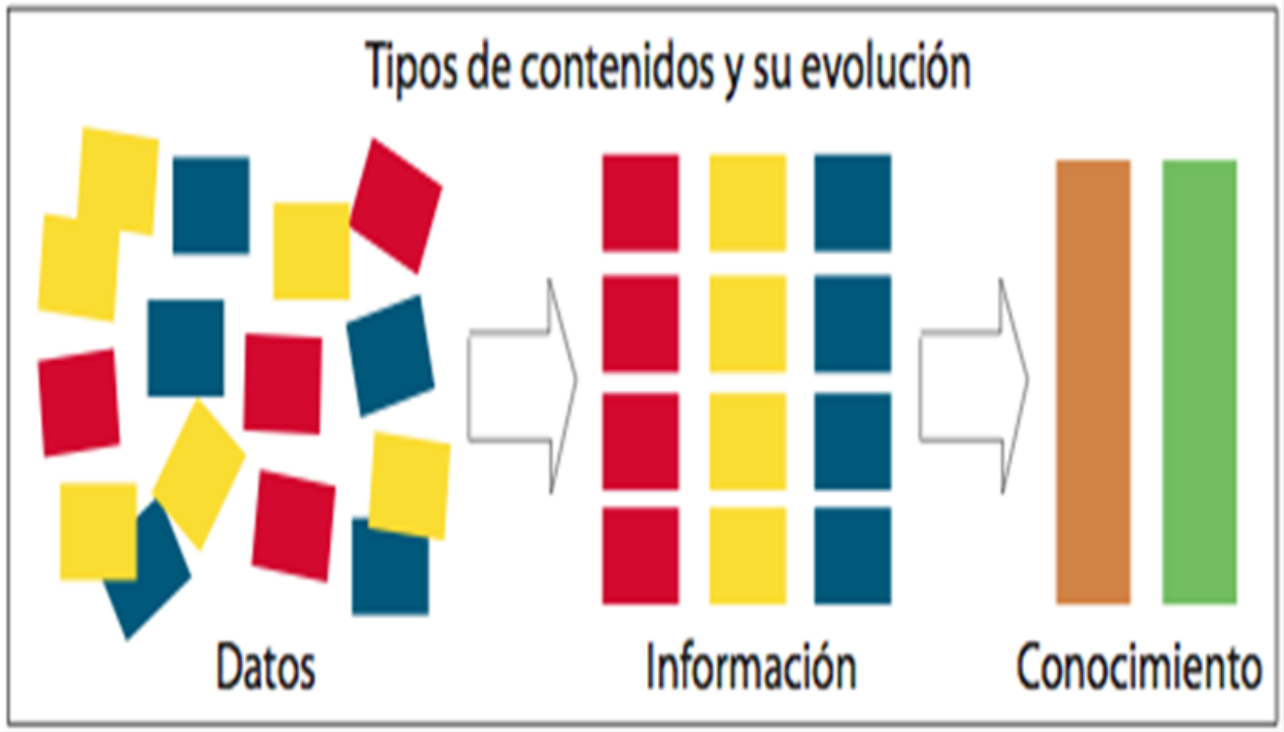
- Ayuda a las organizaciones a entender lo que sucedió en el pasado (el pasado en este contexto puede ser desde hace un minuto o unos pocos años atrás).
- Permite entender la relación entre los clientes y los productos, siendo su objetivo obtener una comprensión del enfoque que se va a adoptar en el futuro: aprender del comportamiento pasado para así influir en los resultados futuros.
- La gran mayoría de las estadísticas que se utiliza entran en esta categoría.
- Por lo general, los datos subyacentes son un recuento de datos a los que se aplica matemáticas básicas. Para todos los propósitos prácticos, hay un número infinito de estas estadísticas.
- Es una etapa preliminar de procesamiento de datos que crea un resumen de los datos históricos para proporcionar información útil y, de esta manera, preparar los datos para su posterior análisis.
- La minería de datos y su tratamiento organiza los datos y hace posible identificar patrones y relaciones en los mismos que de otro modo no ser visibles. De esta forma, la consulta, información y visualización de datos se pueden aplicar para obtener una visión más clara.

¿Qué es Business Intelligence (BI)?

- El término Business Intelligence (BI por sus siglas en inglés) hace referencia al uso de estrategias y herramientas que sirven para transformar información en conocimiento, con el objetivo de mejorar el proceso de toma de decisiones en una empresa.
- BI ayuda a interpretar datos pasados, la ciencia de datos puede analizar los datos pasados (tendencias o patrones) para hacer predicciones futuras.
- BI se utiliza principalmente para informes o análisis descriptivo; mientras que la ciencia de datos se usa más para análisis predictivo o análisis prescriptivo.
- Combina por tanto **información interna y externa** de muy diversa procedencia, desde datos e información propia, a información externa que puede ser guías, informes, recortes de periódico, etc.

Datos, información y conocimiento

- **Datos**: números y hechos sin significado al usuario final
- **Información**: es un conjunto de datos con significado y utilidad
  - Presentada por variables claves que permiten conocer la situación y dar seguimiento
    - Posición en el mercado
    - Innovación
    - Productividad
    - Recursos físicos y financieros
    - Responsabilidad social
- **Conocimiento:** información procesada y presentada de manera ordenada y útil.



La información debe tener temporalidad, relevancia, precisión, disponibilidad y retroalimentación.

Métricas y Clasificación de los Datos

Una métrica es una unidad de medida que proporciona una forma de cuantificar objetivamente el desempeño.

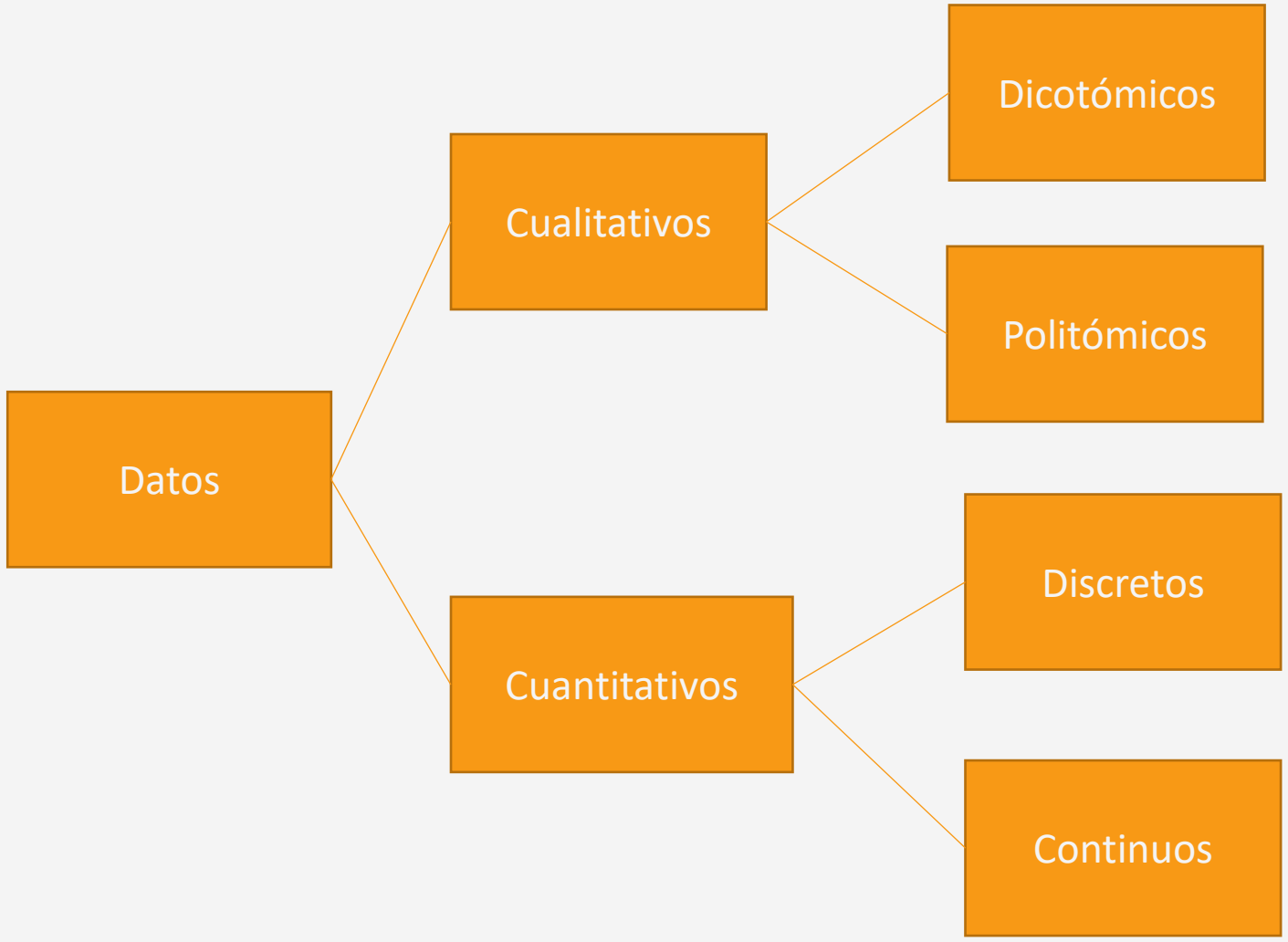
La medición es el acto de obtener datos asociados con una métrica. Las medidas son valores numéricos asociados con una métrica.

Las métricas pueden ser continuas o discretas.

Una métrica discreta es aquella que se deriva de contar algo.

Las métricas continuas se basan en una escala de medición continua.

Tipos de datos según su naturaleza:



Datos por el tipo de escala de medición

Escalas cualitativas:

- **Categóricos (nominales):** se clasifican en categorías según características especificadas. Las categorías no tienen una relación cuantitativa entre sí, pero se suele asignar un número arbitrario a cada categoría para facilitar el proceso de administrar los datos y computar estadísticas. Los datos categóricos suelen ser contados o expresados como proporciones o porcentajes.
- **Ordinales:** permiten ordenar o clasificar según alguna relación a otro. Los datos ordinales son más significativos que los datos categóricos porque los datos se pueden comparar entre sí. Dichos datos son categóricos pero también tienen un orden natural (excelente es mejor que muy bueno) y, en consecuencia, son ordinales. Sin embargo, los datos ordinales no tienen unidades de medida fijas, por lo que no podemos hacer declaraciones numéricas sobre las diferencias entre categorías.

Escalas cuantitativas:

- **De intervalo:** son datos ordinales pero tienen diferencias constantes entre observaciones y tienen puntos cero arbitrarios. A diferencia de los datos ordinales, los datos de intervalo permiten una comparación significativa de rangos, promedios y otras estadísticas.
- **De razones o radios:** Identifican, ordenan, ubican en escala y en función a un cero absoluto a datos o valores. El valor cero representa la ausencia total de la característica a ser medida; son datos continuos y tienen un cero natural.

Tipos de variables y datos

https://www.youtube.com/watch?v=Tb3sgUSd2SQ

Validez y confiabilidad de la información:

- Validez: establece si los resultados obtenidos cumplen todos los requisitos del método de investigación científica.
  - La más importante desde la perspectiva de análisis
  - Establece si lo que se mide es lo que se quiere medir
- Confiabilidad: mide el grado de repetitividad o reproducibilidad de los resultados obtenidos en el estudio bajo iguales condiciones.
  - Tiene que ver con la congruencia y exactitud de los procesos de medición

Factores que afectan la validez y confiabilidad de la información:

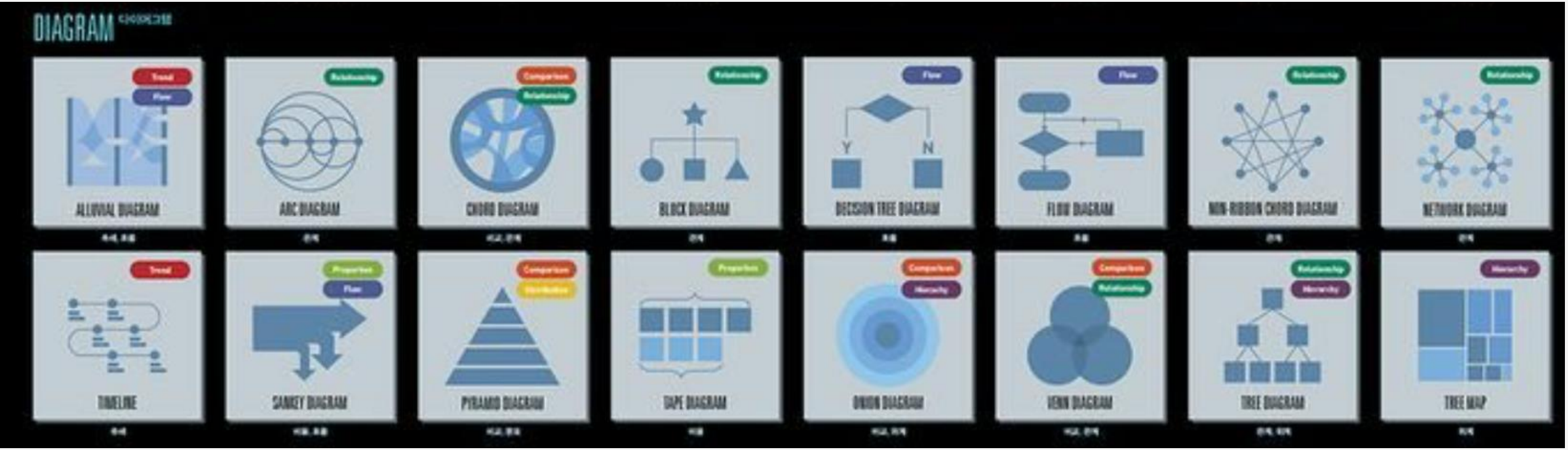
- La improvisación.
- El no estar validados en el contexto donde se aplican.
- El instrumento es inadecuado o no es empático.
- Las condiciones en que se aplica el instrumento.
- Aspectos mecánicos y de diseño.

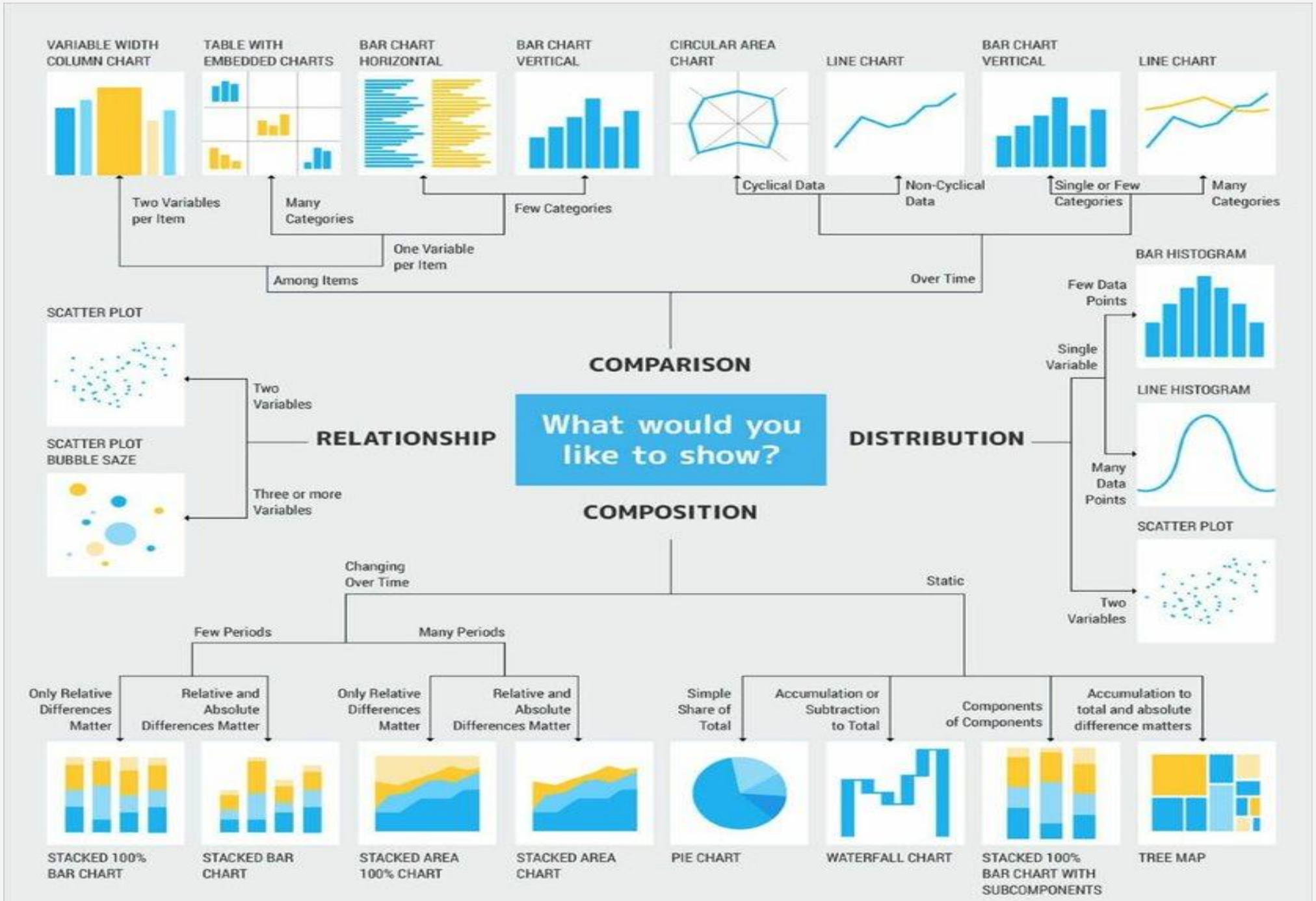
Visualización de datos

- Es la representación gráfica de información y datos.
- Es una herramienta cada vez más importante para darle sentido a las billones de filas de datos que se generan cada día.
- La visualización de datos ayuda a contar historias seleccionando los datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos.
- Una buena visualización cuenta una historia, eliminando el ruido de los datos y resaltando la información útil.
- La visualización eficaz de datos es un delicado equilibrio entre forma y función.
- La gráfica más simple podría ser demasiado aburrida para captar la atención del público o lograr que diga algo importante.
- Por el contrario, la visualización más sorprendente podría fallar por completo a la hora de transmitir el mensaje correcto o podría decir mucho.
- Los datos y los elementos visuales deben trabajar juntos, y hay algo de arte en combinar un gran análisis con una gran narración.

Ejemplos de gráficos y tablas

Ejemplos de diagramas




What would you like to show?
COMPARISON
RELATIONSHIP
DISTRIBUTION
COMPOSITION
VARIABLE WIDTH COLUMN CHART
TABLE WITH EMBEDDED CHARTS
BAR CHART HORIZONTAL
BAR CHART VERTICAL
CIRCULAR AREA CHART
LINE CHART
BAR CHART VERTICAL
LINE CHART
Two Variables per Item
Many Categories
Few Categories
One Variable per Item
Among Items
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
Over Time
SCATTER PLOT
SCATTER PLOT BUBBLE SAZE
Two Variables
Three or more Variables
BAR HISTOGRAM
Few Data Points
Single Variable
LINE HISTOGRAM
Many Data Points
SCATTER PLOT
Two Variables
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Accumulation to total and absolute difference matters
STACKED 100% BAR CHART
STACKED BAR CHART
STACKED AREA 100% CHART
STACKED AREA CHART
PIE CHART
WATERFALL CHART
STACKED 100% BAR CHART WITH SUBCOMPONENTS
TREE MAP

Algunos ejemplos interesantes

- Los 10 mejores blogs de visualización de datos que debes seguir https://www.tableau.com/es-mx/learn/articles/best-data-visualization-blogs

- Review of 20 best big data visualization tools https://bigdata-madesimple.com/review-of-20-best-big-data-visualization-tools/

Aspectos que debe tener en cuenta al evaluar plataformas de BI

- Análisis visual verdadero
  - Le permitirá crear un dashboard atractivo si cuenta con los requisitos o un prototipo que replicar.
- Plataforma integrada
  - Debe ofrecer acceso a los datos, preparación de datos, análisis, colaboración y gobernanza, todo desde una misma plataforma.
- El ritmo de la innovación
  - Debe ofrecer características innovadoras a sus usuarios.
- Flexibilidad y variedad de opciones
  - Todas las funcionalidades se pueden implementar en las instalaciones físicas o en la nube, en IOS, Windows o Linux, con datos en tiempo real o almacenados en la memoria.
- Comunidad
  - Lo ideal sería que la plataforma seleccionada tenga acceso a una comunidad de usuarios que permita resolver dudas y desarrollar nuevas aplicaciones.

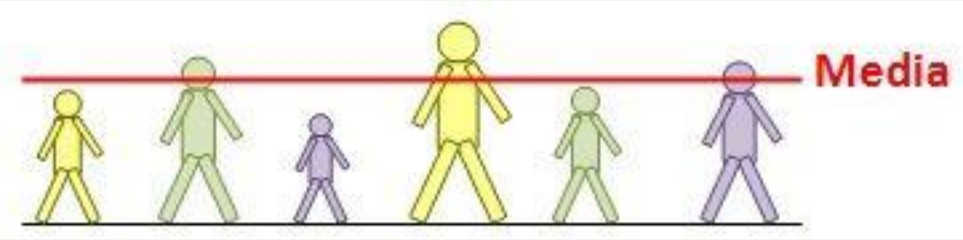
Introducción a la estadística descriptiva

- La **estadística descriptiva** es la técnica matemática que obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso, generalmente con el apoyo de tablas, medidas numéricas o gráficas.
- Una población es un conjunto de seres, individuos, objetos, casos elementos o eventos que presentan determinadas características.
- Una muestra de una población es un subconjunto representativo de esta.
- Un parámetro es un número que se obtiene gracias a una distribución de datos estadísticos y ayuda a organizar la información dada ya sea por una gráfica o una tabla.
- Un estadístico es un valor resumido, calculado, como base en una muestra de observaciones que generalmente, aunque no por necesidad, se considera como una estimación de parámetros de determinada población; es decir, una función de valores de muestra.
- Los principales tipos son:
  - **Centralización**.
  - **Posición**.
  - **Dispersión**.

Medidas de exactitud o tendencia central

- Las **medidas de tendencia central** (o de centralización) son medidas que tienden a localizar en qué punto se encuentra la **parte central** de un **conjunto ordenado** de datos de una variable cuantitativa.
- Media aritmética $\bar{x}$:
- La **media** $\bar{x}$ (también llamada **promedio** o **media aritmética**) de un conjunto de datos (X_1,X_2,…,X_N) es una medida de posición central.
- Se define como el valor característico de la serie de datos resultado de la suma de todas las observaciones dividido por el número total de datos.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$



- Si se trata de los datos (X_1,X_2,…,X_N) de una muestra, se tiene la **media muestral**. Si el conjunto de datos es toda la población, se llama **media poblacional**y y se expesa con la letra griega μ.
- El estadístico media muestral y la medida **media poblacional** son dos conceptos distintos, ya que el primero es un valor estimado a partir de una muestra mientras que el segundo es un valor medido sobre una población. Pero ambos se calculan igual.
- La desventaja de la media aritmética es que si hay valores extremos alejados, no resulta el promedio más indicado.

Media a partir de frecuencias o datos agrupados

- En un conjunto de datos discretos agrupados en frecuencias, podemos calcular el **promedio** o media aritmética a partir de las frecuencias relativas de las observaciones distintas.
- La **frecuencia relativa (f_i)** de un valor X_i es la **proporción** de valores iguales a X_i en el conjunto de datos (X_1, X_2,…, X_N). Es decir, la frecuencia relativa es la frecuencia absoluta dividida por el número total de elementos *N*:

$$Media(X) = \sum_j X_j \cdot f_j$$

Siendo X_j las observaciones distintas,
f_j las frecuencias relativas

$$f_i = \frac{n_i}{N}$$

siendo $(X_1, X_2, ..., X_N)$ el conjunto de datos
y n_i el total de valores igual a X_i

Otros tipos de media:

- **Media geométrica**: se calcula sobre un conjunto de números **estrictamente positivos**. Es la raíz N-ésima del producto de los *N* elementos. Está indicada para calcular medias de **porcentajes, tantos por uno, puntuaciones o índices**. Tiene la ventaja de que no es tan sensible a los valores extremos.

$$MG = \sqrt[N]{X_1 \cdot X_2 \cdot \ldots \cdot X_N}$$

- **Media armónica**: es el **recíproco** de la suma de los recíprocos (donde $1/X_i$ es el recíproco de X_i)) multiplicado por el número de elementos del conjunto. Suele utilizarse principalmente para calcular la media de **velocidades, tiempos o en electrónica**.

$$H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}}$$

- **Media cuadrática**: se define como la raíz cuadrada del promedio de los elementos al cuadrado. La **media cuadrática** es muy útil para variables que toman valores negativos y positivos y su signo no es importante e interesa el valor absoluto del elemento. Por ejemplo, los **errores de medida**, el **valor eficaz** de un parámetro sinusoidal en electricidad, etc.

$$RMS = \sqrt{\frac{X_1^2 + X_2^2 + \ldots + X_N^2}{N}}$$

- **Media ponderada**: consiste en otorgar a cada observación del conjunto de datos unos **pesos** según la importancia de cada elemento. Tiene numerosas aplicaciones, como el cálculo del **IPC (Índice de Precios de Consumo)**, calcular la **nota media de una asignatura** ponderando exámenes, trabajos, etc.

$$MP = \frac{p_1 X_1 + p_2 X_2 + \ldots + p_N X_N}{p_1 + p_2 + \ldots + p_N}$$

siendo $(X_1, X_2, \ldots, X_N)$ el conjunto de datos y $(p_1, p_2, \ldots, p_N)$ los pesos

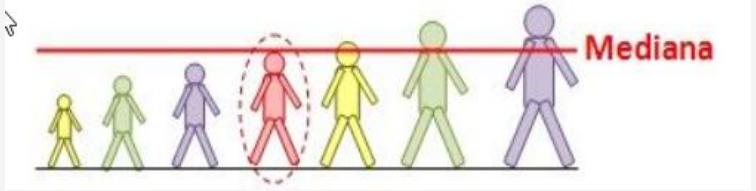
- Cuando el conjunto de pesos es igual a 1, se tiene la media por frecuencias.

Otras medidas de tendencia central

- Visto desde un punto de vista más conceptual, la **media aritmética** es el centro de los datos en el sentido numérico, ya que intenta equilibrarlos por exceso y por defecto. Es decir, si se suman todas las diferencias de los datos a la media el total será cero.

$$\sum_{i=1}^{N}(x_i - \overline{x}) = 0$$

- La **Mediana** (Mediana (X)) es el elemento de un conjunto de datos ordenados (X_1,X_2,…,X_N) que **deja a izquierda y derecha la mitad de valores**.
  - Sea (X_1,X_2,…,X_N) un conjunto de datos ordenado. El cálculo de la **mediana** depende de si el número de elementos *N* es par o impar.
  - Si *N* es **impar**, la mediana es el valor que está al medio, es decir:

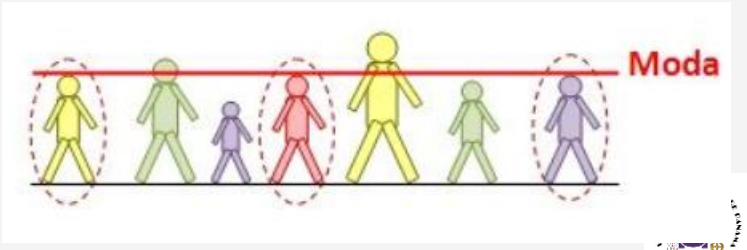


$$Mediana(X) = X_{\frac{N+1}{2}}$$

  - Si *N* es **par**, la mediana es la media de los dos valores del centro, N/2 y N/2+1

$$Mediana(X) = Media(X_{\frac{N}{2}}, X_{\frac{N}{2}+1}) = \frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2}$$

- La **moda** (**Mod(X)**) es el valor más repetido del conjunto de datos, es decir, el valor cuya frecuencia relativa es mayor. En un conjunto puede haber más de una moda.



Medidas de dispersión

- Las **medidas de dispersión** o **medidas de variabilidad** muestran la **variabilidad** de un conjunto de datos, indicando la mayor o menor concentración de datos respecto a las medias de centralización.
- El **Rango** (R) o recorrido estadístico es la diferencia entre el valor máximo y el mínimo de un conjunto de elementos.

$$Rango = (Max) - (Min)$$

- El **rango intercuartílico** (**IQR)** es una estimación estadística de la dispersión de una distribución de datos.
- Consiste en la diferencia entre el **tercer** y el **primer** cuartil, lo que elimina valores extremos.
- Los **cuartiles** son los tres elementos de un conjunto de datos ordenados que dividen el conjunto en cuatro partes iguales.
- Para datos no agrupados. Para el cuartil 1 (Q_1) y cuartil 3 (Q_3) se encuentra su posición mediante los siguientes pasos: (N+1)/4 y 3(N+1)/4
- El rango intercuartílico es altamente recomendable cuando la medida de tendencia central utilizada es la mediana (ya que este estadístico es insensible a posibles irregularidades en los extremos).
- El **cuartil 1** (Q_1) es el percentil 25 (P_{25}).
- El **cuartil 2** (Q_2) es la mediana y el percentil 50 (P_{50}).
- El **cuartil 3** (Q_3) es el percentil 75 (P_{75}).

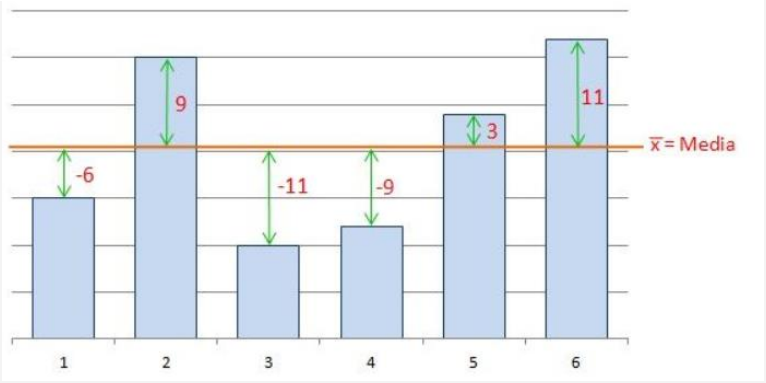
Medidas de dispersión: la varianza

- La **varianza** (S^2) mide la **dispersión** de los datos de una **muestra** respecto a la media, calculando la media de los cuadrados de las distancias de todos los datos.
- Si en vez de tratarse de una muestra, la **varianza** se refiere a la población, el denominador será *N*.
- La fórmula de la **varianza poblacional**, de símbolo σ^2, es:
- Cuanto mayor sea *N* menor será la diferencia entre el resultado de la fórmula S_X^2 y la de σ^2.
- Un inconveniente de la varianza es que sus unidades son las unidades de los datos al cuadrado.

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos y $\overline{x}$ la media

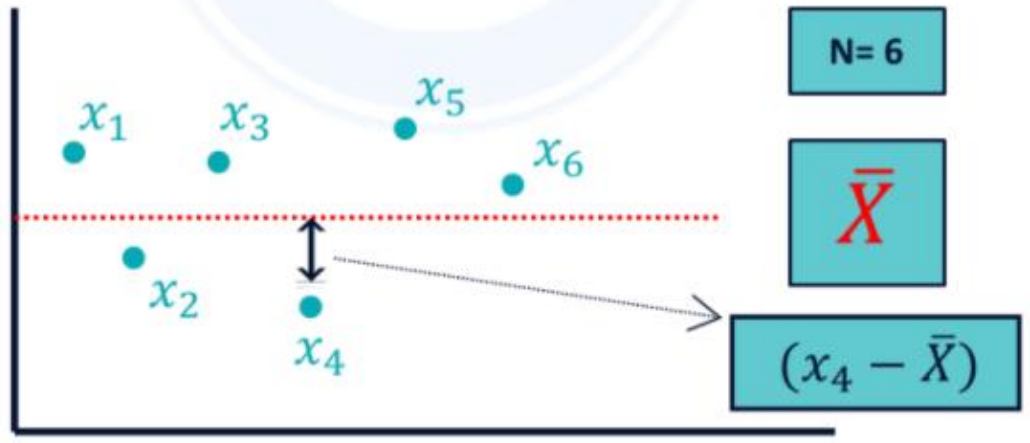
$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N}$$



Medidas de dispersión: la desviación típica o estándar

- Es la medida de dispersión (S) asociada a la media.
- Es una medida que ofrece información sobre la dispersión media de una variable. La desviación estándar es siempre mayor o igual que cero.
- Mide el promedio de las desviaciones de los datos respecto a la media en las mismas unidades de los datos.
- El cuadrado de la desviación típica es la Varianza.

También conocida como desviación típica σ es una medida que ofrece información sobre la dispersión media de una variable.



$$S_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{N-1}}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos

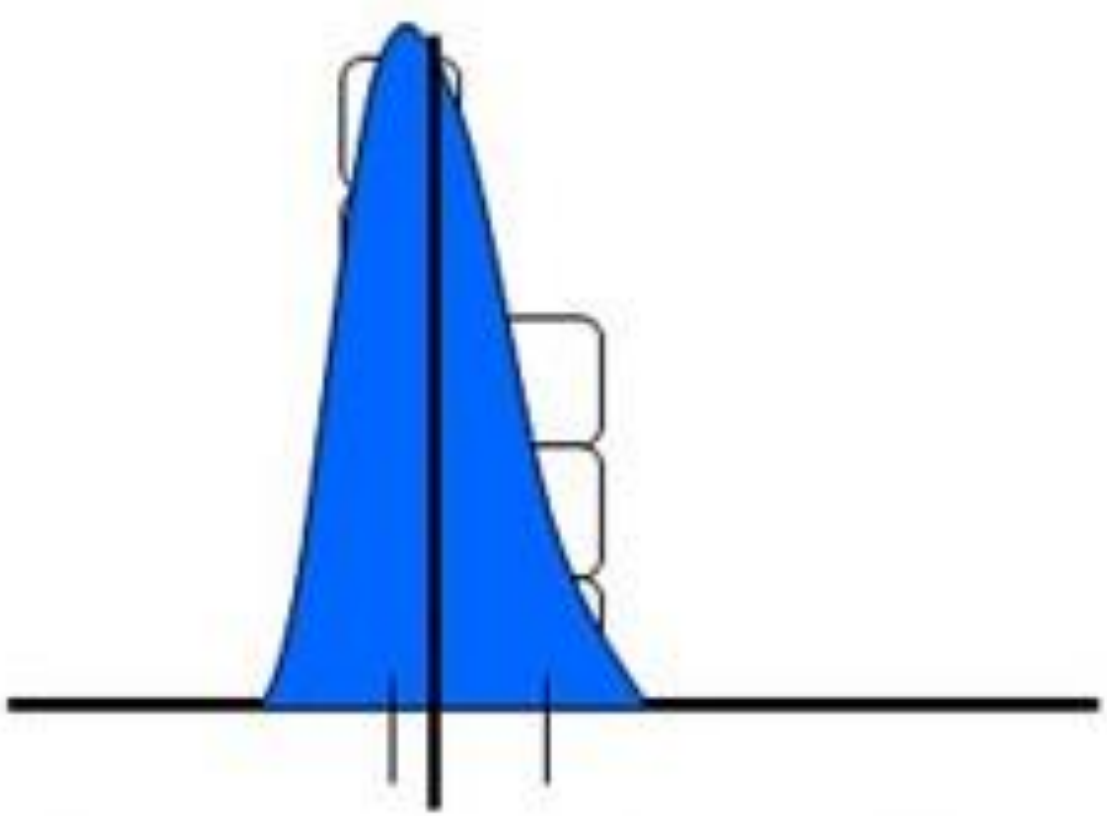
$$\sigma = \sqrt{\frac{\sum_{i}^{N}(X_i - \bar{X})^2}{N}}$$

Coeficiente de variabilidad

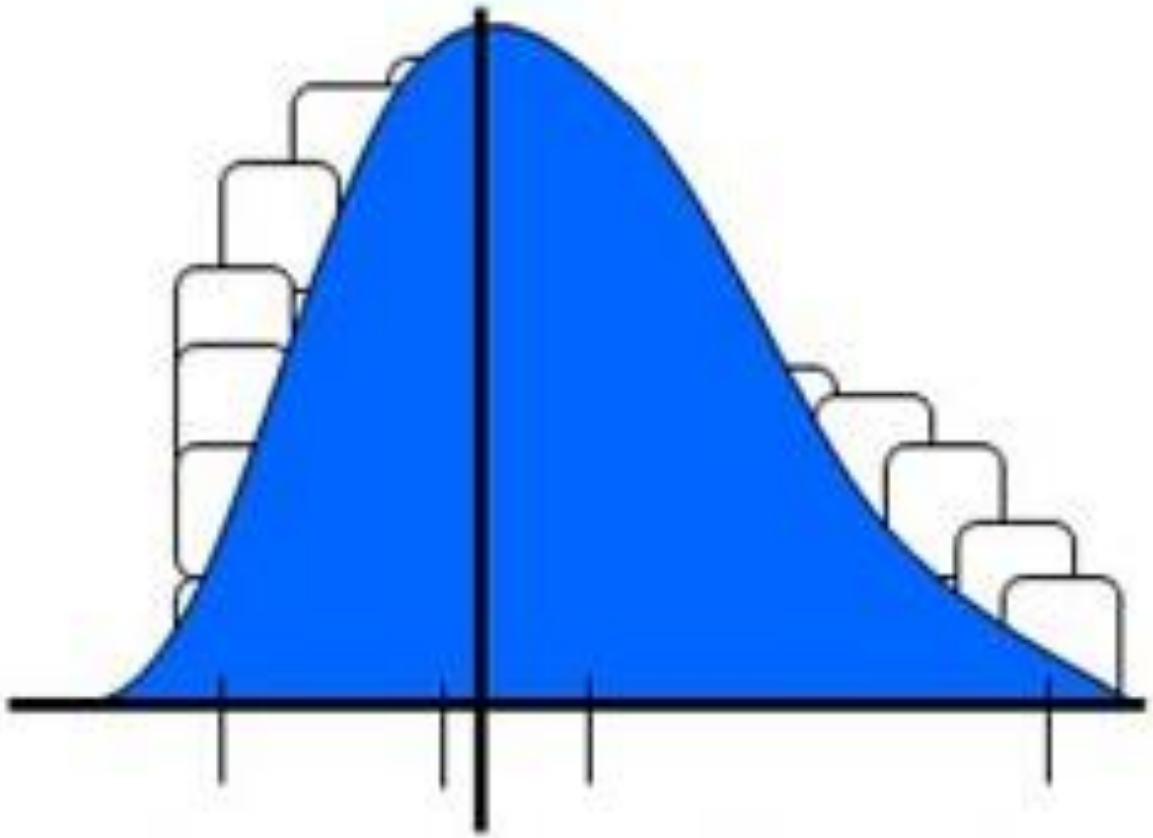
$$CV = \frac{\sigma_x}{|\bar{X}|}$$

- El coeficiente de variación, también denominado como coeficiente de variación de Pearson, es una medida estadística que nos informa acerca de la dispersión relativa de un conjunto de datos.
- Se utiliza para comparar conjuntos de datos pertenecientes a poblaciones distintas.
- Su cálculo tiene en cuenta el valor de la media. Por lo tanto, el coeficiente de variación permite tener una medida de dispersión que elimine las posibles distorsiones de las medias de dos o más poblaciones.

Ejemplo de dos conjuntos de datos con igual media



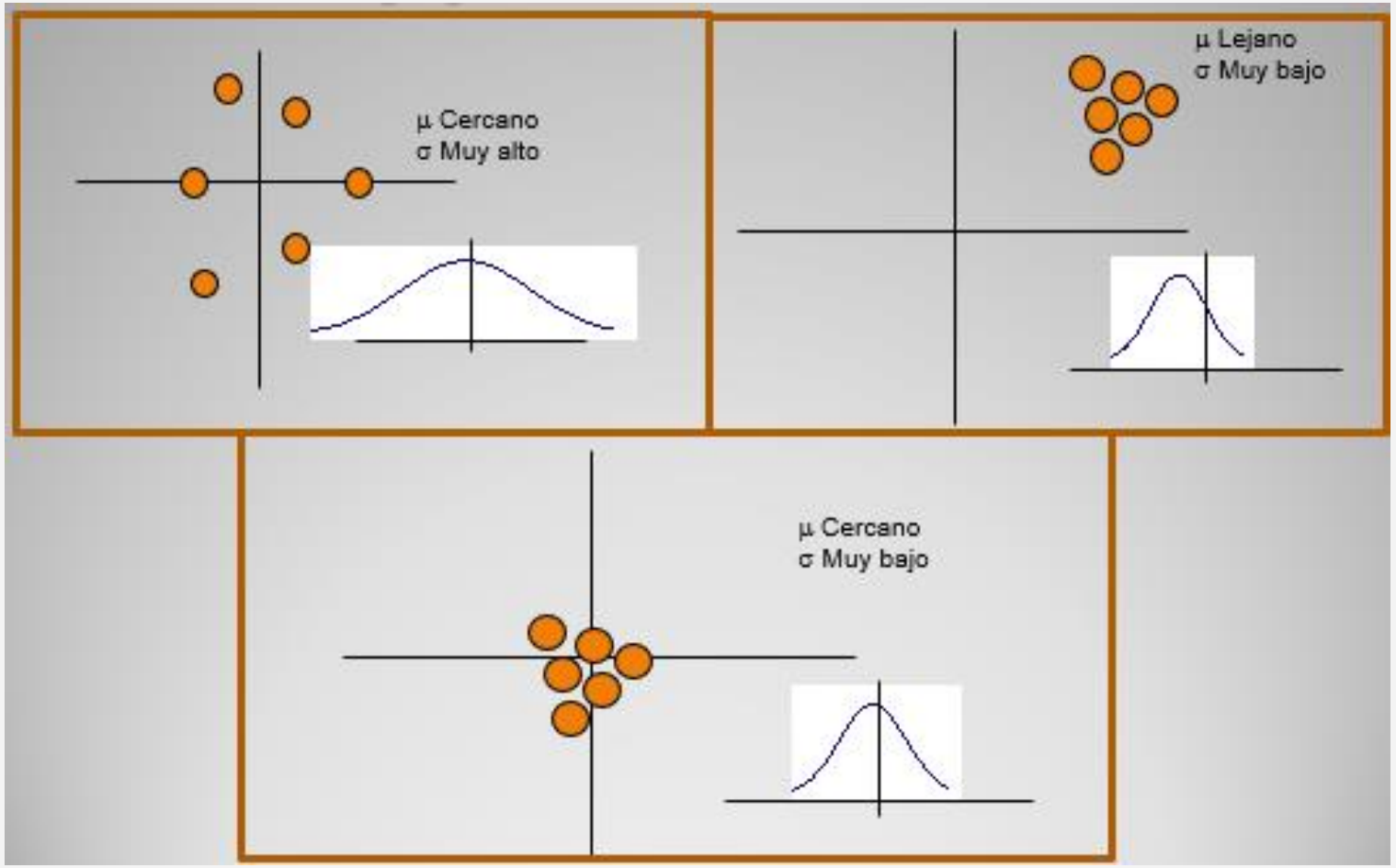
Datos con baja dispersión



Datos con alta dispersión

Exactitud y precisión de la información:

- **Exactitud** es el grado en el cual la información se muestra verdadera o con valores aceptables. Que tan cerca están los datos de una medida de tendencia central.
- **Precisión** es el grado de dispersión en la información.

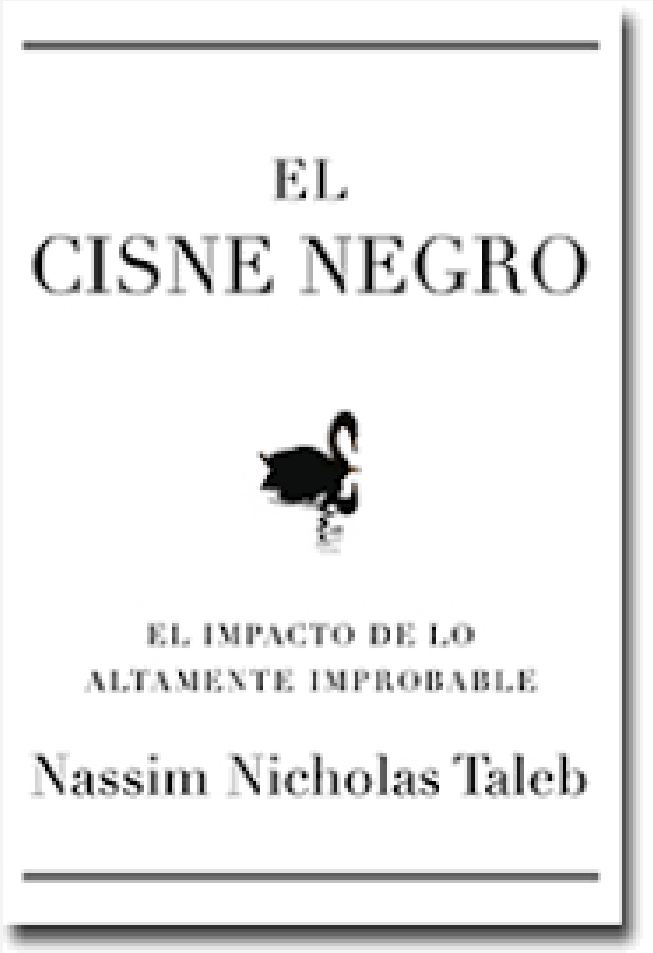


Comportamiento de datos con alta variabilidad

- Como se ha dicho, la **variabilidad** es una medida de la dispersión de lo s datos en una distribución, sea esta teórica o de una muestra.
- Cuanto **mayor** sea el valor de las medidas de variabilidad, **mayor** será la **variabilidad**, y cuanto menor sea, más homogénea será a la media. Cuando es cero quiere decir que todos los datos son iguales.
- Cuanto menor es la **variabilidad**, más homogénea es la muestra de sujetos en la variable.
- En el caso de máxima **homogeneidad**, todos los valores de la variable serán iguales.
- De otro modo, cuanto más o menos dispersión en los **datos**, la muestra es más o menos heterogénea y las puntuaciones difieren entre sí.
- A mayor valor del **coeficiente de variación** mayor heterogeneidad de los valores de la variable; y a menor C.V., mayor homogeneidad en los valores de la variable.

Fenómeno cisne negro

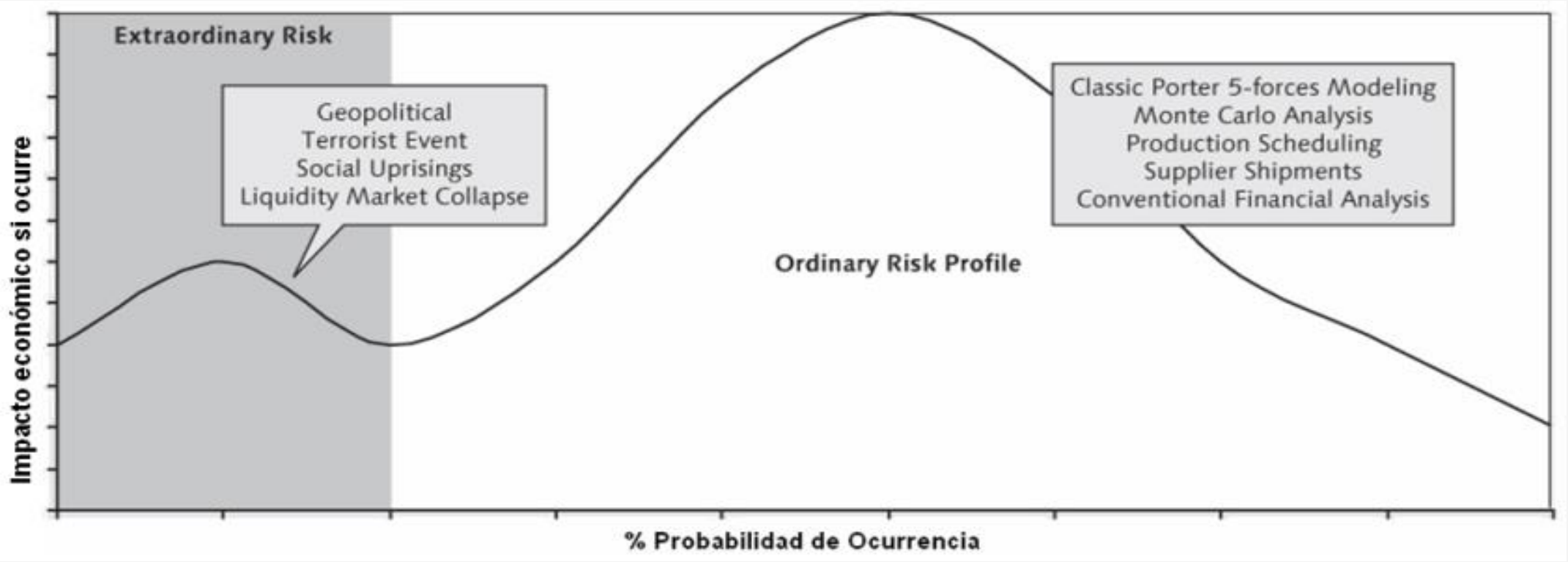
- La **teoría del cisne negro** o **teoría de los sucesos del cisne negro** es una metáfora que describe un suceso *sorpresivo* (para el observador), de gran impacto socioeconómico y que, una vez pasado el hecho, se racionaliza por *retrospección* (haciendo que parezca predecible o explicable, y dando impresión de que *se esperaba que ocurriera*).
- Fue propuesta por Nassim Taleb en su libro de igual título.
- A diferencia del problema filosófico anterior del cisne negro, la teoría del cisne negro se refiere solo a los sucesos inesperados de gran magnitud, consecuencia y su papel dominante en la historia. Estos hechos, considerados atípicos extremos, colectivamente juegan roles mucho más grandes que los sucesos regulares.
- Ejemplos de "cisnes negros" son el inicio de la Primera Guerra Mundial, la gripe de 1918 o los atentados del 11 de septiembre de 2001, Internet, la PC, entre otros.
- Se ha intentado identificar la pandemia de coronavirus de 2020 como un "cisne negro", pero el propio Nassim Taleb lo ha rechazado al considerar que no cumple con los requisitos de su teoría.
- Otros analistas no califican a la pandemia por coronavirus como "cisne negro" sino como "rinoceronte gris" porque era un evento predecible.



Fenómeno cisne negro

- El problema del cisne negro implica:
- El desproporcionado papel de alto impacto, difícil de predecir, y los sucesos extraños que están fuera del ámbito de las expectativas normales de la historia, la ciencia, las finanzas y la tecnología.
- La no computabilidad de la probabilidad de los sucesos raros consecuenciales utilizando métodos científicos (debido a la naturaleza misma de las probabilidades pequeñas).
- Los sesgos psicológicos que hacen a las personas individual y colectivamente ciegas a la incertidumbre e inconscientes al rol masivo del suceso extraño en los asuntos históricos.
- La idea principal no es tratar de predecir sucesos cisne negro, sino construir robustez frente a las actitudes negativas que se producen y poder aprovechar las positivas.
- Taleb sostiene que los bancos y empresas comerciales son muy vulnerables a sucesos cisne negro peligrosos y están expuestos a pérdidas superiores a los pronosticadas por los modelos estadísticos y matemáticos, que él considera defectuosos.

El Cisne Negro en el BI



- Anticiparse al futuro, no es preocuparse en el “cuándo”, sino más bien en el “cómo”.
- Técnicas como la minería de datos o lo estadística, sólo brindará una instantánea de una situación determinada, cuya validez se limita a un espacio de tiempo muy reducido, porque el comportamiento humano y otras tantas variables que conforman cualquier modelo, no obedecen a modelos matemáticos o tendencias duraderas, sino al libre albedrío.
- El Risk Intelligence (RI) considera, no tan sólo no cometer errores, sino saber detectar oportunidades, implicándose más en la cultura de las empresas y en los colaboradores de la organización, porque las decisiones se toman en todos los niveles.

https://www.youtube.com/watch?v=bSfyCO7SOmk&ab_channel=HemisferioDerecho