# Statistics

Statistics is a branch of mathematics that deals with collecting, organizing, analysing, interpreting and presenting data. It involves methods for making decision and drawing conclusions based on the data, as well as techniques for dealing with uncertainty and variability.

## Types of Statistics

1. **Descriptive Statistics:**
   * It involves methods for describing the characteristics of dataset. It includes measures such as central tendency, range, variance, standard deviation and percentiles.
2. **Inferential Statistics**
   * It involves making inferences or generalizations about a population based on a sample data. It includes techniques such as Hypothesis testing, confidence intervals, regression analysis, analysis of variance (ANOVA).

## Central Tendency

It refers to the tendency of data to cluster around a central value or point. It is a way of describing the central point or value in a dataset. Measures of central tendency provide insight into the center of a distribution, helping to summarize and understand overall pattern of the data.

Below are the most common measures of central tendency:

1. **Mean:**
   * The mean is also known as the average, it is calculated by summing up all the values in a dataset and then dividing by the total number of values. It is sensitive to outliers because it considers every value equally.
2. **Median:**
   * The median is the middle value in a dataset when it is ordered from smallest to largest. If there is an odd number of observations, the median is simply the middle value. If there is an even number of observations the median is the average of the two middle values. The median is less influenced by outliers compared to mean and is often used with skewed data**.**
3. **Mode:** 
   * The mode is the value that appears most frequently in a dataset. It is the only measure of central tendency that can be used with nominal data, where number are used to label categories without any inherent order. The mode is also useful for identifying the most common value or category in dataset.

## Variance

* Variance is a measure of how much the values in a dataset vary from the mean. It is calculated by taking the average of the squared differences between each value and the mean. Mathematically, the variance of a dataset with observations and mean is given by:

## Standard Deviation

* Standard deviation is the square root of the variance. It measures the average amount of a variation or dispersion of values from the mean. Standard deviation provides a more interpretable measure of spread compared to variance, as its in the same units as original data.
* Mathematically the standard deviation is given by:

## Range

* Range is the simplest measure of variability and is calculated as the difference between the maximum and minimum values in a dataset. It provides quick indication of the spread of the data but can be sensitive to extreme values (outliers).

## Percentile:

* A percentile is a measure that indicates the percentage of data points that fall below a specific value in a dataset. For example, the 25th percentile (also known as the first quartile, Q1) represents the value below which 25% of the data points fall. Similarly, the 50th percentile (also known as the median) represents the value below which 50% of the data points fall.
* Percentiles can be calculated using the empirical cumulative distribution function (ECDF) of the dataset. Once the data points are sorted in ascending order, the percentile P is the value at the position, where n is the total number of data points.
* Percentiles are useful for understanding the distribution of data and identifying specific cutoff points within the dataset.

## Interquartile Range (IQR):

* The interquartile range (IQR) is a measure of statistical dispersion that describes the range between the first quartile (Q1) and the third quartile (Q3) of a dataset. It is a robust measure of variability that is less affected by outliers compared to the range.
* Mathematically, the IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1): IQR=Q3−Q1.
* The IQR is used to identify the middle 50% of the data and provides information about the spread of the central portion of the dataset. It is often used to detect outliers, with observations lying below Q1−1.5×IQR or above Q3+1.5×IQR considered potential outliers.

## Outliers

Outliers are data points that significantly differ from the rest of the observations in a dataset. They can arise due to various reasons, including measurement errors, data entry mistakes, natural variation or rare-events. Outliers can have a substantial impact on statistical analysis and can distort interpretations and conclusion if not properly identified.

Outliers can be identified by the below methods: -

1. **Standard Deviation Method:**
   * Outliers can be identified as data points that lie beyond a certain number of standard deviations from the mean. For example, any observation that is more than three standard deviations away from the mean may be considered an outlier.
2. **Interquartile Range (IQR) Method:**
   * Outliers can be detected using the interquartile range, which is the difference between the third quartile (Q3) and the first quartile (Q1). Observations that fall below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR are considered outliers.
3. **Box Plot Method:**
   * Box plots visually represent the distribution of data and provide a clear indication of potential outliers. Data points lying outside the "whiskers" of the box plot are often considered outliers.

Once outliers are identified, they can be handled in several ways:

* **Removal:** Outliers can be removed from the dataset if they are deemed to be the result of errors or anomalies. However, this approach should be used cautiously as it can potentially distort the representation of the data and lead to biased results.
* **Transformation:** Data transformation techniques such as log transformation or Winsorization can be applied to reduce the influence of outliers while preserving the overall structure of the data.
* **Robust Methods:** Robust statistical methods, such as using the median instead of the mean or employing robust regression techniques, are less sensitive to outliers and can provide more reliable estimates in the presence of outliers.
* **Reporting:** In some cases, outliers may represent important or meaningful observations. In such instances, it is essential to report the presence of outliers and their potential impact on the analysis and interpretation of results.

Estimates of Variability

Location is just one dimension in summarizing a feature. A second dimension, *variability*,

also referred to as *dispersion*, measures whether the data values are tightly clustered

or spread out. At the heart of statistics lies variability: measuring it, reducing it,

distinguishing random from real variability, identifying the various sources of real

variability, and making decisions in the presence of it.

Standard Deviation and Related Estimates

The most widely used estimates of variation are based on the differences, or *deviations*,

between the estimate of location and the observed data. For a set of data

{1, 4, 4}, the mean is 3 and the median is 4. The deviations from the mean are the

differences: 1 – 3 = –2, 4 – 3 = 1, 4 – 3 = 1. These deviations tell us how dispersed the

data is around the central value.

One way to measure variability is to estimate a typical value for these deviations.

Averaging the deviations themselves would not tell us much—the negative deviations

offset the positive ones. In fact, the sum of the deviations from the mean is precisely

zero. Instead, a simple approach is to take the average of the absolute values of the

deviations from the mean. In the preceding example, the absolute value of the deviations

is {2 1 1}, and their average is (2 + 1 + 1) / 3 = 1.33. This is known as the *mean*

*absolute deviation* and is computed with the formula:

The standard deviation is much easier to interpret than the variance since it is on the

same scale as the original data. Still, with its more complicated and less intuitive formula,

it might seem peculiar that the standard deviation is preferred in statistics over

the mean absolute deviation. It owes its preeminence to statistical theory: mathematically,

working with squared values is much more convenient than absolute values,

especially for statistical models.

With the above estimation we can only evaluate the single numbers to descrive the location or variability of the data. It is also useful to explore how the data is distributed overall by plotting it on graph.

Boxplots

A box plot, also known as a box-and-whisker plot, is a statistical visualization tool that displays the distribution of a dataset along with its central tendency and variability. It consists of a rectangular box (the "box") and two whiskers extending from the box.

The box represents the interquartile range (IQR), which is the middle 50% of the data. It is divided into two parts by a horizontal line, which represents the median (or the 50th percentile) of the dataset. The lower boundary of the box corresponds to the first quartile (Q1), which is the 25th percentile, and the upper boundary corresponds to the third quartile (Q3), which is the 75th percentile.

The whiskers extend from the box to the minimum and maximum values of the dataset that are not considered outliers. The length of the whiskers may vary depending on the method used to determine outliers.

Box plots are particularly useful for comparing the distribution of different datasets or for identifying outliers within a single dataset. They provide a concise summary of the data's distribution and help in understanding its variability and central tendency.

Histograms:

A histogram is a graphical representation of the distribution of numerical data. It consists of a series of adjacent rectangles, or bars, where the width of each bar represents a range of values, and the height represents the frequency or count of observations within that range.

Histograms are commonly used to visualize the frequency distribution of continuous data and to understand its underlying pattern or shape. They provide insights into the central tendency, dispersion, and skewness of the data. Histograms are especially useful for identifying patterns, trends, or anomalies in datasets and are widely employed in exploratory data analysis and statistical analysis

A frequency histogram plots frequency counts on the y-axis and variable values

on the x-axis; it gives a sense of the distribution of the data at a glance.

• A frequency table is a tabular version of the frequency counts found in a

histogram.

• A boxplot—with the top and bottom of the box at the 75th and 25th percentiles,

respectively—also gives a quick sense of the distribution of the data; it is often

used in side-by-side displays to compare distributions.

• A density plot is a smoothed version of a histogram; it requires a function to estimate

a plot based on the data (multiple estimates are possible, of course).

***Data frame***

Rectangular data (like a spreadsheet) is the basic data structure for statistical

machine learning models.

***Feature***

A column within a table is commonly referred to as a *feature*.

*Synonyms*

attribute, input, predictor, variable

***Outcome***

Many data science projects involve predicting an *outcome*—often a yes/ outcome

(in Table 1-1, it is “auction was competitive or not”). The *features* are sometimes

used to predict the *outcome* in an experiment or a study.

*Synonyms*

dependent variable, response, target, output

***Records***

A row within a table is commonly referred to as a *record*.

*Synonyms*

case, example, instance, observation, pattern, sample

## Empirical Rule

## Covariance

* Covariance measures the extent to which two variables vary together. A positive covariance indicates that the variables tend to move in the same direction, while a negative covariance indicates that they move in opposite directions.
* Mathematically, the covariance between two variables and , denoted as is calculated as:

where ​ and ​ are individual data points, and are the means of X and Y respectively, and n is the number of data points.

* The main drawback of covariance is that its value is dependent on the scale of the variables being measured, making it difficult to compare covariances between variables measured in different units.

## Correlation

* Correlation, on the other hand, standardizes the measure of association between two variables by scaling it to a range between -1 and 1. This makes correlation a more interpretable measure of the strength and direction of the relationship between variables.
* The correlation coefficient, denoted as , quantifies the strength and direction of the linear relationship between two variables. A correlation coefficient of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.
* Pearson correlation coefficient is the most common type of correlation coefficient used. It is calculated as:

where ​ and ​ are the standard deviations of variables X and Y respectively.

* Correlation provides a standardized measure of association that is not affected by the scale of the variables, making it easier to compare relationships across different datasets and variables.

## Population

In statistics, a population refers to the entire group of individuals, items, or events that researchers are interested in studying. It is the complete set of all possible elements that share one or more characteristics. The population can be finite, meaning it has a specific, countable number of members, or infinite, meaning it has an unlimited number of members.

## Sample

A sample, on the other hand, is a subset of the population selected for observation or analysis. Sampling involves choosing a smaller group of individuals or items from the larger population and collecting data from them. The purpose of sampling is to make inferences or draw conclusions about the population based on the characteristics of the sample.

### Types of Sampling

* **Random Sampling**:
  + Random sampling involves selecting a sample from the population at random, where each member of the population has an equal chance of being selected. This helps to minimize selection bias and ensure that the sample is representative of the population.
* **Stratified Sampling**:
  + Stratified sampling involves dividing the population into subgroups or strata based on certain characteristics and then randomly selecting samples from each stratum. This ensures that each subgroup is represented in the sample, making it useful when certain subgroups are of particular interest.
* **Cluster Sampling**:
  + Cluster sampling involves dividing the population into clusters or groups and then randomly selecting entire clusters to be included in the sample. This method is useful when it is impractical or expensive to sample individuals directly.
* **Systematic Sampling**:
  + Systematic sampling involves selecting every member of the population after randomly selecting the first member. This method is straightforward and efficient but may introduce bias if there is a systematic pattern in the population.
* **Convenience Sampling**:
  + Convenience sampling involves selecting individuals who are readily available and accessible to the researcher. While convenient, this method can introduce selection bias because it does not ensure that the sample is representative of the population.

Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.

### Sample Mean Versus Population Mean

The symbol (pronounced “x-bar”) is used to represent the mean of a sample from a

population, whereas *μ* is used to represent the mean of a population. Why make the

distinction? Information about samples is observed, and information about large

populations is often inferred from

### Selection Bias

Selection bias is indeed a critical concept in statistics. It occurs when the process of selecting participants or data points for a study systematically favors certain characteristics, leading to results that are skewed or misleading. This bias can occur in various ways, both consciously and unconsciously, and it can seriously compromise the validity and reliability of research findings.

We can guard against this by using a holdout set, and sometimes more than one holdout

set, against which to validate performance. Elder also advocates the use of what

he calls *target shuffling* (a permutation test, in essence) to test the validity of predictive

associations that a data mining model suggests.

Typical forms of selection bias in statistics, in addition to the vast search effect,

include nonrandom sampling cherry-picking data, selection of time intervals that accentuate a particular statistical effect, and stopping an experiment when the results look “interesting.”

### Central Limit Theorem

The CLT states that when independent random variables are added together, their sum tends to be normally distributed, regardless of the distribution of the individual variables themselves. Moreover, as the sample size increases, the distribution of the sample means approaches a normal distribution, even if the original population distribution is not normal.

***Sample statistic***

A metric calculated for a sample of data drawn from a larger population.

***Data distribution***

The frequency distribution of individual *values* in a data set.

***Sampling distribution***

The frequency distribution of a *sample statistic* over many samples or resamples.

***Central limit theorem***

The tendency of the sampling distribution to take on a normal shape as sample

size rises.

***Standard error***

The variability (standard deviation) of a sample *statistic* over many samples (not

to be confused with *standard deviation*, which by itself, refers to variability of

individual data *values*).

### Bootstrap

One easy and effective way to estimate the sampling distribution of a statistic, or of

model parameters, is to draw additional samples, with replacement, from the sample

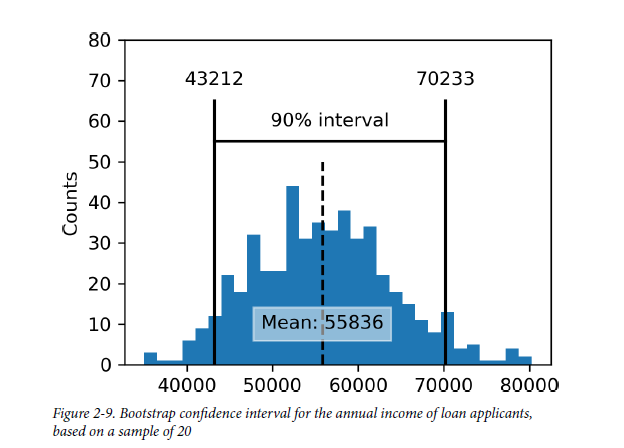
itself and recalculate the statistic or model for each resample. This procedure is called

the *bootstrap*, and it does not necessarily involve any assumptions about the data or

the sample statistic being normally distributed.

### Confidence Interval

A confidence interval is a range of values that is used to estimate the true value of a population parameter, such as a population mean or proportion, with a certain level of confidence. It provides a range of plausible (reasonable) values for the parameter based on sample data.



The percentage associated with the confidence interval is termed the *level of confidence*.

The higher the level of confidence, the wider the interval. Also, the smaller the

sample, the wider the interval (i.e., the greater the uncertainty). Both make sense: the

more confident you want to be, and the less data you have, the wider you must make

the confidence interval to be sufficiently assured of capturing the true value.

## Normal Distribution

The normal distribution, also known as the Gaussian distribution or bell curve, It describes the probability distribution of a continuous random variable that is symmetric and bell-shaped.

Characteristics:

**Symmetry**: The normal distribution is symmetric around its mean, with equal probabilities of values occurring below and above the mean.

**Bell Shape**: The PDF of the normal distribution forms a symmetric bell-shaped curve when plotted, with highest point at the mean and decreasing symmetrically as you move away from the mean.

**Mean, Median, and Mode:** In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

**Standard Deviation:** The spread or dispersion of the data in a normal distribution is described by the standard deviation. Approximately 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations.

***Error***

The difference between a data point and a predicted or average value.

***Standardize***

Subtract the mean and divide by the standard deviation.

***z-score***

The result of standardizing an individual data point.

***Standard normal***

A normal distribution with mean = 0 and standard deviation = 1.

***QQ-Plot***

A plot to visualize how close a sample distribution is to a specified distribution,

e.g., the normal distribution.

## Standard Normal Distribution and Q-Q plots

The standard normal distribution, also known as the Z-distribution or the standard Gaussian distribution, is a specific instance of the normal distribution with a mean of 0 and a standard deviation of 1. It is denoted by N(0,1)N(0,1).

Key characteristics of the standard normal distribution include:

1. **Symmetry:** Like the general normal distribution, the standard normal distribution is symmetric around its mean of 0. The curve is bell-shaped, with equal probabilities of values occurring above and below 0.
2. **Standardization:** Any value from a normal distribution with mean μμ and standard deviation σσ can be converted to a standard normal deviate (Z-score) by subtracting the mean and dividing by the standard deviation: Z=X−μσZ=σX−μ​ This transformation converts the original variable XX to a value on the standard normal distribution, with a mean of 0 and a standard deviation of 1.

## Long-Tailed Distributions

Long-tailed distributions, also known as heavy-tailed distributions, refer to probability distributions where the tails of the distribution extend farther (beyond) out than those of the normal distribution. In other words, these tails refer to higher probability of extreme values (both high and low) compared to what would be expected under a normal distribution.

***Tail***

The long narrow portion of a frequency distribution, where relatively extreme

values occur at low frequency.

***Skew***

Where one tail of a distribution is longer than the other.

## t-Distribution

The t-distribution, also known as Student's t-distribution, is a probability distribution that is similar to the normal distribution but with heavier tails. It is widely used in statistics for hypothesis testing and constructing confidence intervals, particularly when the sample size is small and the population standard deviation is unknown.

Key characteristics of the t-distribution include:

1. **Symmetry:** Like the normal distribution, the t-distribution is symmetric around its mean.
2. **Heavier Tails:** The t-distribution has heavier tails compared to the normal distribution. This means that extreme values are more likely to occur in the tails of the distribution.
3. **Parameter:** The t-distribution is characterized by a single parameter, degrees of freedom (df). The degrees of freedom parameter determines the shape of the distribution. As the degrees of freedom increase, the t-distribution approaches the standard normal distribution.
4. **Shape:** The shape of the t-distribution depends on the degrees of freedom. When the degrees of freedom are small, the t-distribution is more spread out with heavier tails. As the degrees of freedom increase, the t-distribution becomes closer to the standard normal distribution.

## Binomial/Bernoulli Trial & Distribution

***Trial***

An event with a discrete outcome (e.g., a coin flip).

***Success***

The outcome of interest for a trial.

*Synonym*

“1” (as opposed to “0”)

***Binomial***

Having two outcomes.

*Synonyms*

yes/no, 0/1, binary

***Binomial trial***

A trial with two outcomes.

*Synonym*

Bernoulli trial

***Binomial distribution***

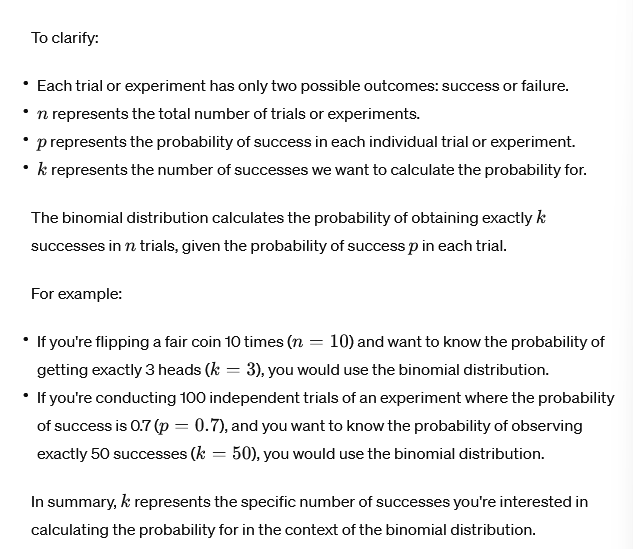
Distribution of number of successes in *x* trials.

*Synonym*

Bernoulli distribution

A binomial trial is an experiment or process that has only two possible outcomes, often referred to as success and failure. Each trial is independent and has the same probability of success, denoted by pp, and the same probability of failure, denoted by 1−p1−p.

The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent binomial trials.



## Chi-Square Distribution

The chi-square distribution is a continuous probability distribution that arises in statistics, particularly in hypothesis testing and confidence interval estimation. It is denoted by and has only one parameter: the degrees of freedom .

Key characteristics of the chi-square distribution include:

1. **Shape:** The shape of the chi-square distribution depends on the degrees of freedom. As increases, the distribution becomes more symmetrical and approaches a normal distribution. However, for small values of , the distribution is skewed to the right.
2. **Support:** The chi-square distribution is defined for non-negative values only (i.e., ).
3. **Parameter:** The degrees of freedom parameter determines the shape of the distribution. It represents the number of independent random variables squared and summed together to form the chi-square variable. The degrees of freedom must be a positive integer ().

## F-Distribution

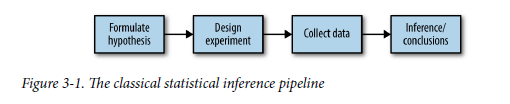
The F-distribution, also known as Fisher's F-distribution, is a continuous probability distribution that is used in the analysis of variance (ANOVA) and in comparing variances of two populations. It is denoted by and has two parameters: the degrees of freedom for the numerator (​) and the degrees of freedom for the denominator ().

Characteristics of F-Distribution

* **Shape:** The shape of the F-distribution depends on the degrees of freedom for the numerator (df1df1​) and the degrees of freedom for the denominator (df2df2​). The F-distribution is always non-negative and skewed to the right.
* **Support:** The F-distribution is defined for non-negative values only (i.e., x≥0x≥0).
* **Parameter:** The F-distribution has two parameters: df1df1​ and df2df2​. df1df1​ represents the degrees of freedom for the numerator, which is associated with the variance of the first sample or group, while df2df2​ represents the degrees of freedom for the denominator, which is associated with the variance of the second sample or group. Both df1df1​ and df2df2​ must be positive integers (df1>0df1​>0, df2>0df2​>0).

## Poisson, Exponential & Weibull Distribution

# Statistical Experiments and Significance Testing



**Statistical Experiments:**

A statistical experiment is a process designed to collect data and test hypotheses about a population. Key components of a statistical experiment include:

1. **Objective:** Clearly define the research question or hypothesis to be tested.
2. **Design:** Plan the experimental design, including the selection of variables, treatments, and sampling methods.
3. **Data Collection:** Collect data according to the experimental design.
4. **Analysis:** Analyze the collected data using appropriate statistical methods to draw conclusions and make inferences.
5. **Interpretation:** Interpret the results of the analysis in the context of the research question and draw conclusions about the population of interest.

**Significance Testing:**

Significance testing is a statistical technique used to determine whether observed differences or associations between variables are statistically significant or merely due to chance. The process typically involves the following steps:

1. **Formulate Hypotheses:** Formulate a null hypothesis (​) and an alternative hypothesis (​). The null hypothesis typically states that there is no effect or no difference between groups, while the alternative hypothesis states that there is an effect or a difference.
2. **Select a Test Statistic:** Choose an appropriate test statistic that measures the difference or association between variables of interest. The choice of test statistic depends on the research question and the type of data being analyzed.
3. **Determine a Significance Level:** Choose a significance level (), which represents the probability of rejecting the null hypothesis when it is actually true. Commonly used significance levels include 0.05 or 0.01.
4. **Calculate the P-value:** Calculate the probability of observing the test statistic (or a more extreme value) under the null hypothesis. This probability is known as the p-value.
5. **Make a Decision:** Compare the p-value to the significance level. If the p-value is less than the significance level (), reject the null hypothesis in favor of the alternative hypothesis. Otherwise, fail to reject the null hypothesis.
6. **Draw Conclusions:** Interpret the results of the significance test in the context of the research question and make conclusions about the population of interest.

***Treatment***

Something (drug, price, web headline) to which a subject is exposed.

***Treatment group***

A group of subjects exposed to a specific treatment.

***Control group***

A group of subjects exposed to no (or standard) treatment.

***Randomization***

The process of randomly assigning subjects to treatments.

***Subjects***

The items (web visitors, patients, etc.) that are exposed to treatments.

***Test statistic***

The metric used to measure the effect of the treatment.