Statistics

- statistics is a branch of maths that helps in analysizing & predicting data.
- example usage:
  1. business: data analysis identifying customer behaviour & demand forecasting
  2. medical: identify effacacy of medicine by conducting clinical trials
  3. surveys, exit poll etc
- statistics helps in analysing past data & forcasting or estimating future possibilities
- Most of the ml algorithms fundamentals are based on statistics

Types of statistics

There are mainly 2 types of statistics:

1. Descriptive: summarizing data with the help of statistics tools like mean, median, 5 number summary etc
2. Inferential: to make predictions or forecasting or inferences about population from sample data

Sample & Population data

Lets consider an example where you have to find out average salary of India, You'll have 2 options:

1. Reach every human being in india, take summation of thier salaries & divide it with the population of india, Thats nearly impossible.
2. Create a subset of population of india which contains few people representing each category like rich, poor, salaried, unemployed, male, female etc & take thier salary & try to estimate/predict/infer the average salary of India.
3. Population is all human being in india & Sample is subset of the population which is the representative of each category.

Things to be careful about while creating sample:

- sample size should be enough to make correct inferences
- It should be random & not bias.. like you're only asking salary from people in your city & leaving everyone which will not give correct inference
- representative from each category

Types of data

Any column's value in a dataset can be of 2 types:

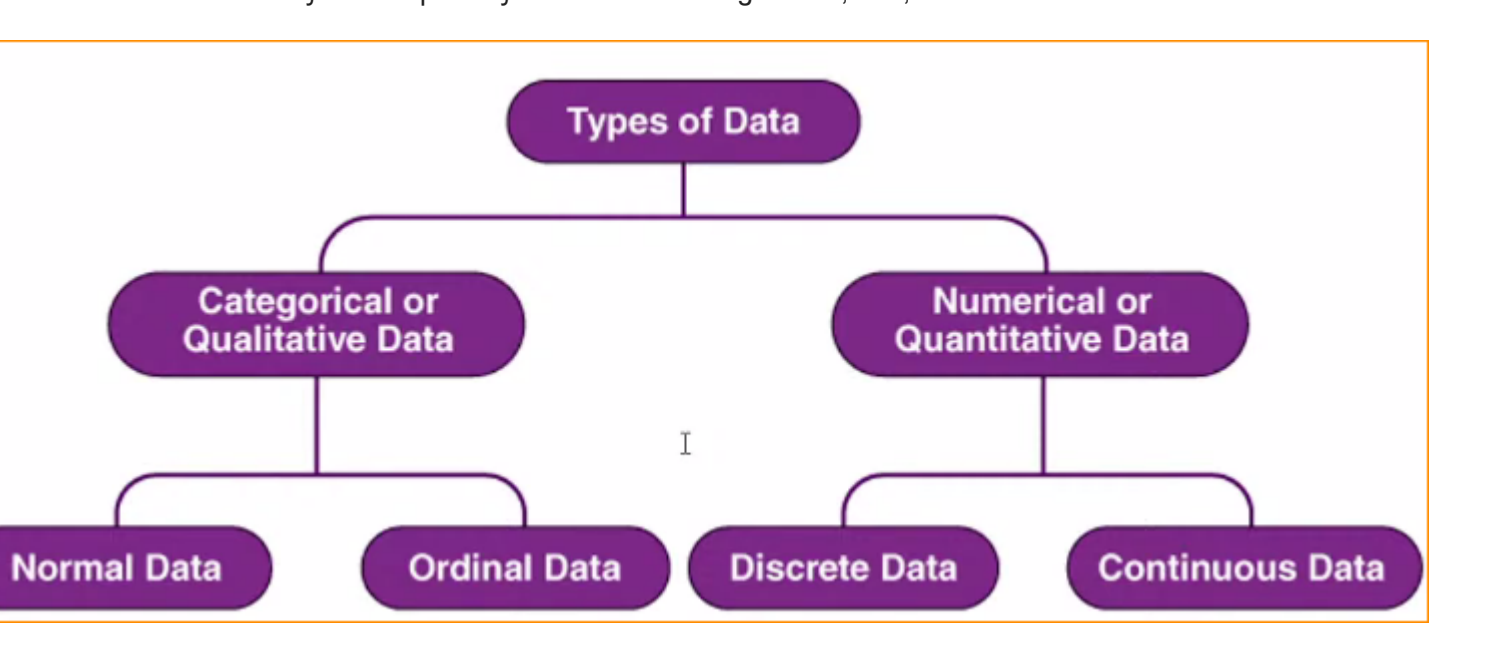
1. Categorical: like divisions first, second, third or cities kanour, delhi, gurgaon etc
2. Numerical: like age: 33,23,43 or weight 75,100,200 etc

Categorical data have 2 types:

1. Nominal: which doesnt have an order like cities kanpur, delhi which doesnt have an order
2. Ordinal: which have an order like division first, second, third..

Numerical Data have 2 types:

1. Discrete: only integer values like age, it will be 34,35,36 but it cant be 35.6, 36.7 etc
2. Continous: can have any value upto any decimal.. like weight 77.5,77.6,77.8 etc



Measure of central tendency

- Its a statistical measure that represents a central value of data.
- It provides a summary of data by identifying a single value which represents whole data.
- like mean, median, mode etc

mean

- sum of all values divided by no of values.
- 1,2,3,4,5,6 = 16/6 = 2.66 is the number which represents the centrality of this data
- population mean will different then sample mean
- BUT if your data have outliar then the value of mean will not be correct representative like avg salary of students in your class: 10000pm, 20000pm, 15000pm, 100000pm.. mean will be 36250.. which doesnt respresents the average salary of every student only bcoz of 1 outliar whose salary is 100000pm

Median

- middle value of the data when arranged in order like the same example:
- 10000pm, 20000pm, 15000pm, 100000pm, when we arrange this data in order it will become 10000pm, 15000pm, 20000pm, 100000pm
- since there are 4 numbers which is even then the middle value will be average of 2 middle numbers: (15000 + 20000)/2 = 17500pm
- this looks like correct representation of average salary
- see how median saves from outliars in data, either it will go in begining or at the end & will not have any impact

Mode

- Mode is the value which appears most frequently in a dataset
- like mode for 1,2,3,4,1,6,7,1,8,9 will be 1
- Its mostly used for categorical or discrete columns
- If all items appear only once then there will be no mode as data is uniform or unimodal

Weighted mean

- The weighted mean is a way to find the average of a set of numbers where some numbers contribute more to the final result than others.
- You multiply each number by its weight (importance or value) and then add up all these products.
- Finally, you divide this sum by the total weight. It gives more importance to numbers that have a greater weight or significance in the calculation.

Trimmed mean

- trimmed mean in very simple terms is a way to find the average (or mean) of a set of numbers by removing a certain percentage of the lowest and highest values.
- For example, if you have a list of numbers like this: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and you're asked to find the trimmed mean with 10% trimmed from both ends, you would remove the lowest and highest values.
- So, removing 10% of the lowest and highest values, you would remove 1 and 10. Then you calculate the mean of the remaining numbers (2, 3, 4, 5, 6, 7, 8, 9), which would be the trimmed mean.
- It's a way to get a more representative average when you have outliers or extreme values that might skew the result.

Measure of Dispersion

- Measure of dispersion describes the spread of data or variability of data
- for example mean of -5, 0, 5 & -100, 0, 100 will be 0.. so how will you diffrentiate?
- we can identify the spread of data to differentiate and those measure which describes the spred of data are measure of dispersion
- It describes how data is distributed around cetral tendency

Range

- range for -5, 0, 5 & -100, 0, 100 will be 10 & 200
- so with the help of range we can diffrentiate these 2 distributions whose mean is same
- substracting high - low of any stock is the range

Dataset X: 2, 4, 6, 8, 10 Dataset Y: -5, -3, -1, 1, 3, 5

Both datasets have the same mean:

For Dataset X: Mean = (2 + 4 + 6 + 8 + 10) / 5 = 30 / 5 = 6 For Dataset Y: Mean = (-5 + -3 + -1 + 1 + 3 + 5) / 6 = 0 / 6 = 0

Now, let's calculate the range:

For Dataset X: Range = 10 - 2 = 8 For Dataset Y: Range = 5 - (-5) = 10

Despite having the same mean, Dataset Y has a larger range than Dataset X. This is because the spread of values in Dataset Y is wider, ranging from -5 to 5, compared to the narrower spread of values in Dataset X, ranging from 2 to 10.

This example illustrates that the range is influenced by the spread of values in the dataset, regardless of whether the mean is the same. In this case, Dataset Y has a wider spread of values, resulting in a larger range, even though both datasets have the same mean.

Variance

- variance is the average of squared differences b/w each data point & mean

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

Where:

- $\bar{x}$ is the mean of the data points, given by: $\bar{x} = \frac{x_1+x_2+...+x_n}{n}$

- so for x1, x2, x3, variance will be: $\text{Variance} = \frac{1}{3}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\right]$

  Where $\bar{x}$ is the mean of the three data points given by: $\bar{x} = \frac{x_1+x_2+x_3}{3}$
- variance also describes the spread but its not exactly spread but its proportional to the spread
- Its describing how far is each data point from its mean
- squared is used "To Make All Values Positive"
- Instead of square, we can also take "mod" & thats called MAD: Mean Absolute Diviation which cant be used in inference like if you know sample mad then you cant infer population mad.
- MAD is less prone to ouliars unlike variance.. if there is a bigger number then taking squared mean will be even bigger number
- Sample & Population Variance is also different just like mean

whats the use of variance?

- The variance is a fundamental measure of variability or spread within a dataset. Its benefits include:
- Quantifying Spread: Variance provides a quantitative measure of how spread out the data points are around the mean. A higher variance indicates that the data points are more spread out, while a lower variance suggests that the data points are closer to the mean.
- Useful in Comparisons: Variance allows for comparisons between different datasets. It helps assess the degree of variability or dispersion in different sets of data. For example, you can compare the variability of sales figures across different regions or the variability of test scores among different groups of students.
- Basis for Further Analysis: Variance is a key component in many statistical analyses and techniques. For instance, it's used in calculating standard deviation, which is another measure of spread commonly used in statistics. Variance is also utilized in hypothesis testing, regression analysis, and other statistical modeling techniques.
- Identifying Outliers: High variance can indicate the presence of outliers or extreme values in the dataset. Outliers can significantly affect the overall distribution of the data and may warrant further investigation.
- Understanding Data Quality: Variance provides insights into the consistency and reliability of the data. A low variance suggests that the data points are relatively consistent and closely clustered around the mean, indicating higher data quality. On the other hand, a high variance may suggest greater variability and potential data quality issues.
- Overall, variance is a valuable statistical tool that helps researchers, analysts, and decision-makers understand and interpret the variability within their data, leading to better-informed decisions and conclusions.

Standard Diviation

- Std is a square root of variation
- The reason why it exists is that its unit is same as data unlike variance where its unit-squared but since we're taking square root of variance, it comes into same unit as data
- Its less prone to outliars bcoz of sqr root

whats the use of Std()?

- let's consider a real-life example involving exam scores:
- Imagine you are a teacher and you have two classes, Class A and Class B, each with 20 students. You want to compare the performance of the two classes on a recent math exam. Here's how standard deviation can help:
- Understanding Variability: After grading the exams, you calculate the average score for each class. Both Class A and Class B have an average score of 75 out of 100. However, to truly understand the performance, you also need to consider how spread out the scores are around the average.
- Comparing Spread: You calculate the standard deviation for the exam scores in each class. Let's say Class A has a standard deviation of 10 points, while Class B has a standard deviation of 5 points.
- Interpreting the Results:
- Class A: A standard deviation of 10 indicates that the scores in Class A are more spread out around the mean of 75. This suggests that there is more variability in performance among the students in Class A.
- Class B: With a standard deviation of 5, the scores in Class B are less spread out around the mean of 75. This indicates that the performance of students in Class B is more consistent or uniform compared to Class A.
- Decision Making: Based on this information, you might conclude that while both classes have the same average score, Class B performed more consistently overall. This could inform your teaching strategies, such as identifying areas where Class A needs more support or recognizing Class B's strengths.
- In this example, standard deviation helps you go beyond just comparing averages and provides insight into the variability or dispersion of the exam scores within each class. This understanding can guide decision-making and interventions to support student learning and improve outcomes.

Coefficient of variation

- The coefficient of variation (CV) is a statistical measure used to compare the variability of data sets with different means. It's particularly useful when comparing the variability of data sets that are measured in different units or have different scales.

$$CV = \left(\frac{\text{Standard Deviation}}{\text{Mean}}\right) \times 100\%$$

Or alternatively:

$$CV = \left(\frac{\text{Standard Deviation}}{\text{Average}}\right) \times 100\%$$

whats the use of CV?

- Imagine you're considering investing in two different portfolios, Portfolio A and Portfolio B, each managed by different investment firms. Both portfolios have provided an average annual return of 8% over the past 5 years.
- Now, let's look at the variability or risk associated with each portfolio using the coefficient of variation (CV).
- Portfolio A:
- Average annual return: 8%
- Standard deviation of annual returns: 4%
- Portfolio B:
- Average annual return: 8%
- Standard deviation of annual returns: 6%
- Using the formula for the coefficient of variation, we can calculate the CV for each portfolio:

  For Portfolio A: $CV_A = \left(\frac{4}{8}\right) \times 100\% = 50\%$

  For Portfolio B: $CV_B = \left(\frac{6}{8}\right) \times 100\% = 75\%$
- Interpreting the results:
- Portfolio A has a coefficient of variation of 50%, indicating that the variability (risk) of its returns is 50% of its mean return. This suggests that Portfolio A has relatively lower risk compared to its average return.
- Portfolio B has a coefficient of variation of 75%, indicating that the variability (risk) of its returns is 75% of its mean return. This suggests that Portfolio B has relatively higher risk compared to its average return.
- Based on this analysis, you might conclude that Portfolio A offers a more favorable risk-return profile compared to Portfolio B. Despite both portfolios having the same average return of 8%, Portfolio A has lower variability or risk associated with its returns, making it potentially a more attractive investment option for investors seeking lower risk.

Titanic Dataset Example:

- Let's assume we've already computed the mean and standard deviation for both age and fare data. Here's how we can interpret the results:
- If the coefficient of variation for age is low, it indicates that the ages of passengers are relatively close to the average age. This suggests that there's less variability in ages among passengers.
- If the coefficient of variation for fare is high, it suggests that the fares paid by passengers vary widely around the average fare. This indicates greater variability in fare prices.

Quantiles

- Quantiles are a way to divide a dataset into equal-sized groups or intervals. Imagine you have a bag of 100 candies, and you want to divide them equally among your friends. You can use quantiles to ensure each friend gets an equal share.
- Here's how you might use quantiles to divide the candies:
- Divide into Quantiles:
- You decide to divide the candies into four quantiles, which means you're splitting them into four equal-sized groups. You count the candies and find that you have 100 in total. Each quantile will contain 25 candies because 100/25 = 4
- Distribute the Candies:
- The first quantile will contain the first 25 candies in the bag. You give these candies to your first friend. The second quantile will contain the next 25 candies. You give these candies to your second friend. Similarly, you give the next 25 candies to your third friend and the final 25 candies to your fourth friend. Equal Distribution:
- Each friend receives an equal number of candies, ensuring fairness in the distribution. So, in simple terms, quantiles help you divide a dataset into equal-sized groups, allowing for fair and balanced comparisons. In the candy example, quantiles helped ensure that each friend received an equal share of candies from the bag.

Types of Quartiles

Quantiles can be divided into various types based on the number of groups they divide the dataset into:

- Quartiles: Quartiles divide the dataset into four equal-sized groups, representing the 25th, 50th (median), and 75th percentiles.
- Quintiles: Quintiles divide the dataset into five equal-sized groups, representing the 20th, 40th, 60th, and 80th percentiles in addition to the median.
- Deciles: Deciles divide the dataset into ten equal-sized groups, representing the 10th, 20th, 30th, ..., 90th percentiles.
- Percentiles: Percentiles divide the dataset into 100 equal-sized groups, representing every percentile from the 1st to the 99th, with the median representing the 50th percentile.

Things to remmember

1. data should be sorted from low to high
2. you're finding the location of an observation 25th quartile, 50th quartile
3. they're not actual value in data
4. all other types can be easily derived from percentiles

Percentile

- Percentile is a way to understand where a particular value falls within a dataset, expressed as a percentage.
- Imagine you're taking a test with 100 questions, and you score 80 out of 100. If someone tells you that your score is at the 80th percentile, it means that your score is equal to or better than 80% of the scores in the group of people who took the same test.
- In simpler terms, if you're at the 80th percentile, it means you did as well as or better than 80% of the people who took the test. Percentiles help you compare your performance to others in a way that's easy to understand.
- If you're at the 50th percentile, you're right in the middle—half of the people scored lower than you, and half scored higher.

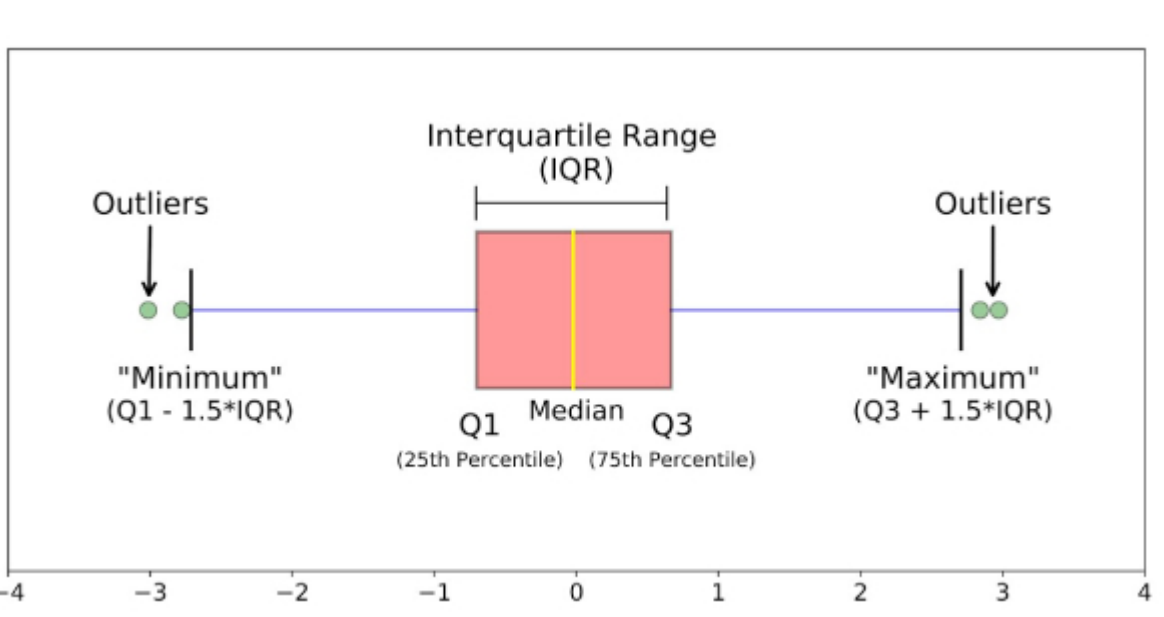
Percentage vs Percentile

- If your score is 90 percentage in exam, It means you scored 90 out of 100
- If your score is 90th percentile in exam, It means you did better then 90 students in exam if there were 100 students.. basically 90th percentile means 90% people score less then you.

- Let's say you have a dataset of exam scores and you want to find the 80th percentile score. $\text{Rank} = \frac{80}{100} \times (100+1) = \frac{80}{100} \times 101 = 80.8$
- First, sort the dataset in ascending order from lowest to highest score.
- Calculate the rank of the percentile you're interested in using the formula: $\text{Rank} = \frac{P}{100} \times (n+1)$
- where P is the percentile (in this case, 80), and n is the number of data points in the dataset.
- If the rank is a whole number (no decimal), take the score at that rank as the percentile value.
- If the rank has a decimal, round up to the nearest whole number and take the score at that rank as the percentile value.

5 Number Summary

- 5 number summary are quartiles which divides data into 4 equal parts
- It provides 5 numbers: min value, 25th percentile, 50th percentile, 75th percentile, max value
- Its used to create box plot



- IQR: Difference b/w 3rd & 1st quartile is "Inter Quartile Range", Box plot box represents IQR

Co-Variance

- To understand the type of linear relation b/w 2 numerical columns
- whether one increases with increase in another or descrease or remain constant
- lets imagine we want to understand the relation b/w studing hours & grades achieved..
- whether studying more will result in higher grades, lower grades or no change in grades
- if covariance is positive then it means studying more will result in higher grades & vice versa or no change in grades
- BUT there's a flaw.. we cant distinguish the magnitude of the relation whether its strongly corelated or weak!!

Corelation

- It quantifies the strength of linear relationship b/w 2 numerical column
- it measures the degree to which 2 variables are related & how they tend to change together
- Its measure using a tool called "Corelation Coefficient" which ranges from -1 to 1
- -1 represents perfect negative corelation
- +1 represents perfect positive corelation
- 0 represents no corelation

In [5]:

```python
import numpy as np

# Create two datasets with different scales
dataset1 = np.array([1, 2, 3, 4, 5])  # Small values
dataset2 = np.array([10, 20, 30, 40, 50])  # Large values

# Calculate covariance
covariance = np.cov(dataset1, dataset2)[0][1]

# Calculate correlation coefficient
correlation_coefficient = np.corrcoef(dataset1, dataset2)[0][1]

print("--- dataset1 & dataset2 ---")
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation_coefficient)


# -----------------------------------------------------------------------


# Create two datasets with different scales
dataset3 = 2*(np.array([1, 2, 3, 4, 5]))  # Small values
dataset4 = 2*(np.array([10, 20, 30, 40, 50]))  # Large values

# Calculate covariance
covariance = np.cov(dataset3, dataset4)[0][1]

# Calculate correlation coefficient
correlation_coefficient = np.corrcoef(dataset3, dataset4)[0][1]


print("--- dataset3 & dataset4 ---")
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation_coefficient)
```

```
--- dataset1 & dataset2 ---
Covariance: 25.0
Correlation Coefficient: 1.0
--- dataset3 & dataset4 ---
Covariance: 100.0
Correlation Coefficient: 1.0
```

Corelation doesnt imply Causation

- "Correlation doesn't imply causation" means that just because two things are related or connected in some way (correlation), it doesn't mean that one thing causes the other (causation).
- Here's a simple example:
- Higher Experienced Employee get Higher Salary.. Thats True but not in all cases.. !!
- Could be Company is giving more salary, Could be company getting exactly what they need in a candidate, Could be urgent hiring is the cause of higher salary.. there could be many reason & not solely experience hence "Corelation doesnt imply Causation"