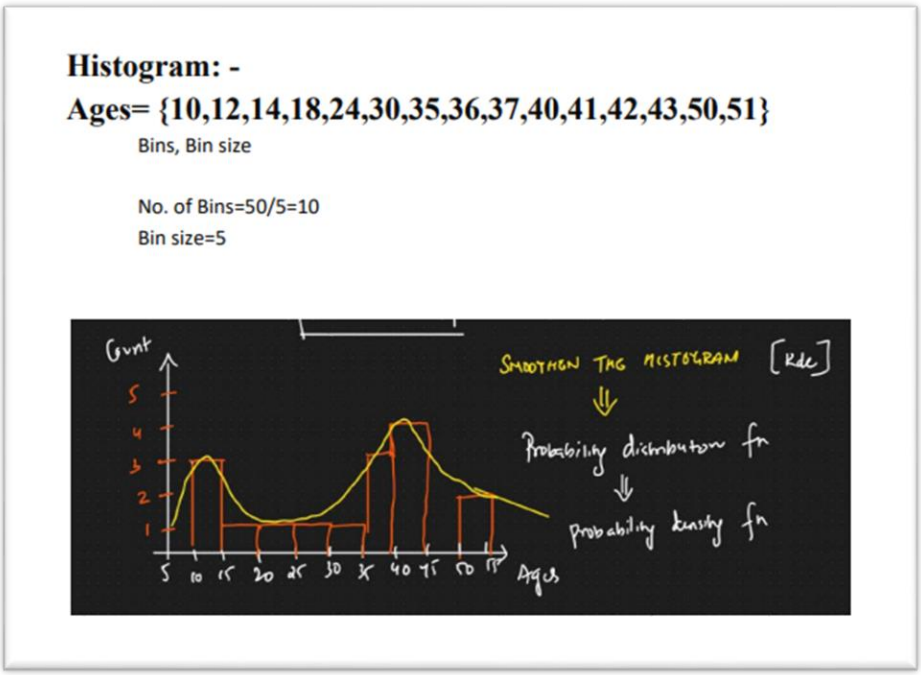
Graphical Techniques in Statistics

- Graphical techniques are powerful tools in statistics for visually representing data, exploring patterns, and communicating findings.
- They provide a visual summary of data distributions, relationships, and trends, making it easier for researchers and analysts to interpret and understand complex information.
- Here are some common graphical techniques used in statistics:

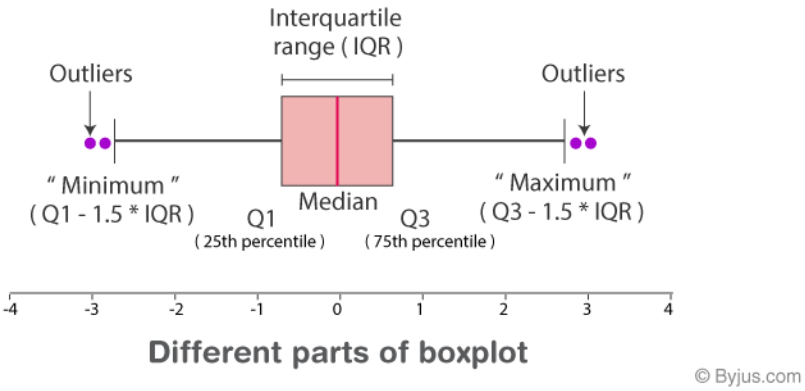
Histograms:

- **Purpose:** Display the distribution of a continuous variable.
- **How it works:** Data is divided into intervals (bins), and the frequency or proportion of observations in each bin is represented by bars.



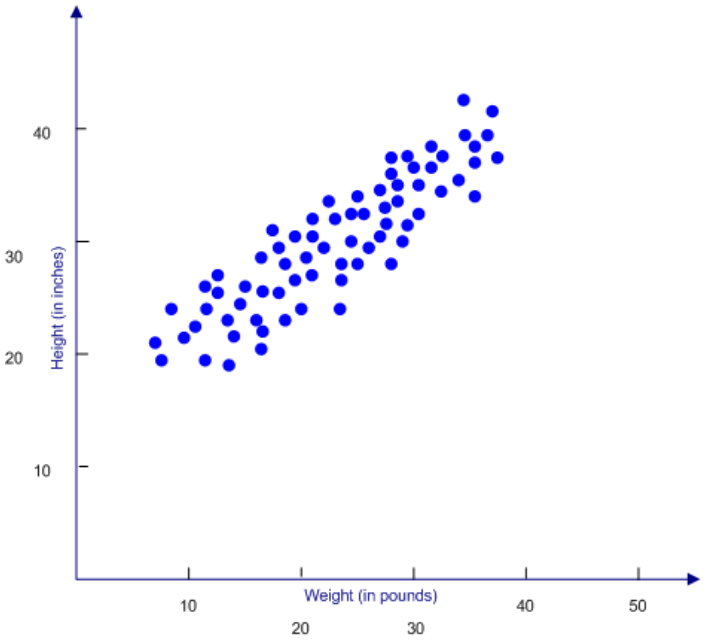
Box Plots (Box-and-Whisker Plots):

- **Purpose:** Show the distribution of a dataset and identify outliers.
- **How it works:** A box is drawn to represent the interquartile range (IQR), and whiskers extend to the minimum and maximum values. Outliers are often marked individually.



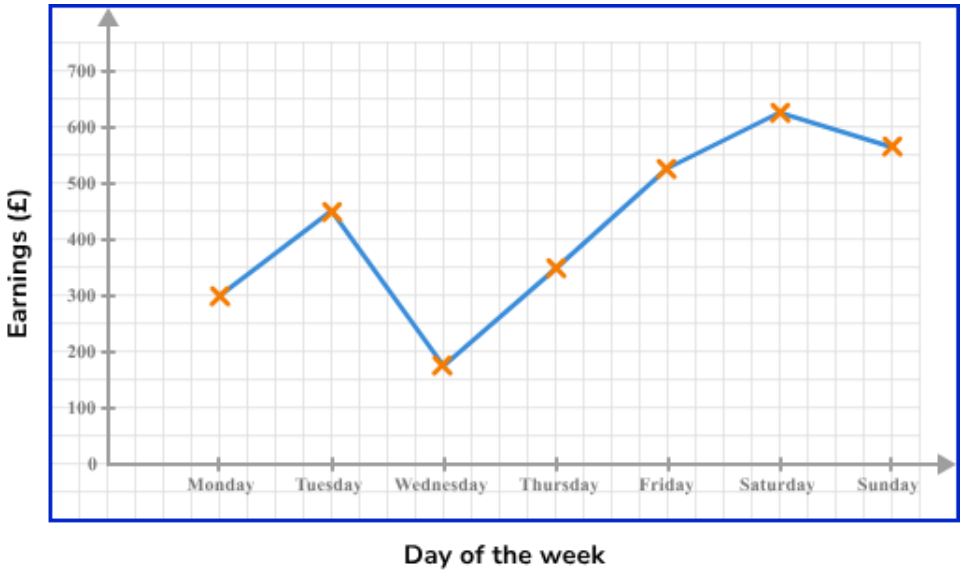
Scatter Plots:

- **Purpose:** Visualize the relationship between two continuous variables.
- **How it works:** Each data point is represented by a dot on the graph, with one variable on the x-axis and the other on the y-axis.



Line Charts:

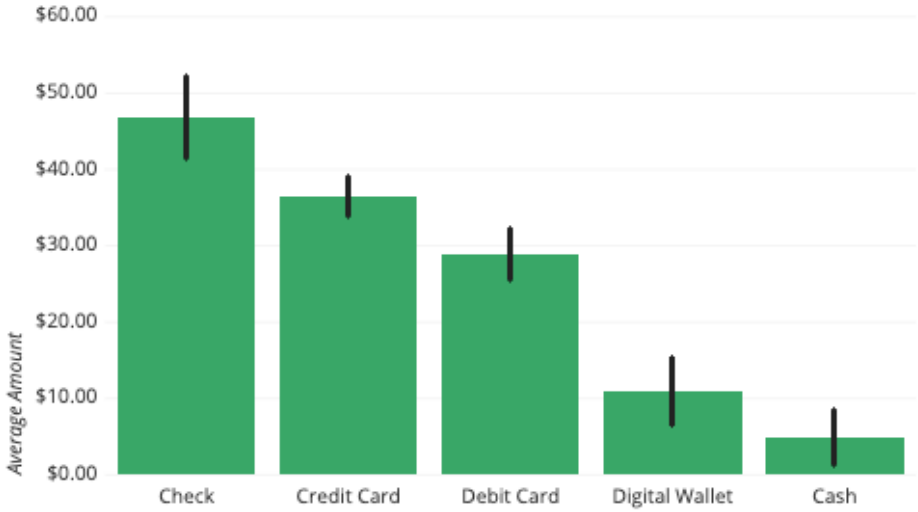
- **Purpose:** Display trends over time or across ordered categories.
- **How it works:** Data points relate to lines to show the progression of values.



Bar Charts:

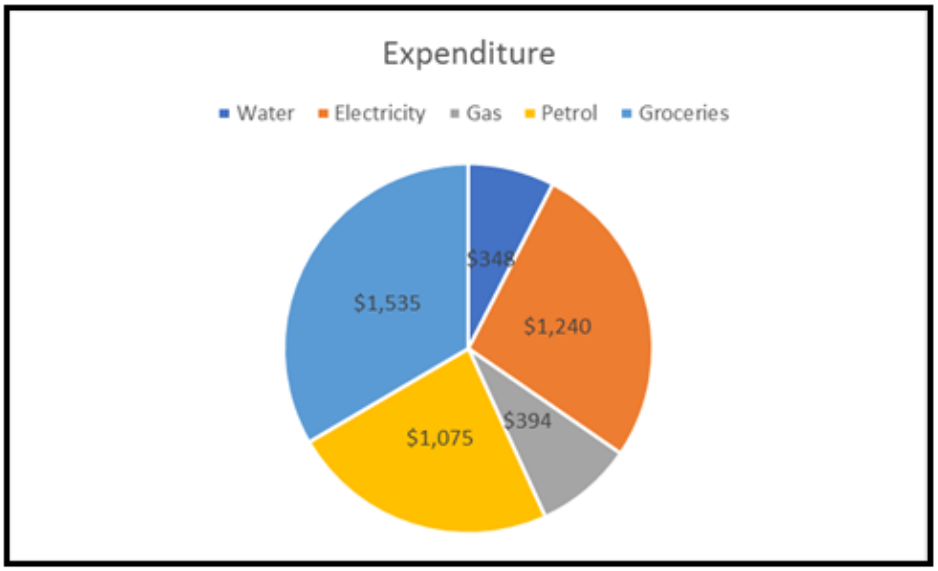
- **Purpose:** Compare the sizes of different categories or groups.

- **How it works:** Bars of equal width represent the values of each category or group.



Pie Charts:

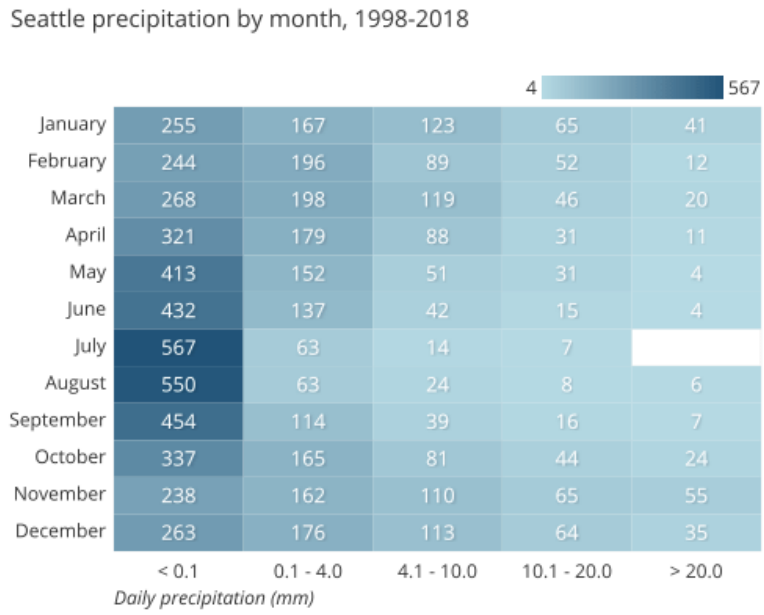
- **Purpose:** Show the proportion of parts to a whole.
- **How it works:** Slices of a circle represent different categories, and the size of each slice corresponds to the proportion of the whole.



Heatmaps:

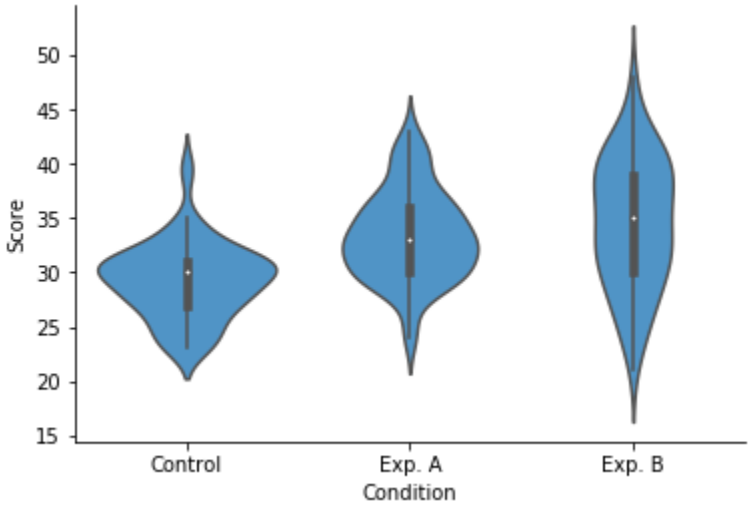
- **Purpose:** Visualize the magnitude of a phenomenon in a matrix format.
- **How it works:** Colors or shades represent the values in a matrix, helping to identify patterns and trends.

Eternaltek



Violin Plots:

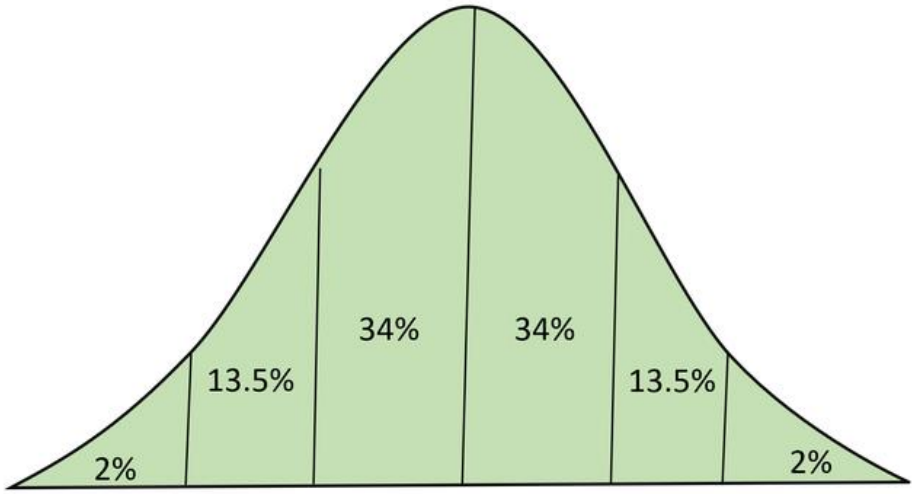
- **Purpose:** Combine aspects of box plots and kernel density plots to show the distribution of a variable.
- **How it works:** The shape of the plot resembles a violin, with wider sections indicating higher density.



QQ Plots (Quantile-Quantile Plots):

- **Purpose:** Assess the normality of a distribution.

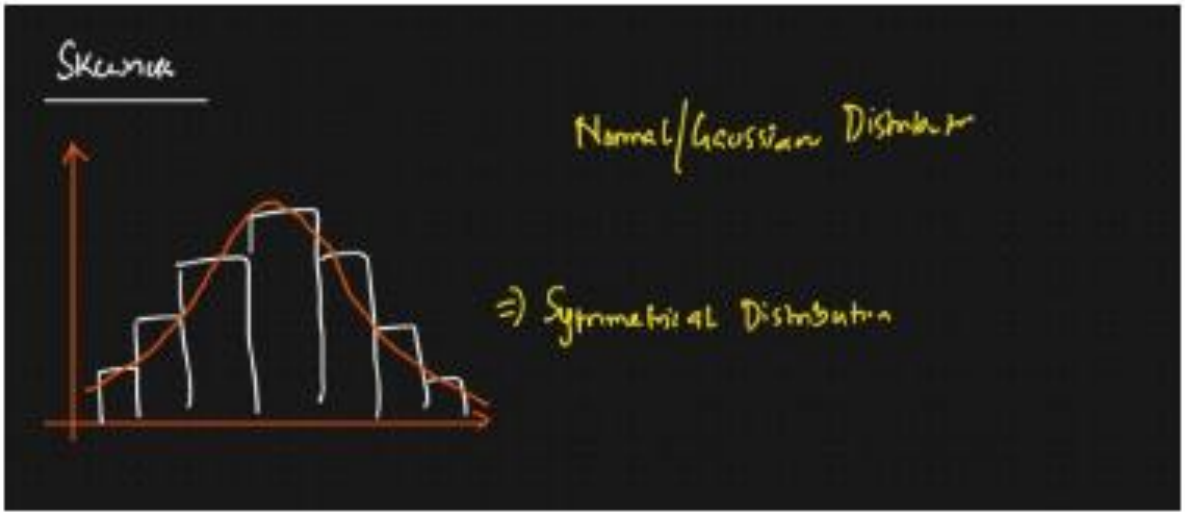
- **How it works:** Points are plotted against theoretical quantiles of a standard normal distribution. A straight line indicates normality.



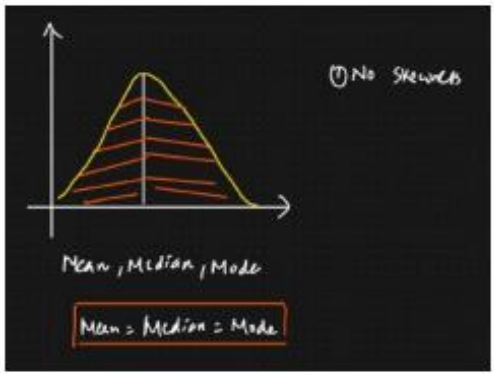
Skewness

- Skewness can be defined as a statistical measure that describes the lack of symmetry or asymmetry in the probability distribution of a dataset.
- It quantifies the degree to which the data deviates from a perfectly symmetrical distribution, such as a normal (bell-shaped) distribution.
- Skewness is a valuable statistical term because it provides insight into the shape and nature of a dataset's distribution.

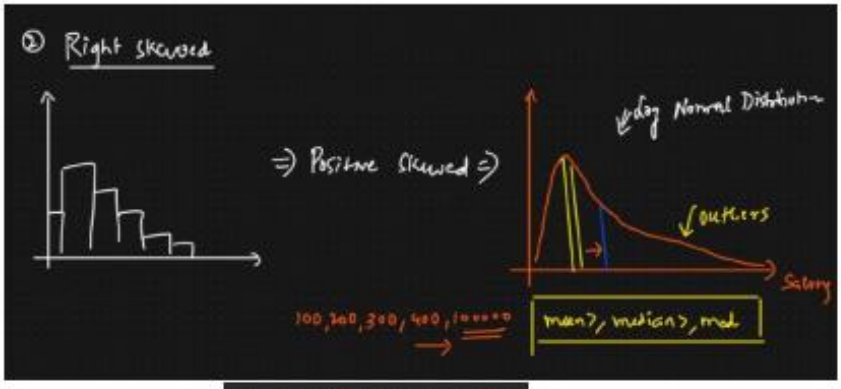
Eternaltek


Skewness
Normal/Gaussian Distribution
⇒ Symmetrical Distribution

A. No Skewed: -



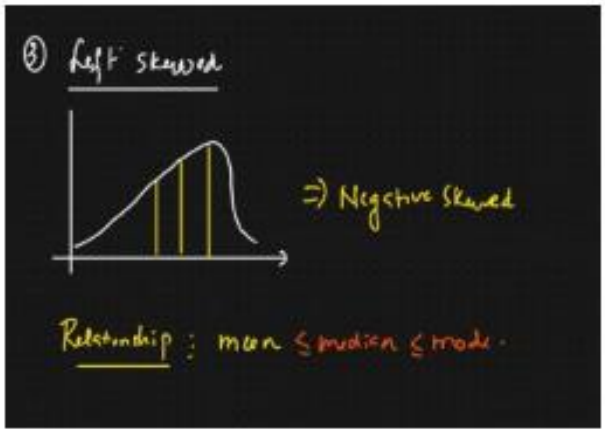
B. Right Skewed: -



Mean > Median > Mode

C. Left Skewed: -

Mean < Median < Mode



sampling Techniques:

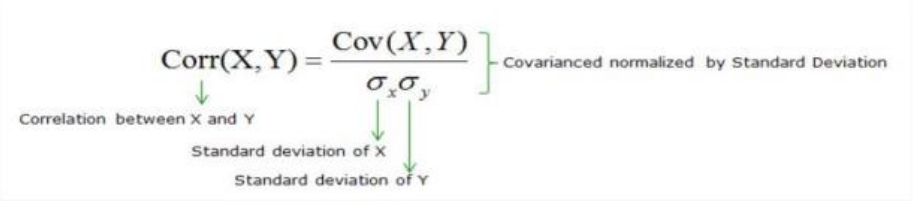
- **Simple Random Sampling:**
  - **Description:** Every individual or element in the population has an equal chance of being selected.
  - **Procedure:** Randomly select individuals from the population without any specific pattern or structure.
- **Stratified Random Sampling:**
  - **Description:** The population is divided into distinct subgroups (strata), and random samples are then taken from each stratum.
  - **Procedure:** Ensure that each subgroup is represented in the final sample, allowing for analysis within each stratum.
- **Systematic Sampling:**
  - **Description:** Individuals are selected at regular intervals from a randomly chosen starting point.
  - **Procedure:** Determine the sampling interval (population size divided by desired sample size) and select every kth individual.
- **Cluster Sampling:**
  - **Description:** The population is divided into clusters, and entire clusters are randomly selected for the sample.
  - **Procedure:** Randomly select clusters and include all individuals within the selected clusters in the sample.

- **Convenience Sampling:**
    - **Description:** Individuals are selected based on their ease of availability and accessibility.
    - **Procedure:** Choose participants who are readily available, which may not result in a representative sample.
- **Random Sampling with Replacement:**
    - **Description:** After an individual is selected, they are placed back into the population before the next selection.
    - **Procedure:** Everyone has an equal chance of being selected in each draw, even if they were selected previously.
- **Random Sampling without Replacement:**
    - **Description:** Once an individual is selected, they are not returned to the population for subsequent selections.
    - **Procedure:** Ensures that everyone is included in the sample only once.

Covariance and Correlation:

- Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.
- The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.
- The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.
- A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables. Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude.
- Positive: An increase in one of the variables results in an increase in the other.
- Negative: The variables are in opposite directions.
- Zero: Then, no relationship exists.

Eternaltek



Covariance explains the joint variability of the variables.

$$\mathbf{Cov(X,Y)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Where

x_i = Data value of x

y_i = Data value of y

$\bar{x}$ = Mean of x

$\bar{y}$ = Mean of y

N = Number of data values