THE LIST OF THE MOST FREQUENT TERMS IN BEGINNER'S STATISTICS

Term	What does it mean?
Alternative hypothesis	Hypothesis where you pose a sentence in which something is changing, you are questioning whether something in the sample is different than in the population itself. Example – Women are, on average, paid less than men.
Box plot	The graphic notion is based on the five-number summary (mean, median, quartiles), and it shows outliers (extreme values very different from other values). It gives a better look at the data than a plain histogram or bar plot.
Confidence interval	A range of values used to estimate a population parameter, with a specified level of confidence.
Correlation	The statistic is there to measure the degree (measured by a coefficient) to which two numeric variables are related. It doesn't show the connection nor does it imply a cause-effect relationship. It shows if the relation between two variables is strong, is it negative or positive.
Correlation (negative)	If a correlation is negative, then it means that if the value of one variable goes up, the value of the other variable goes down.
Correlation (positive)	If a correlation is positive, then it means that if one variable's value goes up, the value of the other variable goes up as well.
Descriptive statistics	Statistics that describe data, with the help of summaries (mean, median, outliers…) and graphic notions.
Distribution	The way in which the values of a random variable are spread or distributed.
Frequency	The number of times a particular value or category appears in a dataset
Hypothesis	The assumption we are testing whether is it true or not. It can be the null hypothesis (nothing is changing) and the alternative hypothesis (different state).

Hypothesis test	The test we are using to check if our hypothesis is true or not.
Hypothesis testing	A method of making decisions or inferences about population parameters based on sample data. It involves testing a null hypothesis against an alternative hypothesis.
Inferential statistics	Statistics that are using samples and population, it is using samples derived from a bigger population, to check some variables in the population. For example – we can't measure the height of every person on the planet or in the country, so we are picking a smaller, representative sample of that planet or country, and then using different tests to check our hypotheses.
Interquartile range (IQR)	IQR is a measure of variability, it is checked as a part of descriptive statistics. Data can be divided into quartiles (4 of them), and IQR is checking the middle 50% of the data around the median. It checks how much data is around the median (variability) because the median tends to be a more safe option than the mean. Mean can be affected by outliers.
Kurtosis	A measure of the tailedness of the distribution of values in a dataset.
Logistic regression	One of the regression analysis, but uses dichotomous (Boolean, binary) variables like yes/no, 0/1, male/female. It checks the relationship between one of those dichotomous variables and other variables that are from another type (nominal, interval…)
Mean	The average value of the variable or dataset
Median	Value separating the values of the variable into two halves – 50% of the sample/population has a value lower than the median, and vice versa. It is not affected by outliers, so it is a safer option to use for explaining the average/middle value of the dataset/variable.
Mode	The value that appears most frequently in a dataset.

Normal distribution	A continuous probability distribution that is symmetrical and bell-shaped, with most of the observations clustering around the central peak.
Null hypothesis	A statement that nothing has changed. Example: H0 – The paycheck of females is the same as the paycheck of males, on average.
Outlier	The extreme value is very different from other values in the dataset. It affects the mean, but not the median. If you have outliers (visible on boxplot or plot), then you have to decide for yourself if you are going to delete that value, if that value is an error, or normal state (happens in medical data).
Parameter	Numbers from the population (mean of the population, the median of the population, etc.)
Percentile	The dataset can be divided into percentiles, which show you a number where a certain percentage of values fall below that number, it shows where one instance is according to other ones. If you have 60th = 120 in exams, then your score was better than the other 60% of other people taking that test, but also – 40% of people were better than you.
Population	The entire set of individuals or items of interest in a statistical study.
Probability	A measure of the likelihood that a particular event will occur.
p-value	The probability of obtaining a test statistic at least as extreme as the one observed, given that the null hypothesis is true. It is used in hypothesis testing to determine the significance of the results.
Quartile	The number that divides the dataset into quarters (median is the 3rd quartile)
Range	The difference between the highest and lowest values in a dataset.
Regression analysis	Analysis in which you are estimating a relationship. Here, you have two types of variables, just like in linear equation (x, y) – dependent and independent. Its result can be in a form of a linear or polynomial equation.

Sample	A subset of the population used to make inferences about the entire population.
Scatterplot	A plot that shows the relationship between two numerical variables, but in a form of x and y coordinates (just like in a coordinate system).
Skewness	Asymmetry of data. If distribution were symmetrical then it would have both tails of the same height, without outliers. The normal distribution has a look of a bell curve. If a distribution you are looking at doesn't look like a bell curve or close to it, it is skewed.
Skewness (negative)	Skewness where the left tail of the distribution is longer than its right tail. In the distribution, you have lower and higher values. If you have lower values (even more – extreme lower values) than higher ones, that is going to affect the mean and it is going to pull the whole distribution to the left (hence a longer left tail).
Skewness (positive)	Skewness where the right tail of the distribution is longer than its left tail. In the distribution, you have lower and higher values. If you have higher values (even more – extreme higher values) than lower ones, that is going to affect the mean and it is going to pull the whole distribution to the right towards those high values (hence a longer right tail).
Standard deviation	It measures the variability of the data from the mean.
Statistic	Numbers from the sample (sample's mean, median…)
Variance	A measure of how much the values in a dataset vary from the mean. It is the average of the squared differences from the mean.
z-score	The number of standard deviations a data point is from the mean. It is used to standardize scores on different scales.