Statistics Interview

Friday, September 15, 2023 11:23 AM

➔ What is Statistics?

Ans: Statistics basically Science of colleting, analysing and interpreting data even presenting data.

➔ What is the central limit theorem?

Ans: The Central Limit Theorem (CLT) is a fundamental concept in statistics that describes the behavior of the distribution of sample means.

➔ Mean vs Median ?

Ans: The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest.

Example:
Data = 11, 15, 78, 33, 5, 6, 8

Mean = $\bar{x} = \frac{\sum x}{n}$

$$\bar{x} = \frac{11+15+78+33+5+6+8}{7}$$

$$\bar{x} = \frac{156}{7}$$

$$\bar{x} = 22.28$$

Data = 11, 15, 78, 33, 5, 6, 8
We have to rearrange the data first
Data = 5, 6, 8, 11, 15, 33, 78
Median = 11

When you will find this type of data set:
Data = 11, 15, 78, 33, 5, 6, 8, 15
Rearrange = 5, 6, 8, 11, 15,15, 33, 78

= 11 + 15 / 2
= 26/2
=13
Median = 13

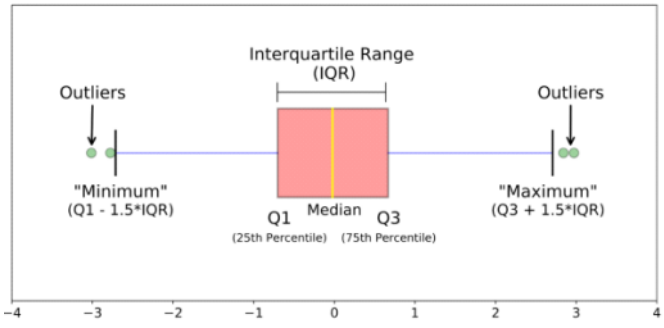
When you will find this type of data set:
Rearrange = 5, 6, 8, 11, 11, 15, 33, 78

At this point you have to take common values just one.

Median = 11

➔ What is a 5-Number Summary ?

Ans: The five-number summary is a set of descriptive statistics that provides information about a dataset



Box Plot

Given:

$$Q_1 = 42$$

$$Q_3 = 50$$

Determine the distance between the quartiles (IQR):

$$IQR = Q_3 - Q_1 = 50 - 42 = 8$$

The lower limit is then the first quartile decreased by 1.5 IQR's:

$$Q_1 - 1.5IQR = 42 - 1.5(8) = 30$$

The upper limit is then the third quartile increased by 1.5 IQR's:

$$Q_3 + 1.5IQR = 50 + 1.5(8) = 62$$

We then note that 65 is higher than the upper limit, which indicates that 65 can be considered to be an outlier.

➔ What is a hypothesis test? How is the statistical significance of an insight determined?

Ans: **Hypothesis testing** is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

Steps

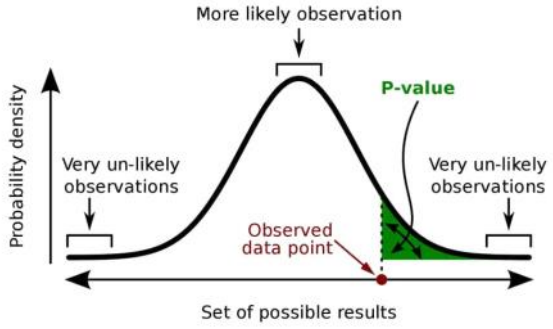
- Define Problem Statement
- Make Initial Hypotheses **(H0- Null Hypothesis)**
- Collect Evidences
- Reject H0 or accept **H1(Alternate Hypothesis)**

➔ What are the characteristics of large numbers in statistics?

Ans: The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as the sample become larger.

➔ What do you think of the phrase 'p-value'?

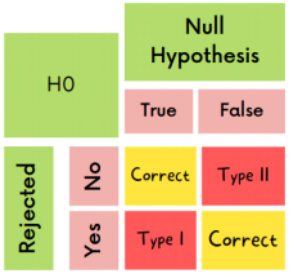
Ans: The phrase "p-value" is a fundamental concept in statistics and hypothesis testing, and it plays an important role in scientific research and decision-making.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

➔ What is the difference between an error of type I and an error of type II ?

Type I Error = Rejecting the null hypothesis when it's **True.**
Type II Error = **Not** Rejecting the null hypothesis when it's **False.**



➔ What are some examples of data sets with non-Gaussian distributions?

Ans: Examples of non-Gaussian distributions include the-

i. **Exponential distribution,**
ii. **Poisson distribution,**
iii. **Log-normal distribution,**
iv. **Weibull distribution,**
v. **Gamma distribution,**
vi. **Chi-square distribution.**

Each distribution has its own characteristics and applications in different fields.

➔ What is the difference between a sample and a population?

Ans: A population is an entire group that you want to draw conclusions about, A sample basically a subset of the population, it's a small portion of the population.

➔ What are the different kinds of variables or levels of measurement?

Quantitative variable: Quantitative variables represent numerical data. **(Numerical)**
Types: Numerical variables are two types

1. Continuous Variable **(Measurement)**
   Examples: height, weight, number of goals scored in a football match, age, length, time, temperature, exam score, etc.

1. Discrete Variable **(Counting or Counted)**
   Examples: Number of Students in a Class, Number of Cars in a Parking Lot, Number of Books on a Shelf, Number of Siblings, Number of Pets, Number of Apples in a Basket, Number of Coins in a Wallet, etc.

Qualitative variable: A qualitative variables represent non-numerical data. **(Categorical, Words)**
Types: Non-numerical variables are two types

1. Ordinal Variable **(Logical Order)**
   Examples: Educational level, Customer Satisfaction Rating, Economic Status, Health Status, Agreement level, Socioeconomic(Lower class, Middle Class and Upper Class) and so on.

2. Nominal Variable **(This one or That one)**
   Examples: Gender, Eye Color, County of Birth, Favorite Color, Blood Type, Marital Status, and so on.

➔ What is the difference between Descriptive and Inferential Statistics?

Inferential Statistics

1. Inferential statistics is a branch of statistics that involves making inferences and drawing conclusions about a population based on a sample of data from that population.

Descriptive Statistics

1. Descriptive statistics is a branch of statistics that involves summarizing, organizing, and presenting data in a meaningful way.

➔ What is sampling and Its Types?

Ans: Sampling allows researchers to use a small group from a larger population to make observations and determinations. Types of sampling include **random sampling, block sampling, judgment sampling, and systematic sampling.** Researchers should be aware of sampling errors, which may be the result of random sampling or bias.

➔ How does the central limit theorem work?

Ans: The central limit theorem says that **the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.** Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

➔ How do you define exploratory data analysis?

Ans: Exploratory data analysis (EDA) is **used by data scientists to analyse and investigate data sets and summarize their main characteristics, Data Cleaning,** often employing data visualization methods.
overall it's help scientist to gain deeper understanding of the data they are working and make the modelling and decision making.

➔ In what situation would the median be a more suitable measure compared to the mean?

Ans: When the data are skewed, the median is more useful because the mean will be distorted by outliers." This answer captures the conventional wisdom I learned in statistics courses.

➔ How might root cause analysis be applied to a real-life situation?

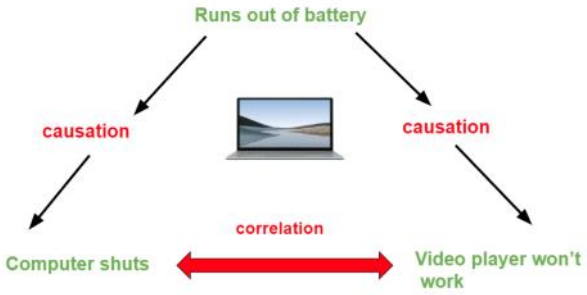
Ans: We can use RCA to modify core processes and system issues in a way that prevents future problems. Instead of just treating the symptoms of a football player's concussion, for example, root cause analysis might suggest wearing a helmet to reduce the risk of future concussions.

➔ What does standard deviation mean?

Ans: In statistics, the standard deviation is **a measure of the amount of variation or dispersion of a set of values**.

➔ Causation vs Correlation ?

Ans: Causation indicates that one event is the result of the occurrence of the other event and Correlation measures the direction and strength of the relationship between two variables.



➔ What are the characteristics of a bell-curve distribution?

Ans: A bell curve distribution, also known as a normal distribution or Gaussian distribution, is a probability distribution that has several distinctive characteristics:

- Symmetry
- Single Peak
- Tails
- Mean, Median, and Mode are the Same

➔ What is your definition of skewness?

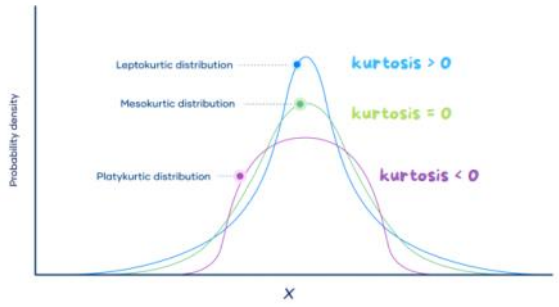
Ans: Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

$$g_1 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3} \qquad \text{Galton skewness} = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3}$$

➔ How do you define kurtosis?

Ans: Kurtosis is a measure of whether the data are heavy tailed or light-tailed relative to a normal distribution.



$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} \quad \text{OR} \quad \text{Kurt} = \frac{\mu_4}{\sigma^4}$$

➔ Left-skewed and right-skewed distributions exist, what are they?

Ans: Left-skewed and right-skewed distributions are specific types of probability distributions that exhibit asymmetry in their shapes. They are also known as negatively skewed and positively skewed distributions, respectively.

➔ What is the relationship between standard error and the margin of error?

Ans: The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.

➔ What does a degree of freedom (DF) represent in statistics?

Ans: In statistics, the term "degrees of freedom" (DF) represents the number of values in the final calculation of a statistic that are free to vary. It is a fundamental concept in hypothesis testing, confidence interval estimation, and the interpretation of various statistical tests. The specific interpretation of degrees of freedom varies depending on the context, but the general idea is that it reflects the amount of variability in a statistical calculation while accounting for constraints imposed by the data.

➔ Is there any significance to outliers in statistics?

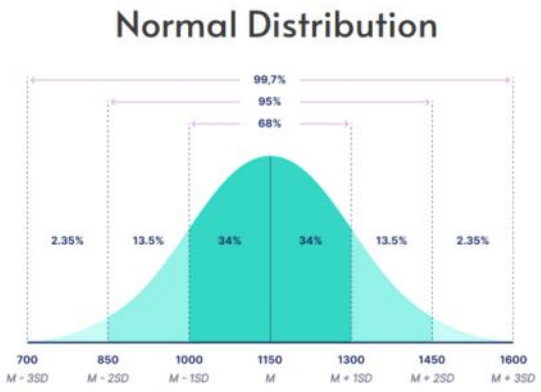
Ans: Yes, outliers in statistics can be significant because they may indicate data anomalies, affect statistical measures, and provide valuable insights or signal important events. They can impact the accuracy and validity of statistical analyses and often require special attention and treatment in data analysis.

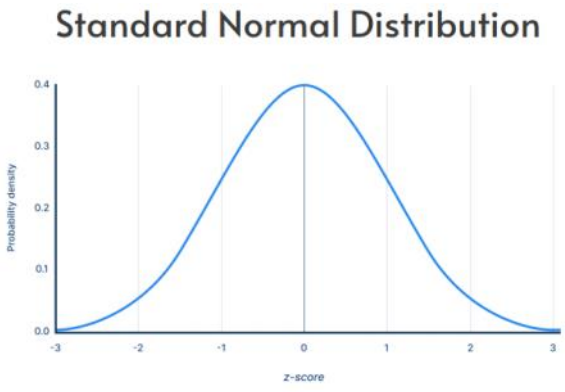
➔ What is the meaning of Central Tendency?

Ans: Central tendency refers to the statistical measure that represents the center or typical value of a dataset. It helps describe where most data points cluster and includes measures like the mean, median, and mode.

➔ How do you define Normal Distribution?

Ans: A normal distribution, also known as a Gaussian distribution, is a symmetric probability distribution characterized by a bell-shaped curve. In a normal distribution, data is centered around a mean, and its variability is described by a standard deviation. Many natural phenomena and random variables in statistics approximately follow this distribution.





$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$Z = \frac{(X - \mu)}{\sigma}$$

➔ How do you define empirical rule?

Ans: The empirical rule, also known as the 68-95-99.7 rule, is a statistical guideline that applies to approximately normal distributions. It states that:

- About 68% of the data falls within one standard deviation of the mean.
- Approximately 95% of the data falls within two standard deviations of the mean.
- Roughly 99.7% of the data falls within three standard deviations of the mean.

The empirical rule provides a quick way to understand the spread and distribution of data when dealing with normally distributed datasets.

➔ What Is the Confidence Interval?

Ans: A confidence interval is a statistical range or interval that estimates the range within which a population parameter, such as a mean or proportion, is likely to fall, with a specified level of confidence. It provides a measure of the uncertainty associated with statistical estimation. Common confidence levels are 95% and 99%, but other levels can be chosen.

➔ Describe the Difference Between Correlation and Autocorrelation?

Ans: Correlation measures the linear relationship between two different variables in a dataset. It quantifies how changes in one variable correspond to changes in another variable, typically expressed as a correlation coefficient (e.g., Pearson correlation).

Autocorrelation, on the other hand, measures the linear relationship between values of the same variable at different time points in a time series. It quantifies how a variable's past values relate to its current or future values, often used in time series analysis. Autocorrelation is also known as serial correlation or lagged correlation.

➔ When should you use a t-test vs a z-test?

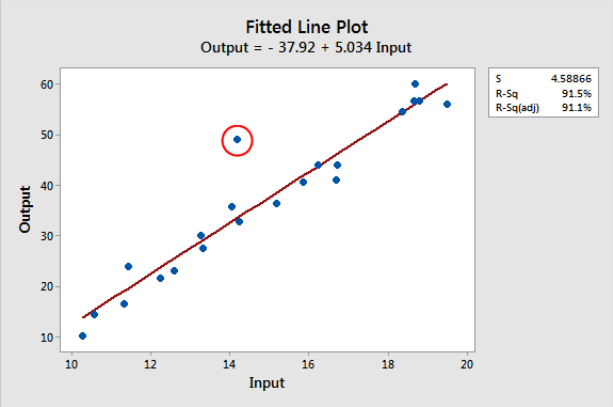
Ans: You should use a t-test when:
- The sample size is small (typically $n < 30$).
- The population standard deviation is unknown and has to be estimated from the sample data.

You should use a z-test when:
- The sample size is large (typically $n \geq 30$).
- The population standard deviation is known or when you are comparing sample means to a known population mean.

➔ What is an outlier? How can outliers be determined in a dataset?

Ans: An outlier is an observation or data point that significantly differs from the majority of the data in a dataset. Outliers can be

determined in a dataset using statistical techniques such as the Interquartile Range (IQR) method, z-scores, or visualization methods like box plots and scatter plots. Outliers are data points that fall below the lower bound or above the upper bound defined by these methods.

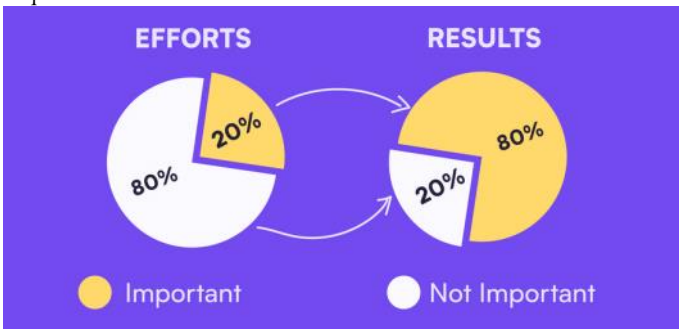


➔ What is the meaning of KPI in statistics?

Ans: KPI stands for "Key Performance Indicator" in statistics. It is a measurable metric used to evaluate the performance or effectiveness of a process, system, or organization. KPIs are essential for monitoring progress, making informed decisions, and achieving specific goals.

➔ What is the Pareto principle?

Ans: The Pareto Principle, also known as the 80/20 rule, states that roughly 80% of the effects come from 20% of the causes. It's a rule of thumb often applied in various fields, suggesting that a minority of inputs or factors often have a disproportionately large impact on outcomes.



➔ What are quantitative data and qualitative data?

Quantitative variable: Quantitative variables represent numerical data. **(Numerical)**

Types: Numerical variables are two types

1. Continuous Variable **(Measurement)**
   Examples: height, weight, number of goals scored in a football match, age, length, time, temperature, exam score, etc.

1. Discrete Variable **(Counting or Counted)**
   Examples: Number of Students in a Class, Number of Cars in a Parking Lot, Number of Books on a Shelf, Number of Siblings, Number of Pets, Number of Apples in a Basket, Number of Coins in a Wallet, etc.

Qualitative variable: A qualitative variables represent non-numerical data. **(Categorical, Words)**

Types: Non-numerical variables are two types

1. Ordinal Variable **(Logical Order)**
   Examples: Educational level, Customer Satisfaction Rating, Economic Status, Health Status, Agreement level, Socioeconomic(Lower class, Middle Class and Upper Class) and so on.

2. Nominal Variable **(This one or That one)**
   Examples: Gender, Eye Color, County of Birth, Favourite Color, Blood Type, Marital Status, and so on.

➔ What is Bessel's correction?

Ans: Bessel's correction is a statistical adjustment made to sample variance and sample standard deviation calculations. It is used to provide a more accurate estimate of the population variance and standard deviation when working with a sample rather than an entire population. Bessel's correction involves dividing the sum of squared differences by (n-1) instead of n, where n is the sample size. This adjustment accounts for the fact that sample data tends to underestimate the population variability, especially for smaller sample sizes.

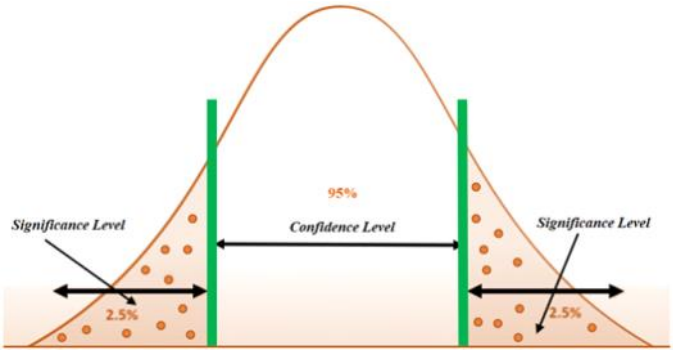
➔ What is the relationship between the confidence level and the significance level in statistics?

Ans: The confidence level and the significance level in statistics are complementary and inversely related.

- Confidence Level (e.g., 95% confidence): It represents the probability that the confidence interval contains the true population parameter (e.g., mean or proportion). Higher confidence levels result in wider intervals.

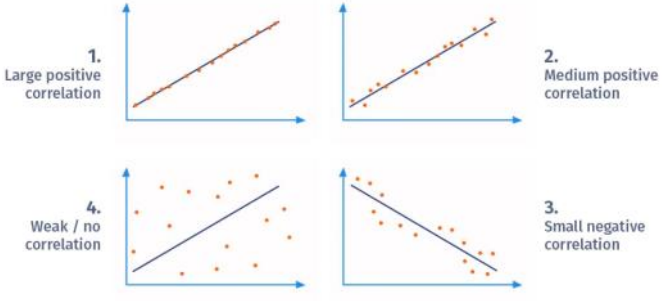
- Significance Level (e.g., 5% significance or alpha): It represents the probability of making a Type I error (incorrectly rejecting a true null hypothesis) in hypothesis testing. It is typically set as \(1 - \text{confidence level}\).

In essence, as you increase the confidence level, you decrease the significance level, and vice versa. The two levels together encompass the entire probability space, with the sum of their probabilities equalling 100%.



➔ What types of variables are used for Pearson's correlation coefficient?

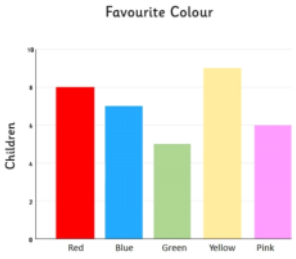
Ans: Pearson's correlation coefficient (also known as Pearson's r) is used to measure the strength and direction of a linear relationship between two continuous variables. Therefore, it is suitable for analysing the correlation between two quantitative or numerical variables. These variables should have numerical values and a relatively linear relationship for Pearson's correlation to be appropriate.



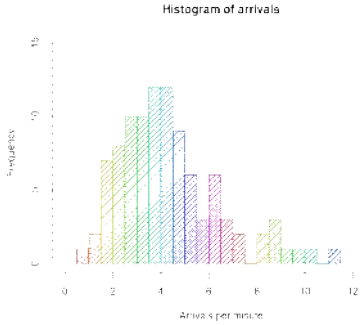
➔ What are the Graphs we have and their uses ?

Ans: There are many types of graphs in data visualization, each with its own use:

1. **Bar Chart:** Displays categorical data with rectangular bars. Useful for comparing discrete categories.

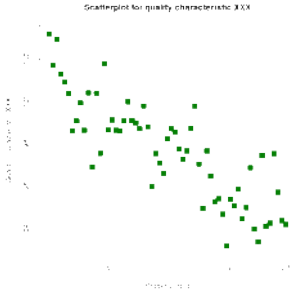


2. **Histogram:** Represents the distribution of continuous data by dividing it into bins. Useful for visualizing data frequency.

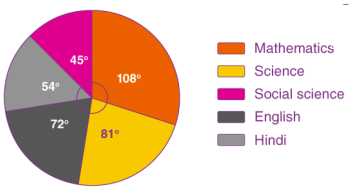


3. **Line Chart:** Connects data points with lines. Useful for showing trends over time or ordered categories.

4. **Scatter Plot:** Displays individual data points on a two-dimensional plane. Useful for visualizing relationships between two continuous variables.

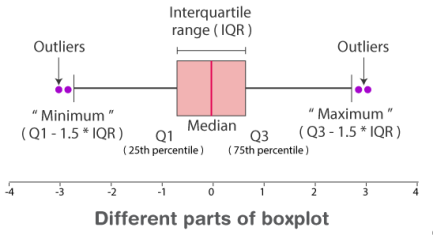


5. **Pie Chart:** Shows the parts of a whole as slices of a circle. Useful for displaying parts of a total percentage.

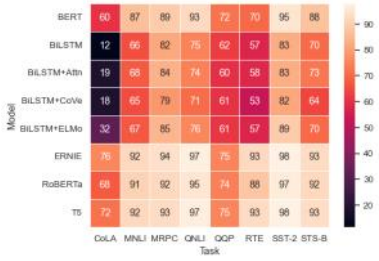


6. **Box Plot (Box-and-Whisker Plot):** Summarizes the distribution of data using quartiles. Useful for identifying outliers and

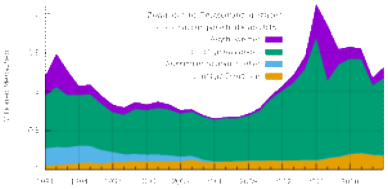
comparing data distributions.



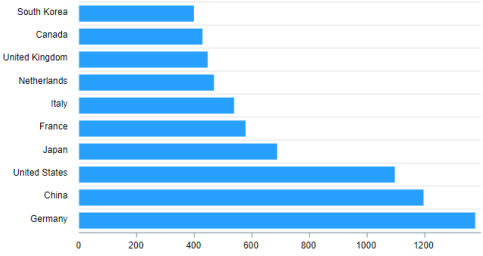
7. **Heatmap:** Uses color to represent data values in a matrix or grid. Useful for visualizing relationships in large datasets.



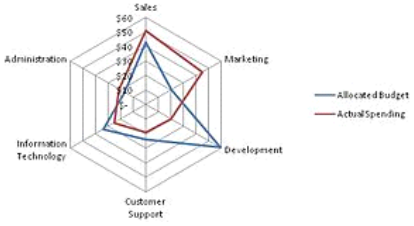
8. **Area Chart:** Similar to a line chart but with the area under the lines filled. Useful for showing cumulative data over time.



9. **Bar (Horizontal) Chart:** A horizontal version of the bar chart. Useful for displaying data labels with long names.

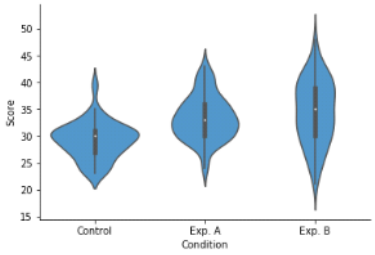


10. **Radar Chart:** Displays multivariate data on a circular grid. Useful for comparing multiple variables across categories.

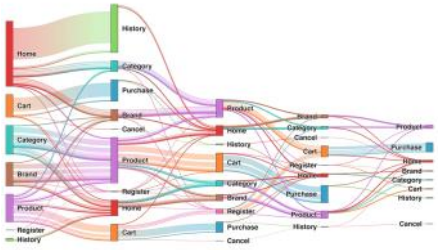


11. **Bubble Chart:** Extends the scatter plot by adding a third dimension, represented by the size of bubbles. Useful for displaying three-variable relationships.

12. **Violin Plot:** Combines a box plot and a kernel density plot to show the distribution of data. Useful for displaying data density.



13. **Sankey Diagram:** Visualizes the flow of data or resources between multiple entities or stages. Useful for understanding processes or networks.

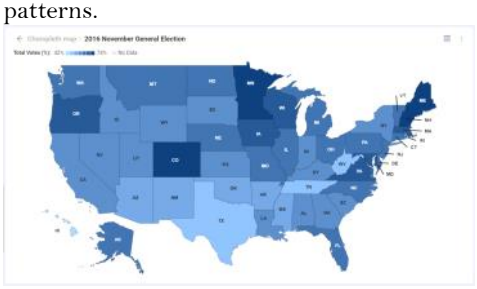


14. **Tree Diagram:** Represents hierarchical data structures like organizational charts or family trees.

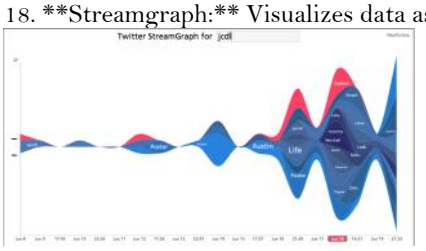
15. **Word Cloud:** Visualizes word frequency in text data, with more frequent words appearing larger. Useful for text analysis.

16. **Network Graph:** Displays relationships between interconnected nodes. Useful for social network analysis and complex systems.

17. **Choropleth Map:** Represents data values for geographic regions using color shading. Useful for displaying regional patterns.



18. **Streamgraph:** Visualizes data as stacked areas, useful for showing changes over time while preserving the total value.



The choice of graph depends on the type of data you have and the specific insights you want to convey. Different graphs are suitable for different purposes and data characteristics.

➔ Null Hypothesis vs Alternative Hypothesis ?

Ans: The null hypothesis (H0) is a statement of no effect or no difference, while the alternative hypothesis (H1 or Ha) is a statement that suggests there is an effect or a difference in a scientific experiment or study. Researchers typically test the null hypothesis to determine if there is enough evidence to reject it in favour of the alternative hypothesis.

➔ What is the benefit of using box plots?

Ans: The benefit of using box plots is that they provide a visual summary of the distribution of data, showing key statistics such as the median, quartiles, and potential outliers. This allows for quick and easy identification of data characteristics, making it useful for data exploration, comparison between groups, and outlier detection in a dataset.

➔ What is the relationship between mean and median in normal distribution?

Ans: In a normal distribution:

- The mean (average) is equal to the median.
- Both the mean and median are located at the center of the distribution.
- The normal distribution is symmetric, so the values are balanced around the mean/median.

➔ **How to convert normal distribution to standard normal distribution?**

Ans:

To convert a normal distribution to a standard normal distribution:

1. Calculate the z-score for each data point in the original distribution using the formula: $z = \frac{X-\mu}{\sigma}$, where X is the data point, μ is the mean, and σ is the standard deviation of the original distribution.
2. The resulting z-scores represent the values in the standard normal distribution, which has a mean (μ) of 0 and a standard deviation (σ) of 1.

➔ **What is the relationship between standard error and margin of error?**

Ans: The margin of error is calculated using the standard error. The standard error measures the variability or uncertainty of a sample statistic (e.g., the mean), while the margin of error quantifies the range within which the true population parameter is likely to fall with a certain level of confidence based on that standard error.

Sheikh Rasel Ahmed
Generative Artificial Intelligence
Follow Me On:
LinkedIn - https://www.linkedin.com/in/shekhnirob1
GitHub - https://github.com/Rasel1435
leetcode - https://leetcode.com/shekh_rasel
Kaggle - https://www.kaggle.com/sheikhraselahmed
YouTube - https://www.youtube.com/@codewithsheikhrasel
Facebook - https://www.facebook.com/rasel1435
Instagram - https://www.instagram.com/shekh_nirob