(equation)
R-squared coefficient of determination

(picture + equation)
Confidence interval

(equation)
Standard error

(picture)
Bimodal (distribution)

(equation)
Z-score (t-score)

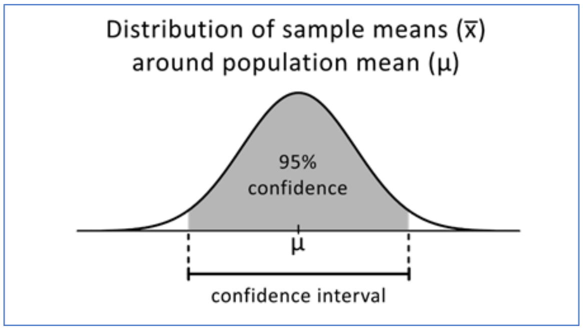
(picture)
Box plot

(picture + equation)
Residual

(picture)
Chi-square test (distribution)

$$\text{Mean:}\quad CI = \bar{x} \pm Z \times SE$$

$$Proportion\text{:}\quad CI = \hat{p} \pm Z \times SE$$



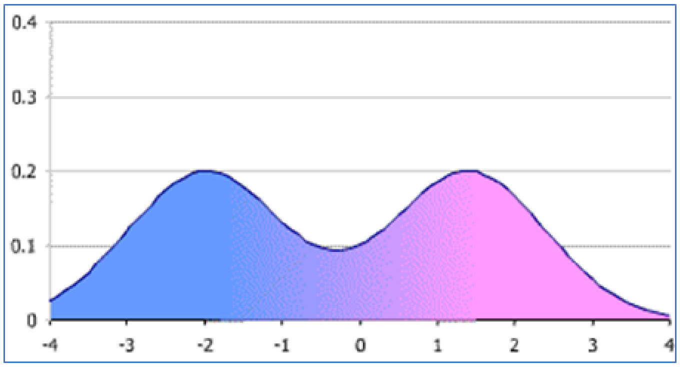
$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

SSE: Sum of squared errors

$$SSE = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

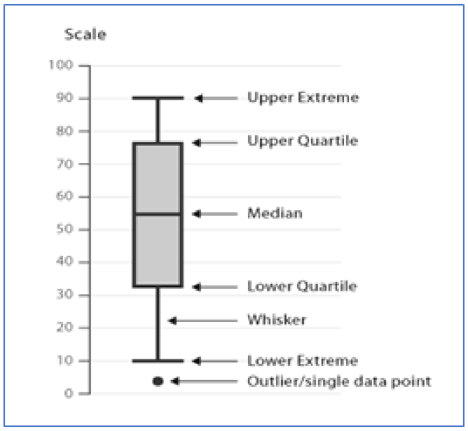
SST: Total sum of squares

$$SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$$



$$\text{Mean:}\quad SE = \frac{\sigma}{\sqrt{n}}$$

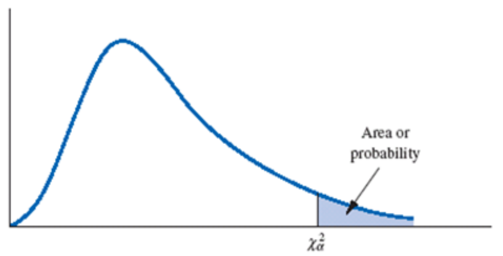
$$Proportion\text{:}\quad SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



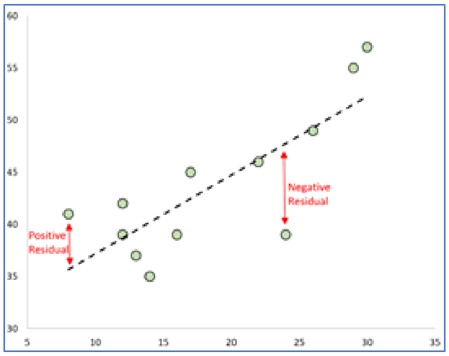
$$\text{Z} = \frac{x - \mu}{\sigma}$$

$$\text{t} - \text{score} = \frac{x - \mu}{SE}$$

TABLE 11.1 SELECTED VALUES FROM THE CHI-SQUARE DISTRIBUTION TABLE*



$$e_i = y_i\text{-}\hat{y}_i$$



(picture)
Cluster sampling

(picture)
Extrapolation

(picture)
Conditions for linear regression
(least squares line)

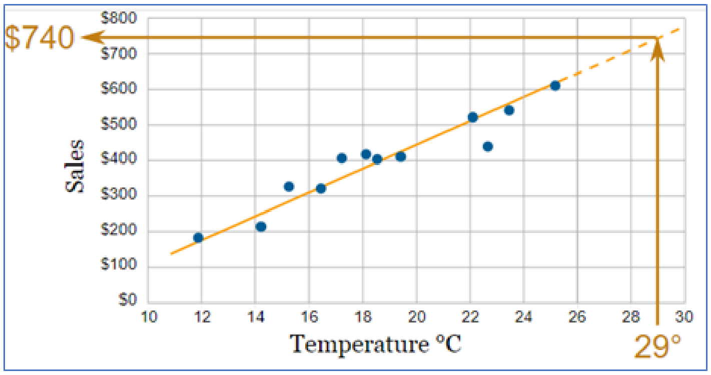
(picture)
Histogram

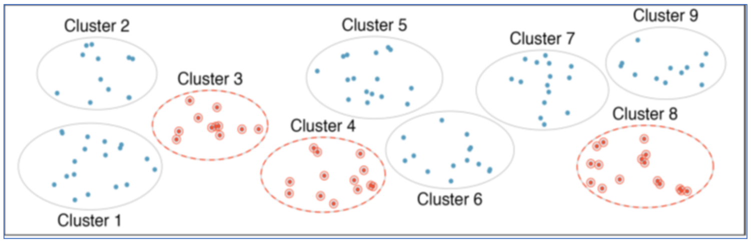
(picture)
Contingency table

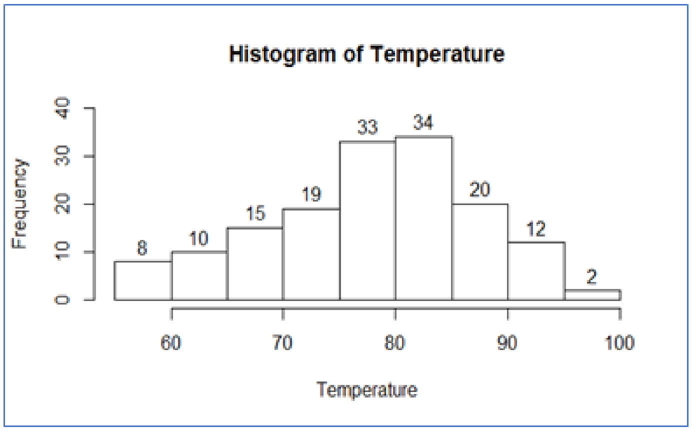
(picture)
Hypothesis test
(one and two-sided)

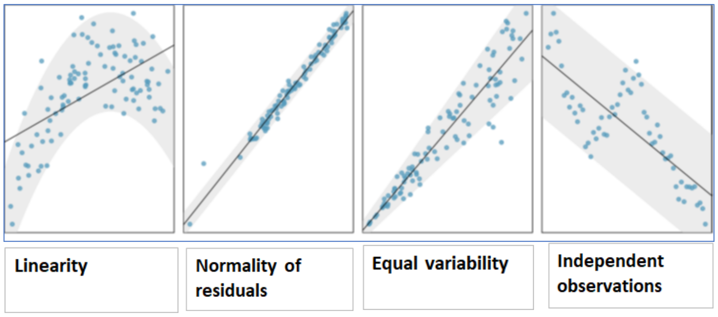
(picture)
Distribution

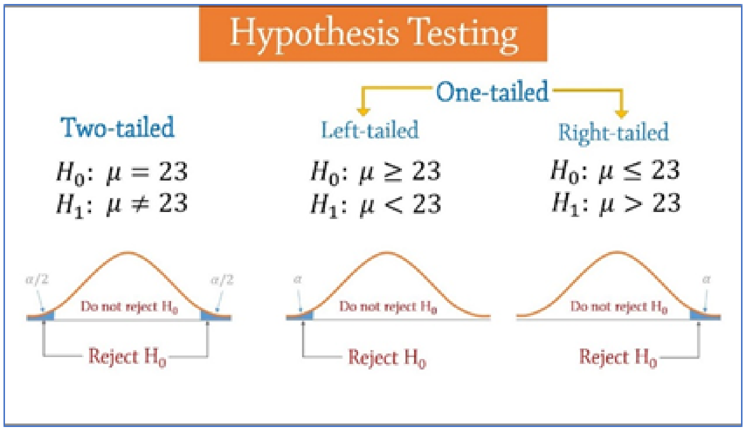
(picture)
Intercept







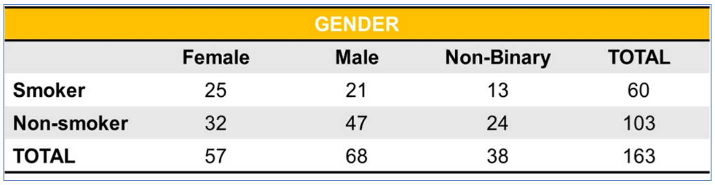




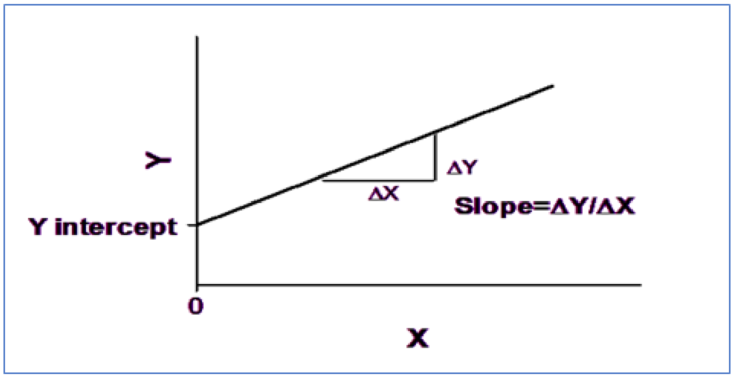
Hypothesis Testing

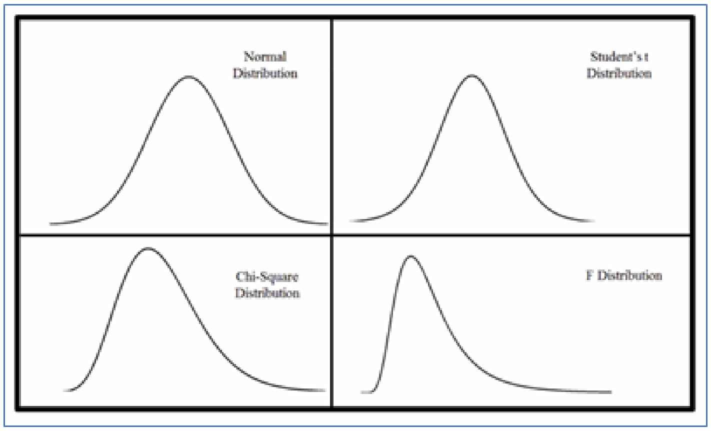
One-tailed

Two-tailed	Left-tailed	Right-tailed
H_0: $\mu = 23$ H_1: $\mu \neq 23$	H_0: $\mu \geq 23$ H_1: $\mu < 23$	H_0: $\mu \leq 23$ H_1: $\mu > 23$



GENDER				
	Female	Male	Non-Binary	TOTAL
Smoker	25	21	13	60
Non-smoker	32	47	24	103
TOTAL	57	68	38	163





(picture)
Interquartile range

(picture)
Linear regression (multiple linear regression)

(picture)
Mean

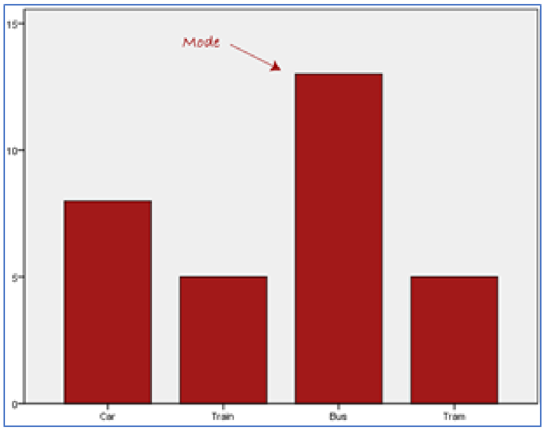
(picture)
Median

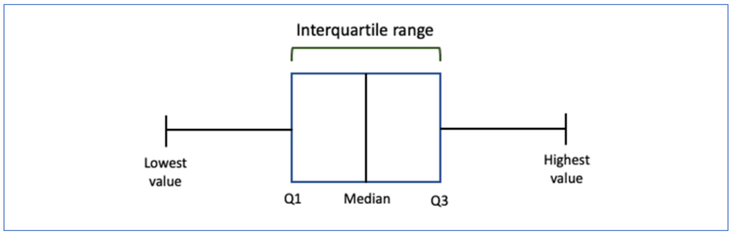
(picture)
Mode

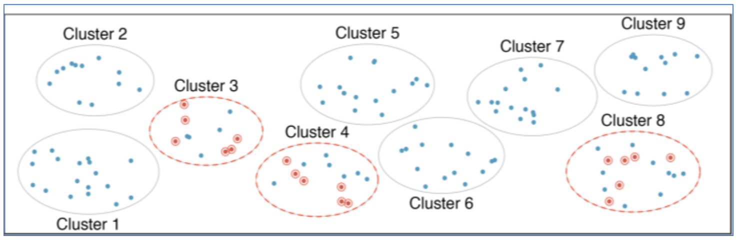
(picture)
Multistage sample

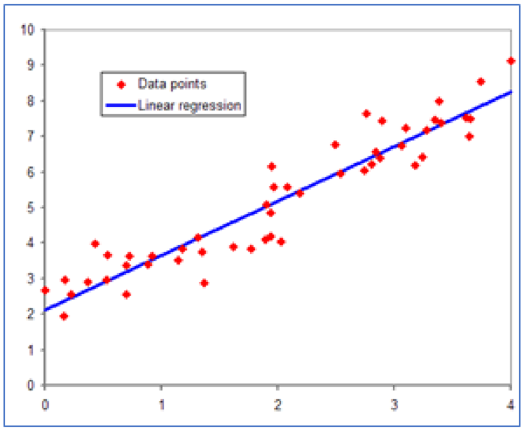
(picture)
Normal (Gaussian) distribution

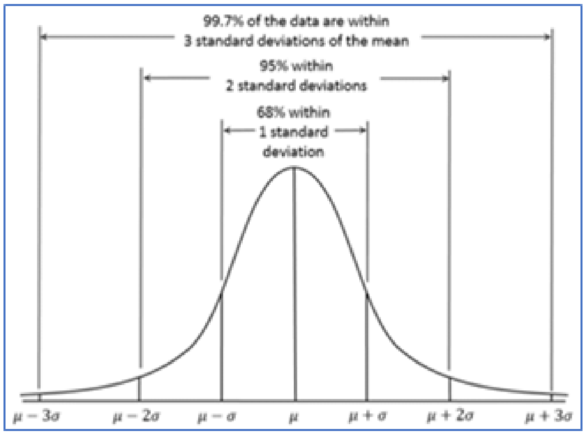
(picture)
Null distribution



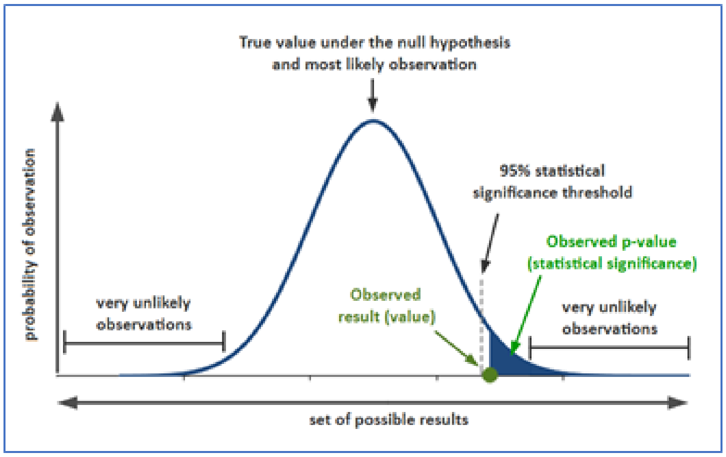


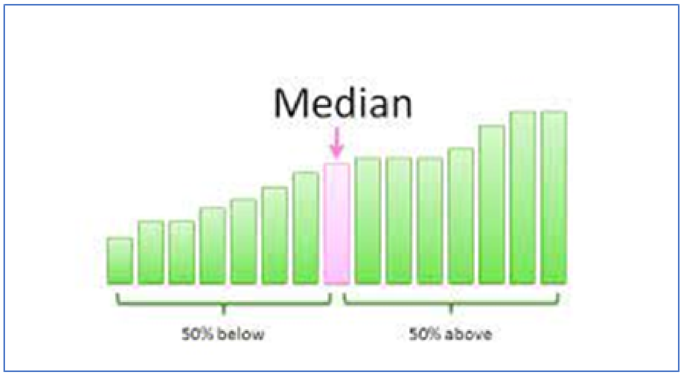






Mean	
Population	μ
Sample	$\bar{x}$





(picture)
Outlier

(picture)
Range (of a distribution)

(picture)
Population

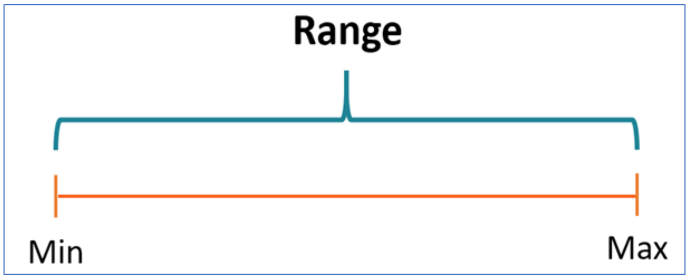
(picture)
Relationship
(positive and negative)

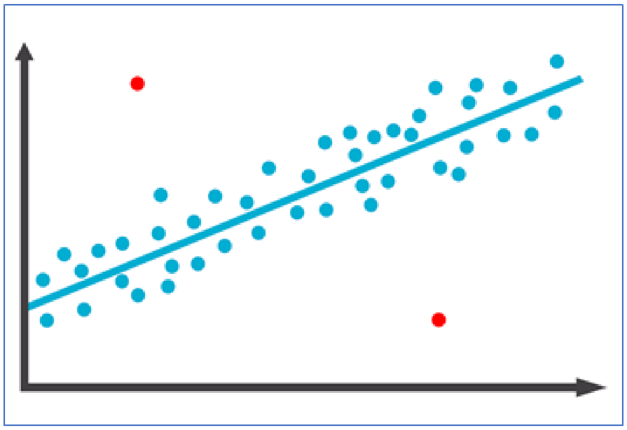
(picture)
Predicted value &
observed value

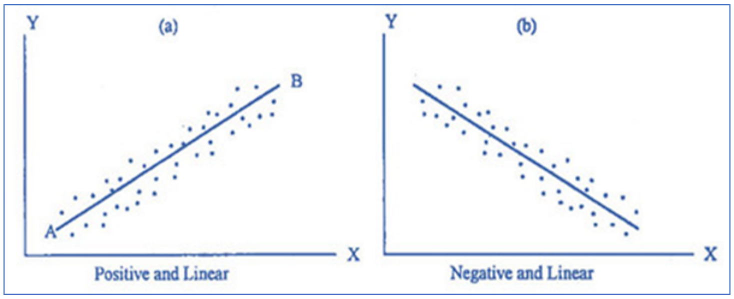
(picture)
Sample, sample size

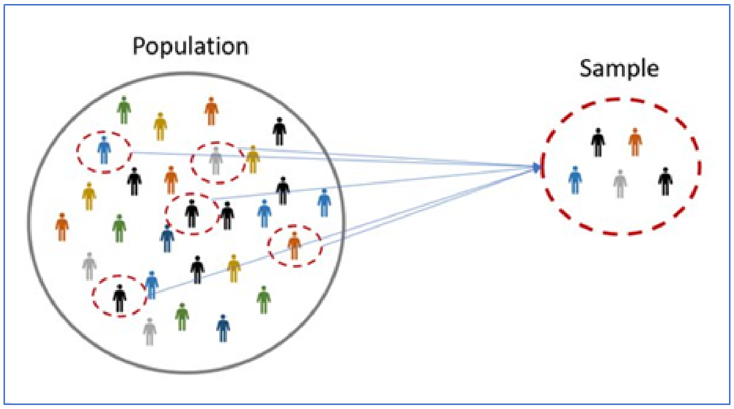
(picture)
Quartile

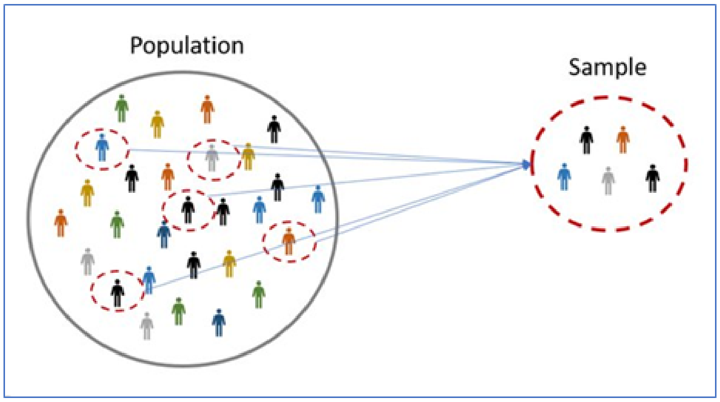
(picture)
Significance level

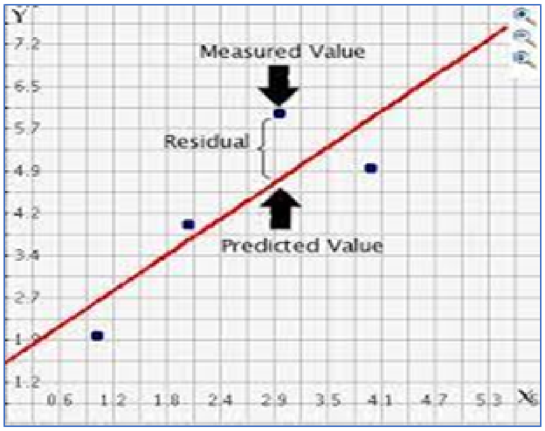

Range
Min
Max

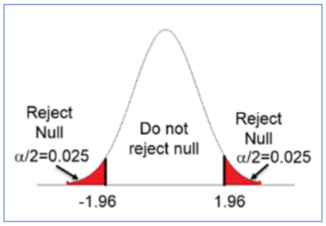


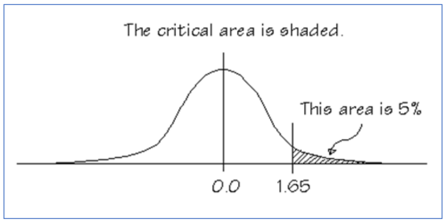

Y
(a)
B
A
X
Positive and Linear
Y
(b)
X
Negative and Linear

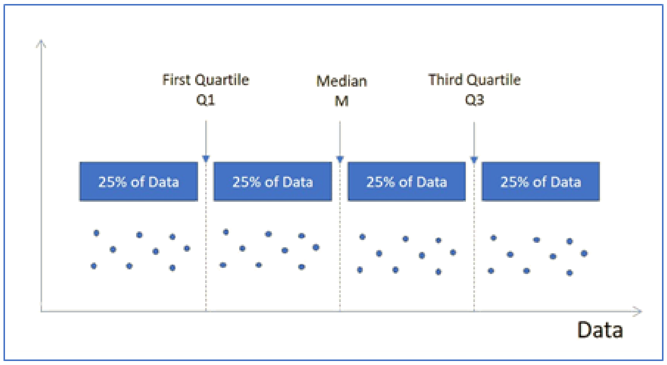

Population
Sample


Population
Sample


Y
7.2
6.5
5.7
4.9
4.2
3.4
2.7
1.9
1.2
Measured Value
Residual
Predicted Value
0.6
1.2
1.8
2.4
2.9
3.5
4.1
4.7
5.3
X


Reject Null α/2=0.025
Do not reject null
Reject Null α/2=0.025
-1.96
1.96


The critical area is shaded.
This area is 5%
0.0
1.65


First Quartile Q1
Median M
Third Quartile Q3
25% of Data
25% of Data
25% of Data
25% of Data
Data

(picture)
Skewness (right/left
skewed distribution)

(picture)
Symmetrical distribution

(picture)
Slope

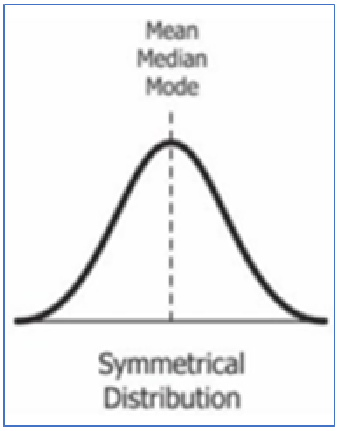
(picture)
T-distribution

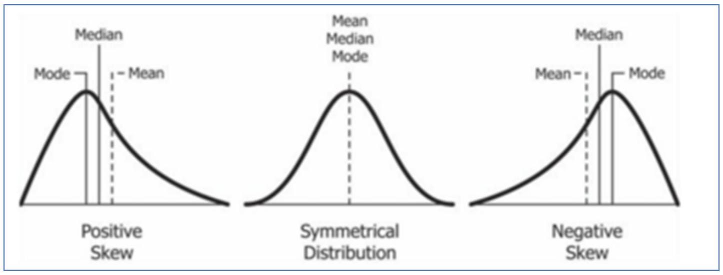
(picture)
Standard deviation

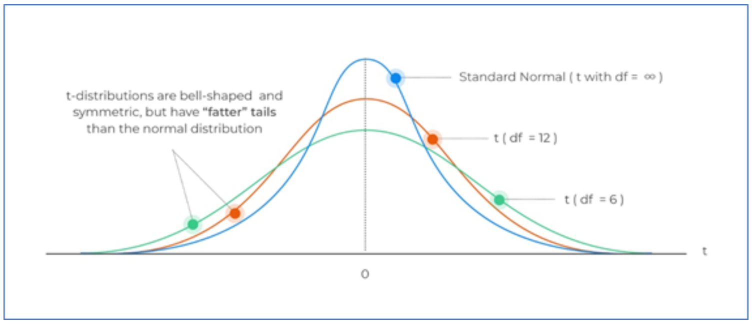
(picture)
Unimodal

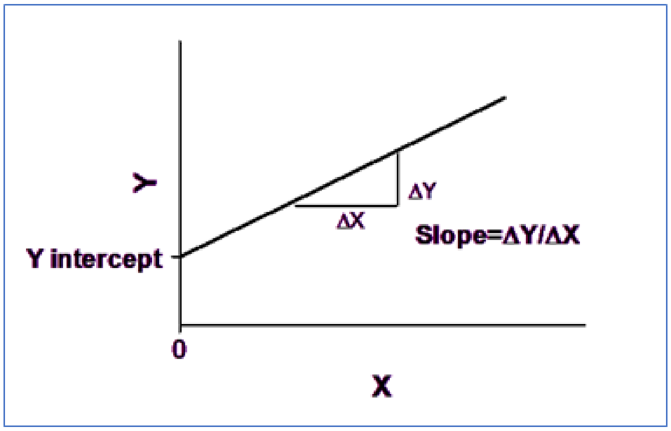
(picture)
Strata
(Stratified sampling)

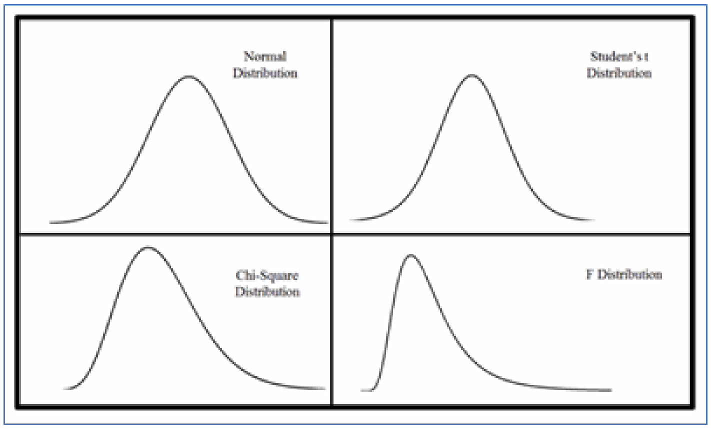
(picture)
Variable



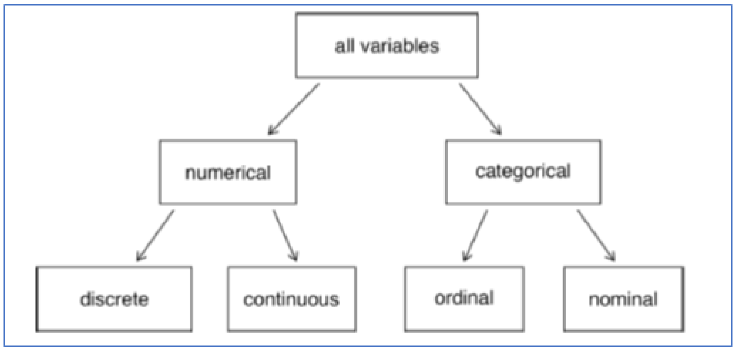


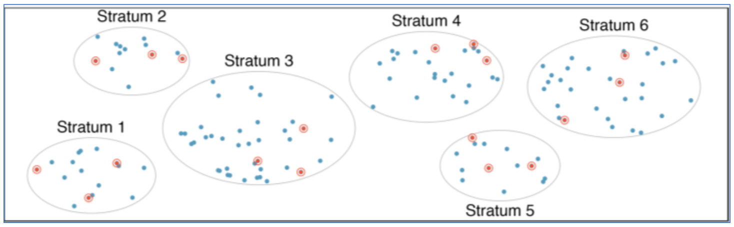






Stndard Deviation	
Population	σ
Sample	s





(picture)
Variance

Bootstrapping,
bootstrap sample

(picture)
p-value

Box plot (*)

Alternative hypothesis,
Ha (research hypothesis)

Categorical (variable)

Bimodal (distribution) (*)

Central Limit Theorem

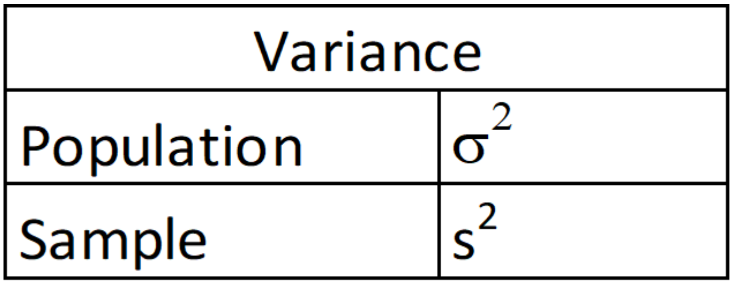
This is a statistical procedure that re-samples from a single dataset to create many simulated samples

Each of these simulated samples has its own properties, such as the mean, median or SD, thus after bootstrapping we can build a sampling distribution for any statistic of our interest.

How bootstrapping works?

1. Equal probability for each original data point in the sample.
2. It is possible to select a data point more than once “with replacement.”
3. The process creates resampled data sets same size as the original.

Bootstrapping is an artificial way of sampling, and it does not create new data.

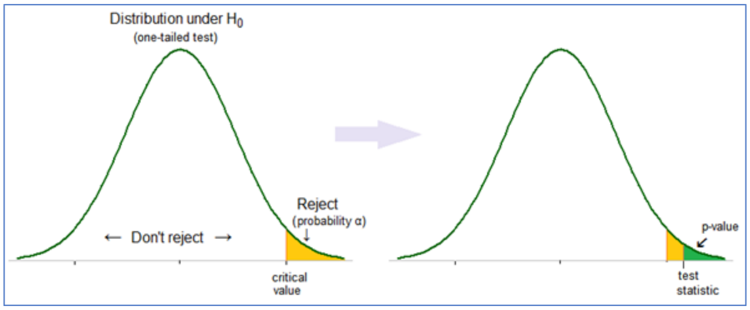


Variance	
Population	σ^2
Sample	s^2

Type of plot to visualize several key statistics (such as the median, the quartiles, extreme values, etc.) and distribution of a variable (skewness, symmetry, outliers, etc).

The length of the the box represents the interquartile range.

The dark line inside the box represents the median.



Also called qualitative variable, it is a variable that can take on one of a limited, and usually fixed, number of possible values. It is not possible to add, subtract or take averages with its values. Examples are education level, ethnicity, gender, etc.

Also called the research hypothesis (H_a). It represents an alternative claim to the null hypothesis and it is often represented by a range of possible values for the value of interest.

What researchers expect to be true.

In general, larger samples are more reflective of population characteristics. This theorem tells us that as the sample size increases the distribution of a sample statistic (such as the sample mean, proportion, or regression slope) approximates a normal distribution, regardless of the distribution of the original variable.

This is a probability/frequency distribution with two different modes. These appear as distinct peaks (local maximum) in the probability density function (or frequency distribution).

Chi-square test (distribution) (*)

Cluster sampling (*)

Conditions for linear regression (least squares line) (*)

Confidence interval (*)

Contingency table (frequency table, cross tabulation) (*)

Continuous (variable)

Control group

Control variable

Also called frequency table or cross tabulation, it summarizes data for two categorical variables. Each value in the table represents the number of times a particular combination of variable outcomes occurred. The independent variable is usually placed among the columns (top row) and the dependent variable is positioned down the rows. Totals in the bottom row or at the far-right column are called the marginal frequencies → they summarize only one variable. To analyze the relationship between two variables in a contingency table we use a Chi-square test.

A test of statistical significance for the relationship among variables in a contingency table. We use this distribution to determine if the relationship in a cross-tabulation is strong enough to infer that the same relationship exists in the whole population.
H_0: There is no relationship between the variables in the population
H_a: A relationship exists in the population
The computation of this statistic is not very complicated but it takes several steps. We can use statistical software to compute this statistic, and the corresponding p-value.

A variable that can only take numerical values, and these values can be broken into decimals or fractions. For example: weight, height, income, speed.

In this kind of sample, we break up the population into many groups, called clusters. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample. For example, considering only 5 schools from a city, and all the students in each of these 5 schools to conduct certain study (the schools being the clusters).

The group that is not assigned to treatment in an experiment. This group receives either no treatment, a standard treatment whose effect is already known, or a placebo (a fake treatment). This group acts as a reference (or base line) for the results to explore the causal effects of the treatment on a research experiment.

LINE

Linearity: The data should show a linear trend.
Independent observations: Observations are independent from each other, and they are representative of the population.
Normality of residuals: the data points vary symmetrically around the line. Balance between points above and below the line, with no obvious outliers.
Equal variability (along the regression line): the variance is roughly constant, as we move along the regression line (homoscedasticity).

Any variable that's held constant in a research study. It is not a variable of interest in the study, but it's controlled because it could influence the outcomes.

This is a range, calculated from the sample, that represent our best guesses of the population parameter (e.g., population mean or proportion). A 95% CI should cover the population mean in 95% of all samples.

Convenience sample

Correlation

Dependent variable
(response variable, outcome)

Discrete (variable)

Distribution (*)

Experiment
(randomized experiment)

Extrapolation (*)

Histogram (*)

In Statistics, this is a function that shows the possible values for a variable and how often they occur, i.e. their frequency. For example if we survey 500 adults and we ask about their age, we can build the age distribution summarizing the number of people that is 18yo, 19yo, 20yo, etc. We can show the distribution on a table or using a histogram.

This is the technique used to build a sample, but where individuals (or observations) who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City citizens (this might be more convenient than stopping people in many different neighborhoods in NY)

These are studies where the researchers assign treatments to cases. The assignment can be done using different techniques, not necessary randomly. The group that is not assigned with any treatment is called the control group.

When individuals are randomly assigned to a group, it is called a randomized experiment. A randomized experiment is used to evaluate the causal relationship between two variables. The explanatory (independent) variable is claimed to have a causal effect on a response (dependent) variable.

A value that describes the strength and direction of the linear relationship between two variables. It takes values between -1 and 1, we denote the correlation by 'r'.

This value has no units and will not be affected by a linear change in the units (e.g., going from inches to centimeters).

It is the regression slope when both 'x' and 'y' have been standardized.

Linear regression is based on a set of data with an upper and lower limit, we do not know how the data outside of our limited window will behave. This is the technique applied to a model to estimate values outside of the realm of the original data. By using a linear model (or any model) to extrapolate data we are assuming that the relationship is valid beyond the limits where data has been collected (or analyzed).

When we suspect one variable might (causally) affect another, we label the first variable the explanatory variable and the second the response variable. Explanatory variable → might affect → response variable.

This variable is usually called 'y', and is plotted in the vertical axis.

Type of plot used to visualize the frequency of cases within groups in a data set (the distribution of a variable). The groups can be individual values or numerical intervals of the same size.

A variable that can only take numerical values with jumps. Typical to count units of any kind when units cannot be broken into fractions, e.g. counting people, counting votes, counting cars, etc.

Hypothesis

Hypothesis test
(one and two-sided) (*)

Independence

Independent variable
(explanatory variable,
predictor)

Indicator variable /
dummy variable

Inferential statistics

Intercept (*)

Interquartile range (*)

Also called dummy variables, they are used to transform categorical variables into a numerical form, so they can be used in linear regression. The most typical transformation is using a 0 - 1 coding for categorical variables with only 2 outcomes. For example if we are analyzing the price of a product, and its condition can be either used or new, the variable is 'condition' and its values are 'new' or 'used'. Then we can assign the values 0: if the product is new and 1: if the product has been used. The interpretation of the slope is the exact same than for a regular variable in linear regression, when x increases in 1 unit, i.e. from 0 to 1, i.e. the value of the slope for the variable 'condition' will indicate the difference in price for the used product (x=1) compared to the reference category (x=0) or the new product.

This is a statement -- about a population or effect -- that we want to test to determine whether it is true or false. Examples of them are: “Eating dark chocolate may aid weight loss”, “Poverty influences domestic terrorism”, “Exposure to sunlight improves your sleep”.

In simple words is a formal proposition of a relationship between dependent and independent variables

The use of quantitative techniques to make generalizations from a sample to a population.

In Hypothesis testing, to reject the Null-hypothesis, the test statistic (z-score, t-score, etc) must be larger than the critical value associated to our level of significance 'α . Alternatively, if the p-value associated to the test statistic) (z-score, t-score, etc) exceeds the 'α value, we fail to reject the null hypothesis. Test statistics are directly related to p-values. The higher the score the lower the p-value.
One tailed test → the hypothesis specifies a direction. We are interested in values larger (or smaller) than a certain value. The level of significance (a value) is concentrated in one tail of the corresponding distribution.
Two tailed test → The hypothesis states that the mean value (or a proportion) differs from the population mean (or proportion) in either direction. It does not state a direction. In this case the significance is distributed among the two tails → we use (α 2) to compare with the p-value.

The point where the regression line intercepts the vertical axis (y-axis). It is the 'y' value when the 'x' value equals 0. Usually called 'α or 'b_0'. In some cases values of x = 0 are not observed or they are irrelevant, so the interpretation of the intercept is merely mathematical but useless from a practical view.
$y=\alpha+\beta x$ (the value 'α' in this case has nothing to do with the level of significance used for hypothesis testing)

Two events are independent if the occurrence of one event does not affect the chances of the occurrence of the other event.

Also called IQR, this is the difference between the 3rd and 1st quartiles (the 75th and 25th percentiles). This is easy to visualize in a box plot, it corresponds to the length of the box. The more variable (spread) the data, the larger the standard deviation and IQR tend to be.

A variable that might affect another. Also this is a variable that isn't changed by the other variables you are trying to measure. We refer to this variable as an explanatory variable or a predictor when it might be found a causal relationship between this (independent) variable and another (dependent) variable.
Ex planatory variable → might affect → response variable.
Also it is usually called 'x', and is plotted in the horizontal axis.

Linear regression (multiple linear regression) (*)

Margin of error

Mean (*)

Median (*)

Mode (*)

Multistage sample (*)

Nominal (variable)

Normal (Gaussian) distribution (*)

The data value with the greatest frequency, i.e. the most repeated value. It can take one or more than one value → see bimodal distribution.
It is not affected by extreme values, and it is always a realistic value.

A statistical technique used for prediction or to evaluate whether there is a linear relationship between two (or more) numerical variables. Also defined as a statistical method for fitting a line to data where the relationship between two variables, 'x' and 'y', can be modeled by a straight line with some error:
$y = b_0 + b_1 x + e$
The values b_0 and b_1 represent the model's intercept and slope, respectively, and the error is represented by 'e'.
A multiple regression model is a linear model with many predictors. In general, we write the model as
$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$, when there are 'k' predictors.
.

This is like a cluster sample, but rather than keeping all observations in each cluster, we would collect a random sample within each selected cluster. In other words we divide the population in groups (clusters), we select only a few clusters, and then in each of the selected clusters we select a random sample.

Usually computed as 2 standard errors. A point estimate plus or minus the margin of error is a 95% confidence interval.

A categorical variable where the values have no natural order -- e.g., race/ethnicity, gender or religion.

The arithmetic average of a set of data points (the sum of the observed values divided by the
number of observations). This is a common way to measure the center of a distribution of data, typically represented by 'μ'   or $\bar{x}$. It is very sensitive to extreme values (outliers).
It can take a not realistic value, for example the mean for the household size in the US in 2021 was 3.13 persons, but we don't have households with 3.13 people.

One of the most common distributions in statistics. It is symmetric, unimodal and bell-shaped. It is determined by the mean and standard deviation of a population. In this distribution about 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. When this distribution has mean 0 and standard deviation 1 is commonly referred to as the standard normal distribution.

This is the middle observation on a set of numbers when they are ranked in order (low to high).
Also the 50th percentile – half of the observations have values below and the other half above it if we rank the observations. If there are an even number of observations, there will be two values in the middle, and the median is taken as the average of these two values. This is not affected by extreme values and typically takes a realistic value.

Null Hypothesis

Null distribution (*)

Numerical (variable)

Observational study
(observational data)

Ordinal (variable)

Outlier (*)

Parameter

Percentile

Variable that measures order but not magnitude. Ordinal variables are categorical.
For example: Rating scales: strongly disagree, disagree, agree and strongly agree.

Also called H_0 is a neutral claim about the population, or about an effect. This hypothesis often represents either a skeptical perspective or a claim of “no difference” or "no effect" to be tested.

These are extreme values in a distribution, i.e. points that are considered unusually high or unusually low. They may have a disproportionate influence on the mean. It is important to analyze if they actually belong to the population we are studying.

This is the sampling distribution of the statistic that would exist if the null hypothesis were true. Therefore, this distribution will always be centered at the value of the parameter given by the null hypothesis.

A measure used to summarize a population, such as a population mean or population proportion. They represent the “true” value of interest. Commonly, the population parameters are unknown.

Variable that can take a wide range of numerical values, and it is typically sensible to add, subtract, or take averages with its values. Examples are height, school size, grades, household size, speed, etc.

This is a number with x% of the observations below and (100 − x%) of the observations above. For example, if the 90th percentile of the SAT scores is 720, this means that 90% of students have a SAT score below 720 and 10% of students above 720. The percentile typically is a realistic value.

A study that is based in observational data, i.e. data where no treatment has been explicitly applied
(observations happen to belong to a group but they where not intentionally assigned to a treatment or control group). Making causal conclusions based on observational data is not recommended. These studies are generally only sufficient to show associations or form hypotheses that can be later checked with experiments.

Population (*)

Random sampling
(simple random sample)

Predicted value &
observed value (*)

Range
(of a distribution) (*)

Quartile (*)

Relationship
(positive and negative) (*)

R-squared (coefficient of
determination) -
adjusted R-squared (*)

Residual (*)

This is the technique used to build a sample, but considering that each case in the population has an equal chance of being included in the sample, and the cases in the sample are not related to each other. The act of taking a this kind of sample helps minimize bias.

The total set of items or individuals that we are concerned about.
Its size is represented by "N" (uppercase). Typically we don't have information for it, thus we use a sample to make inferences about their parameters.

For a distribution, or a variable, this is the difference between the minimum and maximum value in the distribution of numerical values.

The observed value is the actual (real) value that is obtained by observation or by measuring the available data. The predicted value is the value of the variable predicted based on the regression analysis.
y = observed value
$\hat{y}$ = predicted value

<u>Positive relationship</u>: if both variables increase (or decrease) at the same time.
<u>Negative relationship</u>: the independent variable decreases as the dependent variable increases (or vice versa).
<u>No relationship</u> at all: when there is not a clear pattern between both variables.

Each of four equal groups into which a population (sample) can be divided according to the distribution of values of a particular variable. We also use this name to each of the three values of the random variable that divide a population (sample) into four groups. Each of them represents 25% of the observations.

This is the vertical distance between a data point and the regression line. If the data point is greater than the predicted value (the value given by the regression line), then the point will be above the line in the graph, so it will be positive and the predicted value is said to be underestimated. On the contrary, if the data point is smaller than the predicted value, then the point will be below the line in the graph, and this value will be negative and the predicted value is said to be overestimated.

If provided with a linear model, we might like to describe how closely the data cluster around the linear fit. The R-squared of a linear model describes the amount of variation in the outcome variable ('y') that is explained by the least squares line (the 'x' variables). R-squared is calculated as the squared of the correlation.
In multiple linear regression, we use the adjusted R-squared, which considers only the variables that are significant in the model.

Sample, sample size (*)

Significance level (*)

Sampling distribution

Skewness (right/left skewed distribution) (*)

Sampling error

Slope (*)

Sampling variation

Standard deviation (*)

Also called $\boldsymbol{\alpha}$, this value will determine the % of estimates that will fall within the confidence interval.
For hypothesis testing is important to consider if the % or level of significance must be distributed among the two tails, or just one tail of the distribution.

This is a subset (a small fraction) of a population. Its size is represented by "n" (lowercase).

This is an asymmetrical distribution that occurs when more cases fall on one side of the mean, i.e. there is a higher density of data on one side. In this distribution (for a continuous variable) the mean is pulled in the direction of the skew. The side with a smaller density of observations indicates the skew, i.e. if there is a long tail to the right then the distribution is right (or positive) skewed, and if there is a long tail to the left then the distribution is left (or negative) skew. Typical skewed distributions are income or household size.

This is the distribution that we would construct if we sequentially drew many, many samples from the population.
These distributions are not distributions of values of simple variables, they are theoretical (or empirical) distributions of sample statistics, like the mean or a sample proportion.

One of the values to describe a line that indicates how much increases the dependent variable 'y' for a given increment in the independent variable 'x' or how much 'y' will change if we increase 'x' by one unit. Usually called $\boldsymbol{\beta}$ or b.
$y=\boldsymbol{\alpha} + \boldsymbol{\beta}x$

This is the natural difference between the mean (or any statistic) of a single sample and the mean (or any statistic) of the population. In other words we know that the mean of any sample ($\bar{x}$ differs from the mean of the population ($\boldsymbol{\mu}$). The larger the sample the smaller is its value.

Also called SD, it is the typical distance of each observation from the mean.
The SD is calculated as the square root of the variance. In a normal distribution about 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations.

This is the term we use to describe the following. If we consider multiple samples from a population, the mean (or any statistic) for each sample will be different, i.e. there is a variation on the statistic among different samples. This is different than sampling error, in the latter we compare the sample mean with the population mean.

Standard error (*)

Standardize
(standardization)

Statistic / point estimate

Strata
(stratified sampling) (*)

Symmetrical
distribution (*)

Test statistic

Treatment group

Type I error

In this type of distribution, the two sides (or top and bottom in a box plot) look very similar. In this distribution, the mean and the median are equivalent (very similar).

Also called SE, this is also defined as the standard deviation of a sample statistic across all possible samples, i.e. the standard deviation of the sampling distribution. It is a standard for how large the sampling error is likely to be. It decreases when we increase sample size.

This is a summary value that is used to compute the p-value in a hypothesis test. For example the z-score and t-score are test statistics.

Technique to rescale a metric into something more easily interpretable and comparable. In other words standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables. Typically, to apply this technique, you calculate the mean and standard deviation for a variable.

The sets of participants in a research study that are exposed to some manipulation or intentional change in the independent variable of interest (i.e. the patients that receive a medication, or the students that assigned with a tutor). They are an integral part of experimental research design that helps to measure effects as well as establish causality.

A measure (a single number) used to summarize data from a sample. They are typically used to estimate a population parameter. They represent observed values.

Type of error we commit when we reject the Null hypothesis, but in the reality the null hypothesis is true.
This will happen the α% of the times. Increasing the significance level (decreasing a) will decrease the probability of committing this type of error.

This is a sampling technique where the population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. For instance, select 10 students from every single school in a city (school being the strata).

Type II error

Variance (*)

Unimodal (*)

Z-score (t-score) (*)

Unit of analysis

p-value (*)

Variable

A measure of dispersion (indicates how closely the data clusters around the mean).
This is calculated as the arithmetic average of the squared differences of the data values from the mean. It is the square of the standard deviation.

This is the type of error we commit when we accept the null hypothesis as true (we fail to reject it), when in fact is false
Increasing the significance level (decreasing **α**) will increase the probability of committing this type of error.
Increasing the sample size will minimize this type of error

This is the value calculated for an observation that is defined as the number of standard deviations (or standard errors) it falls above or below the mean. It has a standard normal distribution in large samples or t distribution in smaller ones (the t distribution is like the normal distribution, but with heavier tails). For instance, if the observation is 1.5 standard deviations above the mean, then its z-score is 1.5.

This is a probability distribution with only one mode, i.e. we only see one peak in the probability density function.
Examples of these type of distributions are the normal distribution, the T-distribution, the F-distribution or the Chi-squared distribution.

This is the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis were true. We typically use a summary statistic of the data, such as a difference in proportions, to help compute the p-value and evaluate the hypotheses. This summary value that is used to compute the p-value is often called the test statistic. The typical rule to reject the null hypothesis is compare this value with the significance level '.**α** '
If p-values are low → the null must go.

The major entity that you are analyzing in your study.
Can be a person, a school, a country, etc.

This is a quantity that may assume any one of a set of values. They can be of different types: numerical or categorical. Numerical variables can be discrete (household size, class size) or continuous (height, speed, income) and categorical variables can be ordinal (education level, satisfaction rating) or nominal (gender, ethnicity).