Click Here

Statistics formula cheat sheet

Loading... Configure your source and data warehouse in just a few clicks. Set up and maintain your data pipelines without writing a single line of code. Get analytics-ready data in your fingertips. Configure your source and data warehouse in just a few clicks. Set up and maintain your data pipelines without writing a single line of code. Get analytics-ready data in your fingertips. Use Python scripting, dbt models and a low-code GUI to craft precise transformations that ensure your data is always query-ready at the destination. Use Python scripting, dbt models and a low-code GUI to craft precise transformations that ensure your data is always query-ready at the destination. Automatically handle schema drifts and intelligently recover record failures. Get proactive alerts on changes. No manual intervention needed. Automatically handle schema drifts and intelligently recover record failures. Get proactive alerts on changes. No manual intervention needed. Share — copy and redistribute the material in any medium or format for any purpose, even commercially. Adapt — remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms. Attribution — You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits. You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation . No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material. This free statistics formula sheet PDF is a handy reference guide for students, teachers, and professionals who need quick access to essential statistical formulas and equations. Whether you're preparing for an exam or completing a project, this one-page statistics formulas sheet includes commonly used statistical calculations in probability, descriptive stats, and inferential stats. The formulas are presented in a clean, easy-to-read format to help reinforce key concepts and streamline your workflow. Great for printing or viewing on any device. Topics covered include: Statistical formulas PDF for mean, median, mode, variance, and standard deviation Common statistics equations used in basic and advanced stats courses Formulas for probability, Z-scores, correlation, and regression Quick-reference statistics formula sheet for study and review View the sheet directly below or download the PDF for offline access. Download Statistics Formula Sheet (PDF) Download Statistics Formula Sheet (PDF) Need help solving these? Try our Statistical Calculator. Related worksheets: Population Variance and Standard Deviation Worksheet Standard Deviation Practice Worksheet Scatter Plots with a Graphing Calculator – PDF Worksheet Skip to content Optimize your data integration with Hevo! Statistics is both an interesting and a challenging subject to master. So even if you've learned statistics concepts, it can often be helpful to have cheat sheets for quick reference. Cheat sheets can especially be helpful when you're preparing for upcoming exams and interviews or when you need to refresh certain stats concepts before proceeding with statistical analysis. This article is a collection of statistics cheat sheets covering a range of topics from descriptive statistics to hypothesis testing and more. 1. Statistics Cheat Sheet – Stanford University The Statistics Cheat Sheet from Stanford university is a comprehensive reference that covers all the essential stats concepts. This was originally created by the TA team for the CME 106 – Introduction to Probability and Statistics for Engineers course at Stanford. The cheat sheet covers topics such as parameter estimation, confidence intervals, hypothesis testing, regression analysis, and correlation analysis. This should be a concise reference with just enough theory and math to refresh what you've already learned. 2. Probability Cheat Sheet – Stanford University Probability and Statistics go hand in hand. A solid foundation in probability is essential to become proficient in stats—from understanding the population you're working with, effectively sampling from distributions, and more. Also created for the CME 106 course at Stanford, the Probability Cheat Sheet is a quick reference for introductory concepts in probability and random variables. The topics covered include introduction to probability and combinatorics, conditional probability, random variables, expectation and moments, probability distributions, and joint random variables. 3. Statistics Cheat Sheet – MIT If you're looking for a single concise reference to statistics, then the Statistics Cheat Sheet from MIT is for you. This covers basic to intermediate statistics concepts starting from the ground up—from the definition such as population and sample. You can use this cheat sheet to review basic definitions as well as statistical tests—T-test and Chi-Square test—with simple examples. The cheat sheet also covers the basics of probability and random variables. 4. Descriptive Statistics – DataCamp If you've worked with real-world datasets, you've likely used descriptive statistics during exploratory data analysis to understand the data better. The Descriptive Statistics Cheat Sheet from DataCamp is a refresher on basic descriptive statistics when analyzing data. This cheat sheet covers categorical and numerical data, visualizing categorical data and measures to understand numerical data. Further it also covers correlation between variables and interpretation of correlation scores. 5. Hypothesis Testing Cheat Sheet – University of New Mexico Hypothesis testing is a super important concept in statistics but it can also be tricky. The Hypothesis Testing Cheat Sheet by the University of New Mexico is a concise reference. This cheat sheet covers the preliminary concepts you need to know for hypothesis testing. It then covers topics such as steps in significance testing, choosing the right statistical test, and a summary of different statistical tests. Wrapping Up I hope you found this collection of quick reference stats cheat sheets helpful. But remember they're only to refresh essential concepts that you've already learned and understood—and not a substitute for dedicated learning. If you're interested in university stats courses to level up, 5 Free University Courses to Learn Statistics. Whether you're studying for an exam or just want to make sense of data around you every day, knowing how and when to use data analysis techniques and formulas of statistics will help.Being able to make the connections between those statistical techniques and formulas is perhaps even more important. It builds confidence when tackling statistical problems and solidifies your strategies for completing statistical projects.After data has been collected, the first step in analyzing it is to crunch out some descriptive statistics to get a feeling for the data. For example: Where is the center of the data located? How spread out is the data? How correlated are the data from two variables? The most common descriptive statistics are in the following table, along with their formulas and a short description of what each one measures. When designing a study, the sample size is an important consideration because the larger the sample size, the more data you have, and the more precise your results will be (assuming high-quality data). If you know the level of precision you want (that is, your desired margin of error), you can calculate the sample size needed to achieve it. To find the sample size needed to estimate a population mean (μ), use the following formula: In this formula, MOE represents the desired margin of error (which you set ahead of time), and σ represents the population standard deviation. If σ is unknown, you can estimate it with the sample standard deviation, s, from a pilot study; z* is the critical value for the confidence level you need. In statistics, a confidence interval is an educated guess about some characteristic of the population. A confidence interval contains an initial estimate plus or minus a margin of error (the amount by which you expect your results to vary, if a different sample were taken). The following table shows formulas for the components of the most common confidence intervals and keys for when to use them. Critical values (z*-values) are an important component of confidence intervals (the statistical technique for estimating population parameters). The z*-value, which appears in the margin of error formula, measures the number of standard errors to be added and subtracted in order to achieve your desired confidence level (the percentage confidence you want). The following table shows common confidence levels and their corresponding z*-values. Confidence Level z*– value 80% 1.28 85% 1.44 90% 1.64 95% 1.96 98% 2.33 99% 2.58 You use hypothesis tests to challenge whether some claim about a population is true (for example, a claim that 40 percent of Americans own a cellphone). To test a statistical hypothesis, you take a sample, collect data, form a statistic, standardize it to form a test statistic (so it can be interpreted on a standard scale), and decide whether the test statistic refutes the claim. The following table lays out the important details for hypothesis tests. In the field of data science, statistics serves as the backbone, providing the essential tools and techniques for extracting meaningful insights from data. Understanding statistics is imperative for any data scientist, as it equips them with the necessary skills to make informed decisions, derive accurate predictions, and uncover hidden patterns within vast datasets.This article explains the significance of statistics in data science, exploring its fundamental concepts and real-life applications. What are Statistics?Statistics is a branch of mathematics that is responsible for collecting, analyzing, interpreting, and presenting numerical data. It encompasses a wide array of methods and techniques used to summarize and make sense of complex datasets.Key concepts in statistics include descriptive statistics, which involve summarizing and presenting data in a meaningful way, and inferential statistics, which allow us to make predictions or inferences about a population based on a sample of data. Probability theory, hypothesis testing, regression analysis, and Bayesian methods are among the many branches of statistics that find applications in data science.Types of StatisticsThere are commonly two types of statistics, which are discussed below:Descriptive Statistics -Descriptive Statistics Descriptive statistics are tools that help us simplify and organize large chunks of data, making vast amounts of information easier to understand.Inferential Statistics - Inferential StatisticsInferential statistics are techniques that allow us to make generalizations and predictions about a population based on a sample of data. They help us draw conclusions and make inferences about the larger group from which the sample was taken..Descriptive StatisticsMeasure of Central Tendency Mean Median Mode Mean is calculated by summing all values present in the sample divided by total number of values present in the sample. Mean (\mu) = \frac{Sum \, of \, Values}{Number \, of \, Values} Median is the middle of a sample when arranged from lowest to highest or highest to lowest. in order to find the median, the data must be sorted.For odd number of data points: Median = (\frac{n+1}{2})^{th} For even number of data points: Median = Average \, of \, (\frac{n}{2})^{th} value \, and \, its \, next \, value Mode is the most frequently occurring value in the dataset.Each of these measures offers distinct perspectives on the central tendency of a dataset. The mean is influenced by extreme values (outliers), whereas the median is more resilient when outliers are present. The mode is valuable for pinpointing the most frequently occurring value(s) in a dataset.Measure of DispersionRange Mean Absolute Deviation Standard Deviation and VarianceInterquartile Range (IQR)Coefficient of Variation (CV)Z-scoreRange is the difference between the maximum and minimum values of the sample. Mean Absolute Deviation is the average of the absolute differences between each data point and the mean. It provides a measure of the average deviation from the mean.For Mean Absolute Deviation (MAD), the formula is:\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} \left| X_i - \bar{X} \right| Where:X_i are the individual Points.\bar{X} is the mean of data points.n is the number of data points.Standard Deviation is the square root of variance. The measuring unit of S.D. is same as the Sample values' unit. It indicates the average distance of data points from the mean and is widely used due to its intuitive interpretation.\sigma=\sqrt{\sigma^2}=\sqrt{\frac{\Sigma(X-\mu)^2}{n}) Variance is a measure of how spread-out values from the mean by measuring the dispersion around the mean.\sigma^2=\frac{\Sigma(X-\mu)^2}{n} Interquartile Range (IQR) is the range between the first quartile (Q1) and the third quartile (Q3). It is less sensitive to extreme values than the range.IQR = Q_3 -Q_1 To compute IQR, calculate the values of the first and third quartile by arranging the data in ascending order. Then, calculate the mean of each half of the dataset.Coefficient of Variation (CV) is the ratio of the standard deviation to the mean, expressed as a percentage. It is useful for comparing the relative variability of different datasets.CV = (\frac{\sigma}{\mu}) * 100 The Z-score measures the number of standard deviations a data point is from the mean of a dataset, providing a standardized way to assess its relative position within the distribution. Z=\frac{X-\mu}{\sigma} Quartiles divides the dataset into four equal parts:Q1 is the median of the lower 25%Q2 is the median (50%)Q3 is the median of the upper 25% of the dataset.Measure of ShapeKurtosisKurtosis is a statistical measure that describes the shape of a distribution's tails in relation to its overall shape. It indicates whether data points are more or less concentrated in the tails compared to a normal distribution. High kurtosis signifies heavy tails and possibly outliers, while low kurtosis indicates light tails and a lack of outliers.Types of KurtosisSkewnessSkewness is the measure of asymmetry of probability distribution about its mean.Skewness is a statistical measure that describes the asymmetry of a distribution around its mean. A distribution can be:Positively skewed (right-skewed): The right tail (higher values) is longer or fatter than the left tail. Most of the data points are concentrated on the left, with a few larger values extending to the right.Negatively skewed (left-skewed): The left tail (lower values) is longer or fatter than the right tail. Most of the data points are concentrated on the right, with a few smaller values extending to the left.Types of Skewness Right Skew:Also known as positive skewness.Characteristics:Longer or fatter tail on the right-hand side (upper tail).More extreme values on the right side.Mean > Median.Indicates a distribution that is skewed towards the left.Left Skew:Also known as negative skewness.Characteristics:Longer or fatter tail on the left-hand side (lower tail).More extreme values on the left side.Mean < Median.Indicates a distribution that is skewed towards the right.Zero Skew:Also known as symmetrical distribution.Characteristics:Symmetric distribution.Left and right sides are mirror images of each other.Mean = Median.Indicates a distribution with no skewness.Types of Skewed dataCovariance and CorrelationCovariance Correlation Covariance measures the degree to which two variables change together.Cov(x,y) = \frac{\sum(X_i-\overline{X})(Y_i - \overline{Y})}{n} Correlation measures the strength and direction of the linear relationship between two variables. It is represented by correlation coefficient which ranges from -1 to 1. A positive correlation indicates a direct relationship, while a negative correlation implies an inverse relationship. Pearson's correlation coefficient is given by:\rho(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} Regression coefficientRegression coefficient is a value that represents the relationship between a predictor variable and the response variable in a regression model. It quantifies the change in the response variable for a one-unit change in the predictor variable, holding all other predictors constant. In a simple linear regression model, the regression coefficient indicates the slope of the line that best fits the data. In multiple regression, each coefficient represents the impact of one predictor variable while accounting for the effects of other variables in the model.The equation is y=\alpha+ \beta * x , wherey is the dependent variable,x is the independent variable,\alpha is the intercept,\beta is the regression coefficient.Regression coefficient, \beta = \frac{\sum(X_i-\overline{X})(Y_i - \overline{Y})}{\sum(X_i-\overline{X})^2} Probability Probability FunctionsProbability Mass Functions Probability Density FunctionProbability Mass Function is a concept in probability theory that describes the probability distribution of a discrete random variable. The PMF gives the probability of each possible outcome of a discrete random variable.The Probability Density Function describes the likelihood of a continuous random variable falling within a particular range. It's the derivative of the cumulative distribution function (CDF).Cumulative Distribution FunctionEmpirical Distribution FunctionThe Cumulative Distribution Function gives the probability that a random variable will take a value less than or equal to a given value. It's the integral of the probability density function (PDF).The Empirical Distribution Function is a non-parametric estimator of the cumulative distribution function (CDF) based on observed data. For a given set of data points, the EDF represents the proportion of observations less than or equal to a specific value. It is constructed by sorting the data and assigning a cumulative probability to each data point.Bayes TheoremBayes' Theorem is a fundamental principle in probability theory and statistics that describes how to update the probability of a hypothesis based on new evidence. The formula is as follows:P(A|B)=P(B)P(B|A)·P(A), whereP(A|B): The probability of event A given that event B has occurred (posterior probability).P(B|A): The probability of event B given that event A has occurred (likelihood).P(A): The probability of event A occurring (prior probability).P(B): The probability of event B occurring. Probability DistributionsDiscrete Distribution Uniform Distribution Binomial DistributionPoisson DistributionThe uniform distribution represents a constant probability for all outcomes in a given range.f(X)=\frac{1}{b-a} For the same previous dataset, assuming the bus arrives uniformly between 5 and 18 minutes so the probability of waiting less than 15 minutes: P(X