Statistics for Data Science and Business Analysis

Glossary

Word	Definition
population	The collections of all items of interest to our study; denoted N.
sample	A subset of the population; denoted n.
parameter	A value that refers to a population. It is the opposite of statistic.
statistic	A value that refers to a sample. It is the opposite of a parameter.
random sample	A sample where each member is chosen from the population strictly by chance
representative sample	A sample taken from the population to reflect the population as a whole
variable	A characteristic of a unit which may assume more than one value. Eg. height, occupation, age etc.
type of data	A way to classify data. There are two types of data - categorical and numerical.
categorical data	A subgroup of types of data. Describes categories or groups.
numerical data	A subgroup of types of data. Represents numbers. Can be further classified into discrete and continuous.
discrete data	Data that can be counted in a finite matter. Opposite of continuous.
continuous data	Data that is 'infinite' and impossible to count. Opposite of discrete.
levels of measurement	A way to classify data. There are two levels of measurement - qualitative and quantitative, which are further classed into nominal & ordinal, and ratio & interval, respectively.
qualitative data	A subgroup of levels of measurement. There are two types of qualitative data - nominal and ordinal.
quantitative data	A subgroup of levels of measurement. There are two types of quantitative data - ratio and interval.
nominal	Refers to variables that describe different categories and cannot be put in any order.
ordinal	Refers to variables that describe different categories, but can be ordered.
ratio	A number that has a unique and unambiguous zero point, no matter if a whole number or a fraction
interval	An interval variable represents a number or an interval. There isn't a unique and unambiguous zero point. For example, degrees in Celsius and Fahrenheit are interval variables, while Kelvin is a ratio variable.
frequency distribution table	A table that represents the frequency of each variable.
frequency	Measures the occurrence of a variable.
absolute frequency	Measures the NUMBER of occurrences of a variable.
relative frequency	Measures the RELATIVE NUMBER of occurrences of a variable. Usually, expressed in percentages.
cumulative frequency	The sum of relative frequencies so far. The cumulative frequency of all members is 100% or 1.
Pareto diagram	A type of bar chart where frequencies are shown in descending order. There is an additional line on the chart, showing the cumulative frequency.
histogram	A type of bar chart that represents numerical data. It is divided into intervals (or bins) that are not overlapping and span from the first observation to the last. The intervals (bins) are adjacent - where one stops, the other starts.
bins (histogram)	The intervals that are represented in a histogram.
cross table	A table which represents categorical data. On one axis we have the categories, and on the other - their frequencies. It can be built with absolute or relative frequencies.
contigency table	See cross table.
scatter plot	A plot that represents numerical data. Graphically, each observation looks like a point on the scatter plot.

	measures of central tendency	Measures that describe the data through 'averages'. The most common are the mean, median and mode. There is also geometric mean, harmonic mean, weighted-average mean, etc.
	mean	The simple average of the dataset. Denoted μ.
	median	The middle number in an ordered dataset.
	mode	The value that occurs most often. A dataset can have 0, 1 or multiple modes.
	measures of asymmetry	Measures that describe the data through the level of symmetry that is observed. The most common are skewness and kurtosis.
	skewness	A measure that describes the dataset's symmetry around its mean.
	sample formula	A formula that is calculated on a sample. The value obtained is a statistic.
	population formula	A formula that is calculated on a population. The value obtained is a parameter.
	measures of variability	Measures that describe the data through the level of dispersion (variability). The most common ones are variance and standard deviation.
	variance	Measures the dispersion of the dataset around its mean. It is measured in units squared. Denoted σ2 for a population and s2 for a sample.
	standard deviation	Measures the dispersion of the dataset around its mean. It is measured in original units. It is equal to the square root of the variance. Denoted σ for a population and s for a sample.
	coefficient of variation	Measures the dispersion of the dataset around its mean. It is also called 'relative standard deviation'. It is useful for comparing different datasets in terms of variability.
	univariate measure	A measure which refers to a single variable.
	multivariate measure	A measure which refers to multiple variables.
	covariance	A measure of relationship between two variables. Usually, because of its scale of measurement, covariance is not directly interpretable. Denoted σxy for a population and sxy for a sample.
	linear correlation coefficient	A measure of relationship between two variables. Very useful for direct interpretation as it takes on values from [-1,1]. Denoted ρxy for a population and rxy for a sample.
	correlation	A measure of the relationship between two variables. There are several ways to compute it, the most common being the linear correlation coefficient.
	distribution	A function that shows the possible values for a variable and the probability of their occurrence.
	Bell curve	A common name for the normal distribution.
	Gaussian distribution	The original name of the normal distribution. Named after the famous mathematician Gauss, who was the first to explore it through his work on the Gaussian function.
	to control for the mean/std/etc	While holding a particular value constant, we change the other variables and observe the effect.
	standard normal distribution	A normal distribution with a mean of 0, and a standard deviation of 1
	z-statistic	The statistic associated with the normal distribution
	standardized variable	A variable which has been standardized using the z-score formula - by first subtracting the mean and then dividing by the standard deviation
	central limit theorem	No matter the distribution of the underlying dataset, the sampling distribution of the means of the dataset approximate a normal distribution.
	sampling distribution	the distribution of a sample.
	standard error	the standard error is the standard deviation of the sampling distribution. It takes the size of the sample into account
	estimator	Estimations we make according to a function or rule
	estimate	The particular value that was estimated through an estimator.
	bias	An unbiased estimator has an expected value the population parameter. A biased one has an expected value different from the population parameter. The bias is the deviation from the true value.

	efficiency (in estimators)	in the context of estimators, the efficiency loosely refers to 'lack of variability'. The most efficient estimator is the one with the least variability. It is a comparative measure, e.g. one estimator is more efficient than another.
	point estimator	A function or a rule, according to which we make estimations that will result in a single number.
	point estimate	A single number that is derived from a certain point estimator.
	interval estimator	A function or a rule, according to which we make estimations that will result in an interval. In this course, we will only consider confidence intervals. Another instance that we don't discuss are also credible intervals (Bayesian statistics).
	interval estimate	A particular result that was obtained from an interval estimator. It is an interval.
	confidence interval	A confidence interval is the range within which you expect the population parameter to be. You have a certain probability of it being correct, equal to the significance level.
	reliability factor	A value from a z-table, t-table, etc. that is associated with our test.
	level of confidence	Shows in what % of cases we expect the population parameter to fall into the confidence interval we obtained. Denoted 1 - α. Example: 95% confidence level means that in 95% of the cases, the population parameter will fall into the specified interval.
	critical value	A value coming from a table for a specific statistic (z, t, F, etc.) associated with the probability (α) that the researcher has chosen.
	z-table	A table associated with the Z-statistic, where given a probability (α), we can see the value of the standardized variable, following the standard normal distribution.
	t-statistic	A statistic that is generally associated with the Student's T distribution, in the same way the z-statistic is associated with the normal distribution.
	a rule of thumb	A principle which is approximately true and is widely used in practice due to its simplicity.
	t-table	A table associated with the t-statistic, where given a probability (α), and certain degrees of freedom, we can check the reliability factor.
	degrees of freedom	The number of variables in the final calculation that are free to vary.
	margin of error	Half the width of a confidence interval. It drives the width of the interval.
	hypothesis	Loosely, a hypothesis is 'an idea that can be tested'
	hypothesis test	A test that is conducted in order to verify if a hypothesis is true or false.
	null hypothesis	The null hypothesis is the one to be tested. Whenever we are conducting a test, we are trying to reject the null hypothesis.
	alternative hypothesis	The alternative hypothesis is the opposite of the null. It is usually the opinion of the researcher, as he is trying to reject the null hypothesis and thus accept the alternative one.
	to accept a hypothesis	The statistical evidence shows that the hypothesis is likely to be true.
	to reject a hypothesis	The statistical evidence shows that the hypothesis is likely to be false.
	one-tailed (one-sided) test	Tests which determine if a value is lower (or equal) or higher (or equal) to a certain value are one-sided. This is because they can only be rejected on one side.
	two-tailed (two-sided) test	Tests which determine if a value is equal (or different) to a certain value are two-sided. This is because they can be rejected on two sides - if the parameter is too big or too small.
	significance level	The probability of rejecting the null hypothesis, if it is true. Denoted α. You choose the significance level. All else equal, the lower the level, the better the test.
	rejection region	The part of the distribution, for which we would reject the null hypothesis.
	type I error (false positive)	This error consists of rejecting a null hypothesis that is true. The probability of committing it is α, the significance level.
	type II error (false negative)	This error consists of accepting a null hypothesis that is false. The probability of committing it is β.
	power of the test	Probability of rejecting a null hypothesis that is false (the researcher's goal). Denoted by 1- β.
	z-score	The standardized variable associated with the dataset we are testing. It is observed in the table with an α equal to the level of significance of the test.

	μ0	The hypothesized population mean.
	p-value	The smallest level of significance at which we can still reject the null hypothesis given the observed sample statistic.
	email open rate	A measure of how many people on an email list actually open the emails they have received.
	causation	Causation refers to a causal relationship between two variables. When one variable changes, the other changes accordingly. When we have causality, variable A affects variable B, but it is not required that B causes a change in A.
	GDP	Gross domestic product is a monetary measure of the market value of all final goods and services produced for a specific country for a period.
	regression analysis	A statistical process for estimating relationships between variables. Usually, it is used for building predictive models.
	linear regression model	A linear approximation of a causal relationship between two or more variables.
	dependent variable (ŷ)	The variable that is going to be predicted. It also 'depends' on the other variables. Usually, denoted y.
	independent variable (xi)	A variable that is going to predict. It is the observed data (your sample data). Usually, denoted x1, x2 to xk.
	coefficient (βi)	A numerical or constant quantity placed before and multiplying the variable in an algebraic expression.
	constant (βo)	This is a constant value, which does not affect any independent variable, but affects the dependent one in a constant manner.
	epsilon (ε)	The error of prediction. Difference between the observed value and the (unobservable) true value.
	regression equation	An equation, where the coefficients are estimated from the sample data. Think of it as an estimator of the linear regression model
	b0, b1,..., bk	Estimates of the coefficients βo, β1, … βk.
	regression line	The best-fitting line through the data points.
	residual (e)	Difference between the observed value and the estimated value by the regression line. Point estimate of the error (ε).
	b0	The intercept of the regression line with the y-axis for a simple linear regression.
	b1	The slope of the regression line for a simple linear regression.
	SAT	The SAT is a standardized test for college admission in the US.
	GPA	Grade point average
	ANOVA	Abbreviation of 'analysis of variance'. A statistical framework for analyzing variance of means.
	SST	Sum of squares total. SST is the squared differences between the observed dependent variable and its mean.
	SSR	Sum of squares regression. SSR is the sum of the differences between the predicted value and the mean of the dependent variable. This is the variability explained by our model.
	SSE	Sum of squares error. SSE is the sum of the differences between the observed value and the predicted value. This is the variability that is NOT explained by our model.
	r-squared (R2)	A measure ranging from 0 to 1 that shows how much of the total variability of the dataset is explained by our regression model.
	OLS	An abbreviation of 'ordinary least squares'. It is a method for estimation of the regression equation coefficients.
	regression tables	In this context, they refer to the tables that are going to be created after you use a software to determine your regression equation.
	multivariate linear regression	Also known as multiple linear regression. There is a slight difference between the two, but are generally used interchangeably. In this course, it refers to a linear regression with more than one independent variable.
	adjusted r-squared	A measure, based on the idea of R-squared, which penalizes the excessive use of independent variables.

	F-statistic	The F-statistic is connected with the F-distribution in the same way the z-statistic is related to the Normal distribution.
	F-test	A test for the overall significance of the model.
	assumptions	When performing linear regression analysis, there are several assumptions about your data. They are known as the linear regression assumptions.
	linearity	Refers to linear.
	homoscedasticity	Literally means the same variance.
	endogeneity	In statistics refers to a situation, where an independent variable is correlated with the error term.
	autocorrelation	When different error terms in the same model are correlated to each other.
	multicollinearity	Refers to high correlation.
	omitted variable bias	A bias to the error term, which is introduced when you forget to include an important variable in your model.
	heteroscedasticity	Literally means a different variance. Opposite of homoscedasticity.
	log transformation	A transformation of a variable(s) in your model, where you substitute that variable(s) with its logarithm.
	semi-log model	One part of the model is log, the other is not.
	log-log model	Both parts of the model are logarithmical.
	serial correlation	Autocorrelation.
	cross-sectional data	Data taken at one moment in time.
	time series data	A type of panel data. Usually, time series is a sequence taken at successive, equally spaced points in time, e.g. stock prices.
	day of the week effect	A well-known phenomenon in finance. Consists in disproportionately high returns on Fridays and low returns on Mondays.