Statistics is broadly divided into two main branches: **descriptive statistics** and **inferential statistics**. As a data science student, understanding this distinction is fundamental.

Descriptive Statistics

Descriptive statistics summarizes and describes the main features of a dataset. It aims to provide a clear and organized picture of the data, such as its central tendency, variability, and distribution. Think of it as summarizing a large book into a short, easy-to-read summary. It doesn't make any conclusions or predictions about a larger population; it simply describes the data you have.

Common examples of descriptive statistics include:

- **Measures of Central Tendency:** These are values that describe the center of a data set.
    - **Mean:** The average of all the values.
    - **Median:** The middle value when the data is ordered.
    - **Mode:** The value that appears most frequently.
- **Measures of Variability:** These describe how spread out the data is.
    - **Range:** The difference between the highest and lowest values.
    - **Standard Deviation:** A measure of how much the data deviates from the mean.

Standard deviation (σ or s) is a measure of the **amount of variation or dispersion** in a set of values. In simple terms, it tells you how spread out your data is from the **mean** (average).

- A **low standard deviation** means the data points are clustered closely around the mean, indicating they are very similar to each other.
- A **high standard deviation** means the data points are spread out over a wider range, indicating greater variability.

σ=N∑i=1N(xi−μ)2

-
    - **Variance:** The average of the squared differences from the mean.

variance is a measure of how spread out the data points are in a dataset. It is calculated as the average of the squared differences from the mean.

A **low variance** indicates that the data points are very close to the mean, meaning they are quite similar to each other. A **high variance** indicates that the data points are widely spread out from the mean, showing a high degree of variability.

Population Variance (σ2):

- $\sigma2=N\sum(xi-\mu)2$

Sample Variance (s2):

- $s2=n-1\sum(xi-\bar{x})2$

Descriptive statistics are often presented visually using tools like charts and graphs.

Inferential Statistics

Inferential statistics uses a sample of data to make predictions, generalizations, or inferences about a larger population. This is where you go beyond just describing the data you have and start to draw conclusions about a broader group. For example, if you wanted to know the average height of all adults in a country, you couldn't measure every single person. Instead, you would measure a random sample of people and use inferential statistics to estimate the average height of the entire population.

Key concepts in inferential statistics include:

- **Hypothesis Testing:** A formal procedure for investigating a claim about a population.
- **Regression Analysis:** A set of statistical processes for estimating the relationships among variables.
- **Confidence Intervals:** A range of values that likely contains the true value of a population parameter.

Inferential statistics allows you to test hypotheses and draw conclusions with a certain level of confidence or probability.

Covariance

Covariance measures the direction of the relationship between two variables. It tells you whether two variables tend to move in the same direction or in opposite directions.

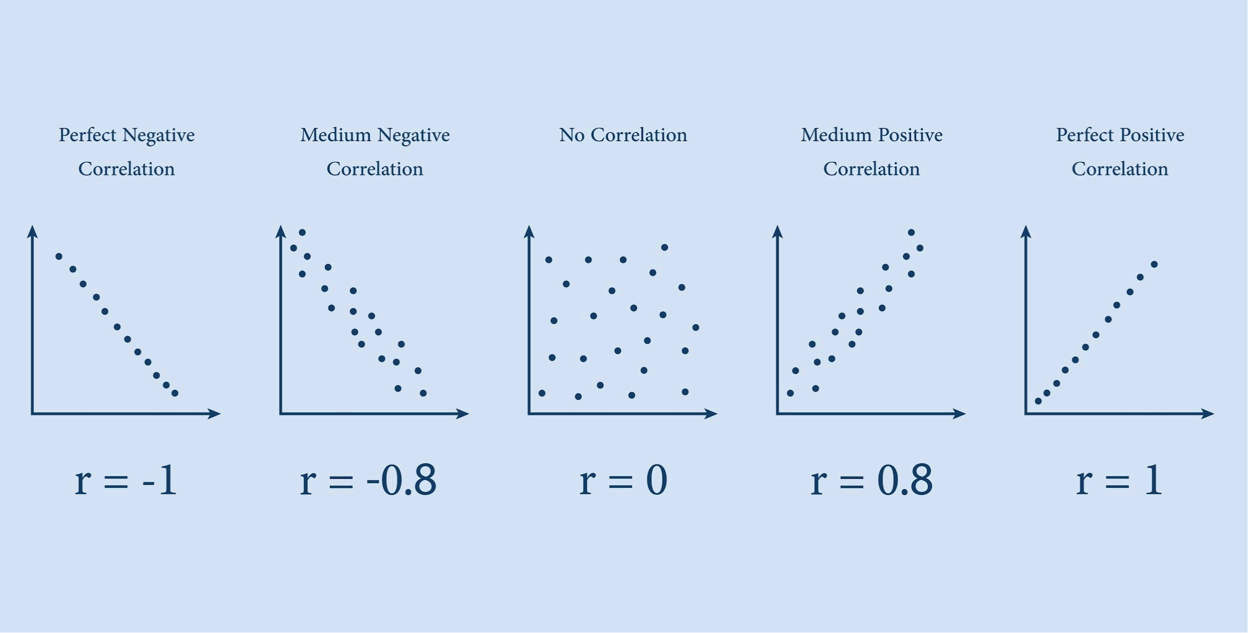
- **Direction only**:
  - A **positive covariance** means that as one variable increases, the other tends to increase as well.
  - A **negative covariance** means that as one variable increases, the other tends to decrease.
  - A **covariance close to zero** suggests no linear relationship.

- **Scale-dependent**: The value of covariance is not standardized. Its magnitude depends on the units of the variables being measured. For example, if you measure height in centimeters instead of meters, the covariance value will change dramatically, even though the relationship between the variables remains the same. This makes it difficult to compare the covariance between different pairs of variables.
- **Formula:**
  $\text{Cov}(X,Y)=n-1\sum(Xi-X^{-})(Yi-Y^{-})$

Correlation

Correlation measures both the **direction and strength** of the linear relationship between two variables. It is essentially a normalized version of covariance, which makes it a much more useful metric.

- **Standardized value**: Correlation is a unit-free value that always ranges from **-1 to +1**. This makes it easy to interpret and compare across different datasets.
  - A correlation of **+1** indicates a perfect positive linear relationship.
  - A correlation of **-1** indicates a perfect negative linear relationship.
  - A correlation of **0** indicates no linear relationship.



- **Scale-independent**: Because correlation is standardized, it is not affected by changes in the units of measurement. The correlation between height in inches and weight in pounds would be the same as the correlation between height in centimeters and weight in kilograms.

- **Formula:**
  $\text{Corr}(X,Y)=\sigma X\sigma Y\text{Cov}(X,Y)$
  where σX and σY are the standard deviations of X and Y, respectively.

While **correlation** and **causation** can exist at the same time, it's crucial to understand that they are not the same thing. Mistaking correlation for causation is a common fallacy, especially in data science.

Correlation

Correlation simply means that there is a relationship or an association between two variables. When one variable changes, the other variable tends to change as well. This relationship can be positive, negative, or non-existent. For example:

- **Positive Correlation:** As the number of ice cream sales increases, the number of sunburns also tends to increase.
- **Negative Correlation:** As the number of hours spent exercising increases, a person's weight tends to decrease.
- **No Correlation:** The number of times you've watched a movie has no relationship to the price of a gallon of milk.

A correlation coefficient, which ranges from -1 to 1, is used to measure the strength and direction of this relationship.

Causation

Causation, or cause and effect, means that one event is directly responsible for another event. A change in one variable is proven to cause a change in another. This is much more difficult to prove than correlation because it requires demonstrating that all other potential factors have been controlled for.

A classic example is a randomized controlled trial in medicine, where one group gets a new drug and a control group gets a placebo. If the group with the new drug shows a statistically significant improvement, it's possible to conclude that the drug *caused* the improvement.

The Key Difference

The main point is that **correlation does not imply causation**.

Going back to the ice cream and sunburn example, there is a strong correlation between the two, but eating ice cream does not cause sunburns. Both are caused by a third, unseen variable: hot, sunny weather. The weather causes people to buy more ice cream and also causes them to get more sunburned.

How to Prove Causation

To prove causation, you need to go beyond simply observing data. The gold standard for establishing a causal relationship is a **controlled experiment** or **randomized controlled trial**. These methods allow you to manipulate one variable (the independent variable) while keeping all other factors constant to see if it causes a change in another variable (the dependent variable).

https://g.co/gemini/share/4ca84c51563f