COLOMBIA POTENCIA DE LA VIDA

TIC

UNIVERSIDAD SERGIO ARBOLEDA

Análisis de Datos
Nivel Integrador

TALENTO TECH

UNIVERSIDAD SERGIO ARBOLEDA

PREPARACIÓN SEMANA 2

ESTADÍSTICA EN EL ANÁLISIS DE DATOS

Recordemos algunos conceptos básicos de estadística para luego aplicarlos en Python

La estadística es una rama es una rama de las matemáticas que se ocupa de recopilar, organizar, analizar, interpretar, presentar y describir conjuntos de datos. Su objetivo principal es extraer información significativa de los datos para tomar decisiones informadas. La estadística se utiliza en una variedad de campos, desde la investigación científica hasta la toma de decisiones empresariales, para entender patrones, tendencias y variaciones en los datos.

Propósito en el Análisis de Datos:

1. **Descripción de Datos:** La estadística proporciona herramientas para resumir y describir características importantes de un conjunto de datos. Esto incluye medidas como la media, mediana, moda, desviación estándar y otras que ayudan a comprender la distribución y variabilidad de los datos.

ESTADÍSTICA EN EL ANÁLISIS DE DATOS

2. **Inferencia Estadística:** La estadística permite hacer inferencias sobre una población completa basándose en una muestra representativa. Los intervalos de confianza y las pruebas de hipótesis son herramientas cruciales para realizar afirmaciones con un cierto grado de certeza.
3. **Toma de Decisiones:** Muchas decisiones en la vida cotidiana y en entornos profesionales se toman mediante el análisis estadístico. Desde estrategias empresariales hasta políticas públicas, la estadística proporciona la base para decisiones fundamentadas.
4. **Predicción y Modelado:** A través de técnicas como la regresión, la estadística ayuda a prever el comportamiento futuro de variables basándose en patrones pasados. Esto es esencial en campos como la economía, la ciencia de datos y la investigación.

Tipos de Datos: Cualitativos y Cuantitativos

1. **Datos Cualitativos:**
   - Son atributos no numéricos que describen cualidades o características.
   - Ejemplos: colores, estados civiles, opiniones.
   - Se pueden clasificar en nominales (sin orden) u ordinales (con orden).
2. **Datos Cuantitativos:**
   - Representan cantidades numéricas y se pueden medir.
   - Ejemplos: edad, ingresos, temperatura.
   - Se dividen en discretos (valores distintos y separados) y continuos (valores en un rango).

Tipos de Datos: Cualitativos y Cuantitativos

DIFERENCIAS ENTRE DATOS CUANTITATIVOS Y CUALITATIVOS

Datos cuantitativos	Datos cualitativos
Asociado con números.	Asociados a los detalles.
Se implementa cuando los datos son numéricos.	Se implementa cuando los datos pueden ser segregados en grupos bien definidos.
Los datos recolectados pueden ser analizados estadísticamente.	Los datos recogidos sólo pueden ser observados y no evaluados.
Ejemplos: Altura, peso, tiempo, precio, temperatura, etc.	Ejemplos: Aromas, Apariencia, Belleza, Colores, Sabores, etc.

Tomado de: https://www.eduardvelazquez.com/3-ventajas-recolectar-datos-cuantitativos/

MEDIDAS DESCRIPTIVAS

Entre las medidas descriptivas tenemos

1. **Media:**
    - La media es el promedio aritmético de un conjunto de datos. Se calcula sumando todos los valores y dividiendo el resultado por el número total de observaciones.
    - Fórmula: $\bar{X} = \frac{\sum_{i-1}^{n} X_i}{n}$
    - La media es sensible a valores extremos, por lo que puede no ser representativa si existen valores atípicos en el conjunto de datos.
2. **Mediana:**
    - La mediana es el valor que se encuentra en el centro de un conjunto de datos ordenados. Divide el conjunto en dos partes iguales, siendo el punto medio.
        - Es menos sensible a valores extremos que la media y es útil cuando se trabaja con distribuciones asimétricas.

MEDIDAS DESCRIPTIVAS

Entre las medidas descriptivas tenemos

1. **Media:**
    - La media es el promedio aritmético de un conjunto de datos. Se calcula sumando todos los valores y dividiendo el resultado por el número total de observaciones.
    - Fórmula: $\bar{X} = \frac{\sum_{i-1}^{n} X_i}{n}$
    - La media es sensible a valores extremos, por lo que puede no ser representativa si existen valores atípicos en el conjunto de datos.
2. **Mediana:**
    - La mediana es el valor que se encuentra en el centro de un conjunto de datos ordenados. Divide el conjunto en dos partes iguales, siendo el punto medio.
        - Es menos sensible a valores extremos que la media y es útil cuando se trabaja con distribuciones asimétricas.
3. **Moda:**
    - La moda es el valor que aparece con mayor frecuencia en un conjunto de datos.
    - Puede haber una moda (unimodal), más de una moda (multimodal) o ninguna moda si todos los valores son distintos.

Desviación Estándar y Varianza

1. **Desviación Estándar:**
   - Mide la dispersión de los datos en torno a la media. Una desviación estándar pequeña indica que los datos tienden a estar cerca de la media, mientras que una desviación estándar grande indica mayor dispersión.
   - Fórmula: $\sigma = \sqrt{\frac{\sum_{i-1}^{n}(X_i - \bar{X})^2}{n}}$

Varianza:

- Es el cuadrado de la desviación estándar y proporciona una medida de la variabilidad total de un conjunto de datos.
- Fórmula: $(\sigma^2) = \frac{\sum_{i-1}^{n}(X_i - \bar{X})^2}{n}$

Desviación Estándar y Varianza

Interpretación de la Dispersión

- Una baja desviación estándar/varianza indica que los datos tienden a estar cerca de la media, lo que sugiere una mayor homogeneidad.
- Una alta desviación estándar/varianza señala una mayor dispersión de los datos, indicando heterogeneidad o variabilidad significativa.
- Al comparar dispersión entre conjuntos de datos, es crucial entender las unidades en las que se miden las variables. La desviación estándar es más interpretable cuando las unidades son las mismas que los datos originales.
- Interpretar la dispersión es esencial para entender la variabilidad y la consistencia de los datos, proporcionando una visión más completa de la distribución y permitiendo tomar decisiones informadas basadas en la estabilidad o variabilidad inherente a los datos analizados.

Desviación Estándar y Varianza

Interpretación de la Dispersión

- Una baja desviación estándar/varianza indica que los datos tienden a estar cerca de la media, lo que sugiere una mayor homogeneidad.
- Una alta desviación estándar/varianza señala una mayor dispersión de los datos, indicando heterogeneidad o variabilidad significativa.
- Al comparar dispersión entre conjuntos de datos, es crucial entender las unidades en las que se miden las variables. La desviación estándar es más interpretable cuando las unidades son las mismas que los datos originales.
- Interpretar la dispersión es esencial para entender la variabilidad y la consistencia de los datos, proporcionando una visión más completa de la distribución y permitiendo tomar decisiones informadas basadas en la estabilidad o variabilidad inherente a los datos analizados.

Tipos de Datos: Cualitativos y Cuantitativos

Media

(Promedio) Suma de datos dividido entre la cantidad de los mismos.

$$\bar{x} = \frac{\sum x}{N}$$

Moda

Dato que mas se repite. Si son dos es bimodal, si son 3 es trimodal.

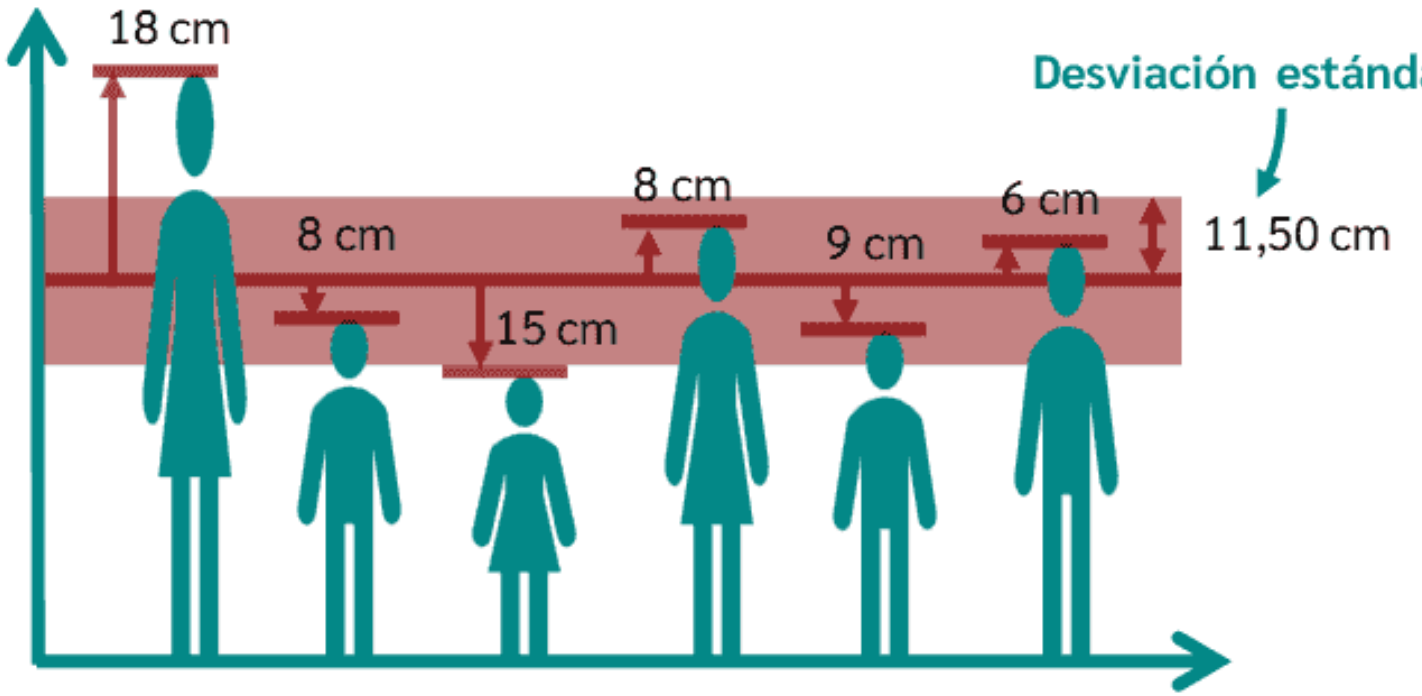
Mediana

Dato central. Si son dos se saca la media de estos.

med_x

Tomado de: https://www.geogebra.org/m/Mwbsz33s

Tipos de Datos: Cualitativos y Cuantitativos



Tomado de: https://datatab.es/tutorial/dispersion-parameter

DISTRIBUCIONES ESTADÍSTICAS

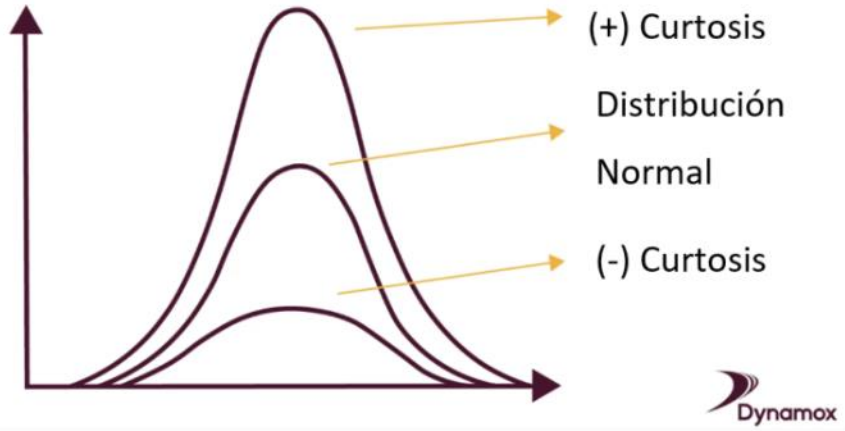
Distribución Normal y sus propiedades

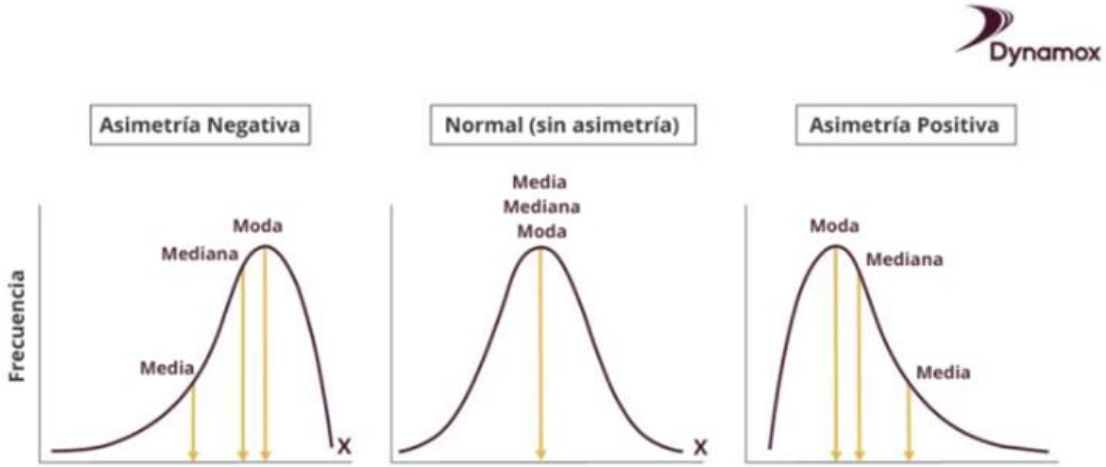
1. **Distribución Normal:**
    - La distribución normal, también conocida como la distribución de campana de Gauss, es una distribución simétrica alrededor de su media.
    - Sus propiedades clave son la media (μ) y la desviación estándar (σ), que determinan completamente su forma.
    - En una distribución normal estándar (media = 0, desviación estándar = 1), alrededor del 68% de los datos caen dentro de 1 desviación estándar, el 95% dentro de 2 desviaciones estándar y el 99.7% dentro de 3 desviaciones estándar.

DISTRIBUCIONES ESTADÍSTICAS

2. **Skewness(Simetría):**
    - La skewness mide la asimetría de una distribución. Indica si la cola de la distribución se inclina más hacia la derecha (skewness positiva) o hacia la izquierda (skewness negativa) en relación con la media.
    - Una skewness de 0 sugiere simetría, valores positivos indican asimetría hacia la derecha, y valores negativos, asimetría hacia la izquierda.
    - La skewness es útil para entender la forma general de la distribución y cómo se distribuyen los valores a lo largo de ella.
3. **Kurtosis:**
    - La kurtosis mide la "taconicidad" de una distribución, es decir, qué tan puntiaguda o achatada es en comparación con una distribución normal.
    - Una kurtosis mayor que 3 indica colas más pesadas y una distribución más puntiaguda (leptocúrtica), mientras que una kurtosis menor que 3 indica colas más ligeras y una distribución más achatada (platicúrtica).
        - La kurtosis es esencial para entender la concentración de datos alrededor de la media y cómo se distribuyen los valores extremos.

Tipos de Datos: Cualitativos y Cuantitativos





Tomado de: https://dynamox.net/es/blog/metricas-de-analisis-de-vibraciones-curtosis-y-asimetria-skewness

DISTRIBUCIONES ESTADÍSTICAS

Aplicaciones Prácticas:

- Comprender la distribución normal y sus propiedades es crucial en estadística inferencial, ya que muchos métodos se basan en la suposición de normalidad.
- La skewness y kurtosis son herramientas valiosas para diagnosticar la forma de una distribución y pueden guiar la elección de técnicas estadísticas adecuadas.
- En la interpretación de datos, identificar desviaciones de la distribución normal puede indicar patrones interesantes o inusuales en los datos.
- Estas medidas son fundamentales en la modelización estadística y en la evaluación de la validez de ciertos métodos estadísticos en diferentes escenarios.

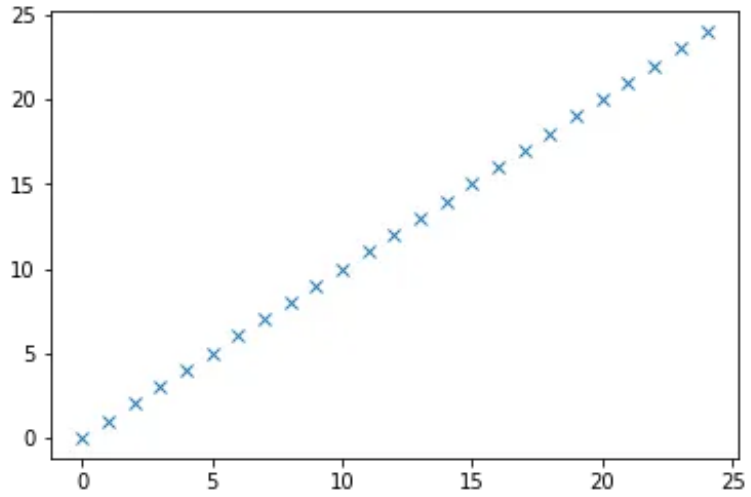
CORRELACIÓN Y COVARIANZA

Covarianza:

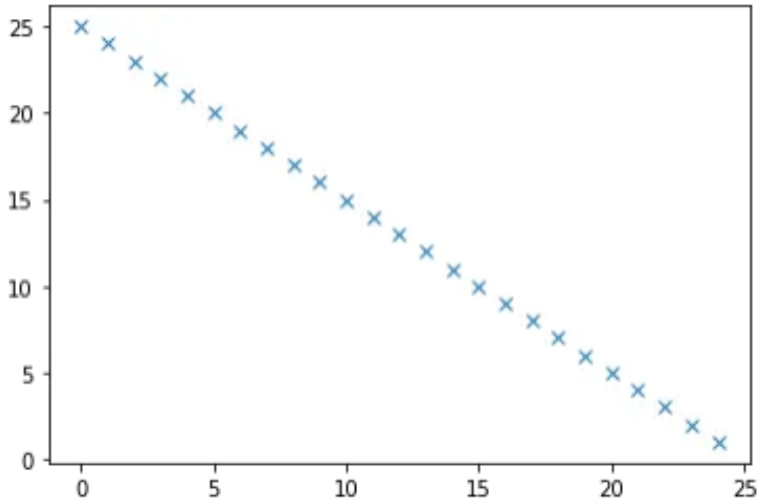
- La covarianza es una medida que evalúa cómo dos variables aleatorias varían juntas. Indica si los valores de una variable tienden a aumentar o disminuir simultáneamente con los de la otra.
- La covarianza ($Cov(X,Y)$) se calcula como la media del producto de las desviaciones de las variables respecto a sus medias.
- Fórmula: $Cov(X,Y) = \frac{\sum_{i-1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n}$
- Si la covarianza es positiva, las variables tienden a aumentar juntas. Si es negativa, una variable tiende a disminuir cuando la otra aumenta. Sin embargo, la magnitud de la covarianza no es fácil de interpretar.

CORRELACIÓN Y COVARIANZA

Covarianza positiva



Covarianza negativa



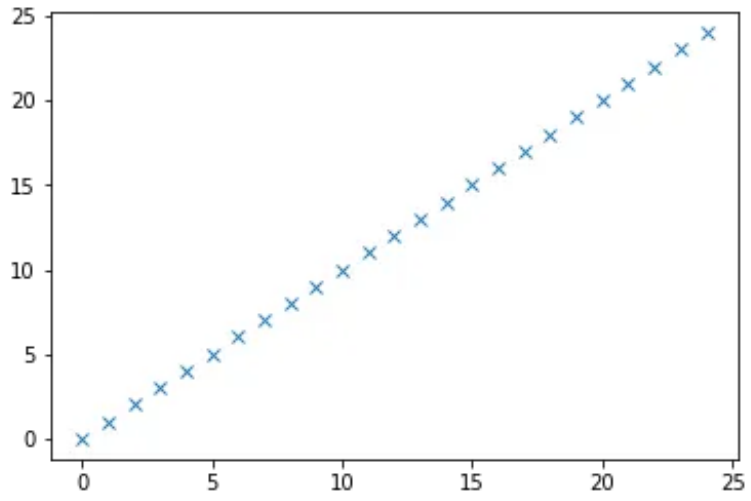
CORRELACIÓN Y COVARIANZA

Correlación:

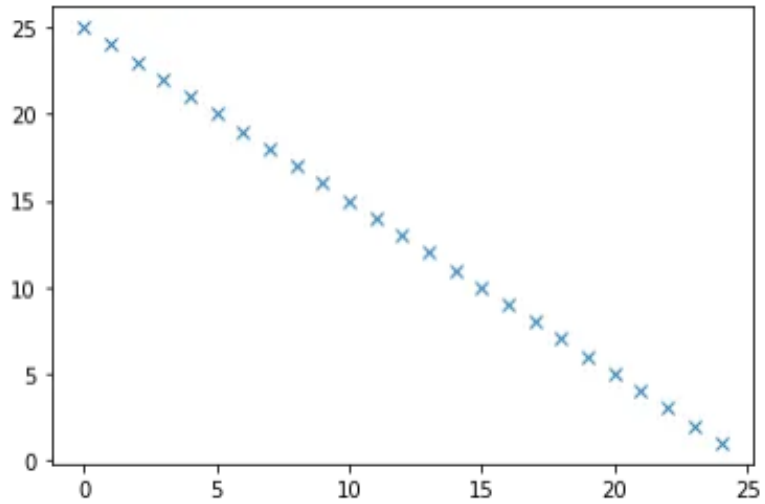
- La correlación es una versión estandarizada de la covarianza, que mide la fuerza y dirección de la relación lineal entre dos variables.
- La correlación *Corr*(*X*,*Y*)) se calcula dividiendo la covarianza entre el producto de las desviaciones estándar de las dos variables.
- Fórmula: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma X.\sigma Y}$ donde σX y σY son las desviaciones estándar de X e Y, respectivamente.
- La correlación varía entre -1 y 1. Un valor de 1 indica una correlación perfecta positiva, -1 indica una correlación perfecta negativa, y 0 indica ausencia de correlación.

CORRELACIÓN Y COVARIANZA

Correlación positiva



Correlación negativa



CORRELACIÓN Y COVARIANZA

Interpretación:

- Una correlación cercana a 1 sugiere una fuerte relación positiva, donde las variables tienden a aumentar juntas.
- Una correlación cercana a -1 indica una fuerte relación negativa, donde una variable tiende a disminuir cuando la otra aumenta.
- Una correlación cercana a 0 indica una débil o nula relación lineal.

CORRELACIÓN Y COVARIANZA

Consideraciones Importantes:

- La correlación no implica causalidad. Puede haber una relación aparente entre dos variables, pero esto no significa que una causa la otra.
- La correlación es sensible a valores atípicos, por lo que es importante examinar gráficamente la relación y considerar la posibilidad de transformar los datos si es necesario.
- Es crucial entender que la correlación evalúa solo relaciones lineales. Puede haber relaciones no lineales que no se reflejen en la correlación.

CORRELACIÓN Y COVARIANZA

Aplicaciones Prácticas:

- La correlación y la covarianza son fundamentales en estadística y análisis de datos para comprender la relación entre variables.
- Son útiles en la selección de variables para modelos predictivos, identificando qué variables están fuertemente relacionadas.
- Ayudan a entender la estructura de datos y a explorar patrones que puedan guiar análisis más profundos.
- En la toma de decisiones, la comprensión de la correlación puede ser esencial para prever el impacto de cambios en una variable sobre otra.

¿Preguntas?