# Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 3

Ritesh Gupta · Follow
15 min read · Mar 17



**Statistics** is a branch of mathematics that deals with the collection, analysis, interpretation, and presentation of numerical data. In data science, statistics is used to extract insights and meaningful information from large amounts of data.

Credit: vecteezy

*In this article, we will cover Advance statistics. If you missed Part 1 & Part 2, you can find it here.*

*Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 1*

*Mastering the Fundamentals of Statistics for Data Science -Basic to Advance Level- Part 2*

## QQ plot

A QQ plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a sample with a theoretical distribution. It is a type of probability plot that can help to assess whether a sample of data comes from a specific distribution, such as a normal distribution, or to identify any differences between two distributions.

In a QQ plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the sample data follow the theoretical distribution, the points on the plot should fall along a straight line. However, if the sample data differ from the theoretical distribution, the points on the plot will deviate from the straight line.

QQ plots are particularly useful for testing the assumption of normality in statistical analyses, as they can help to identify whether the sample data is normally distributed or not. They can also be used to compare other types of distributions, such as exponential, uniform, or Poisson distributions, to a sample of data.

### Bernoulli Distribution And Binomial Distribution

Binomial Distribution is also a discrete probability distribution that models the probability of a series of independent binary events with only two possible outcomes. It describes the number of successes in a fixed number of trials. The distribution is characterized by two parameters: n, the number of trials, and p, the probability of success in each trial. The distribution function is as follows:

P(X=x) = nCx * p^x * (1-p)^(n-x)

where X is the random variable that represents the number of successes in n trials, and nCx represents the number of ways to choose x items from a set of n items.

In summary, the Bernoulli Distribution is used to model a single binary event, while the Binomial Distribution is used to model the probability of a series of independent binary events with a fixed number of trials.

## Log Normal Distribution

A log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. In other words, if you take the natural logarithm of a random variable X and the result is normally distributed, then X is said to have a log-normal distribution.

The log-normal distribution is often used to model the behavior of variables that are only positive, such as income, stock prices, and other financial variables. The distribution has a positive skewness, meaning that it is skewed to the right and has a long tail.

The probability density function (PDF) of a log-normal distribution is given by:

$$f(x) = (1 / (x * \sigma * \sqrt{2 * \pi})) * \exp(-((\ln(x) - \mu)^2) / (2 * \sigma^2))$$

where x is the random variable, mu is the mean of the logarithm of x, sigma is the standard deviation of the logarithm of x, and pi is the mathematical constant pi.

The cumulative distribution function (CDF) of a log-normal distribution is not available in a closed form, but it can be approximated using numerical methods.

## Power Law Distribution

A power law distribution is a type of probability distribution that describes a relationship between two variables in which one variable is proportional to a power of the other. The power law distribution is also known as a Pareto distribution or a long-tailed distribution.

In a power law distribution, the frequency of an event is proportional to its magnitude raised to a negative power. This means that a few large events are much more common than many small events, and the distribution has a long tail that extends to infinity. Power law distributions are used to describe a wide range of phenomena, such as the distribution of city sizes, the number of citations for scientific papers, and the popularity of websites.

The probability density function (PDF) of a power law distribution is given by:

f(x) = (alpha-1) * x^(-alpha)

where x is the random variable, and alpha is the exponent of the power law distribution. The exponent alpha is usually greater than 1, and it determines the shape of the distribution.

The cumulative distribution function (CDF) of a power law distribution is given by:

F(x) = 1 — (x/x_min)^(-alpha+1)

where x_min is the minimum value of x in the distribution. The CDF is useful for calculating the probability that a random variable is less than or equal to a given value.

Power law distributions are important in many fields, including economics, physics, and computer science, because they describe the behavior of complex systems that exhibit scale invariance, meaning that their properties are independent of scale.

## Boxcox Tranform

The Box-Cox transform is a mathematical transformation that is commonly used to transform non-normal data into approximately normal data. The transform is named after statisticians George Box and David Cox, who developed it in 1964.

The Box-Cox transform involves applying a power transformation to the data, where the power parameter, λ, is estimated from the data. The transformation can be written as:

y(lambda) = (x(lambda)^(lambda) — 1) / lambda

where x is the original data and y is the transformed data. The parameter λ can take on any value, including negative values, although values close to zero are typically avoided.

The Box-Cox transform is often used in statistics to normalize the distribution of data so that it meets the assumptions of certain statistical

models, such as linear regression. It can also be used to stabilize the variance of the data.

The implementation of the Box-Cox transform can vary depending on the software or programming language used, but there are many packages and functions available that can apply the transformation automatically.

## All Transformation Techniques

Transformation techniques in statistics refer to the process of applying a mathematical function to a variable or set of variables to change their distribution or shape. This can be helpful in various statistical analyses, such as hypothesis testing, data visualization, and regression analysis. Some of the most common transformation techniques used in statistics are:

1. Log transformation: Taking the logarithm of a variable can be helpful in dealing with variables that have skewed or exponentially distributed data. This can make it easier to identify patterns and relationships in the data.

2. Square root transformation: Similar to the log transformation, taking the square root of a variable can be useful in dealing with skewed data. This transformation can also make it easier to visualize the data.

3. Box-Cox transformation: This is a family of power transformations that can be applied to data with various distributions. The Box-Cox transformation is a method for finding the most appropriate transformation parameter for a given dataset.

4. Z-score transformation: This transformation involves subtracting the mean of a variable from each observation and then dividing by the standard deviation. This results in a new variable with a mean of zero and a standard deviation of one.

5. Min-Max normalization: This transformation involves scaling the data to a range of 0 to 1 by subtracting the minimum value from each observation and then dividing by the range of the variable.

6. Quantile normalization: This transformation involves matching the distribution of a variable to a standard normal distribution by ranking the observations and assigning them new values based on their rank.

7. Fourier transformation: This is a technique used to transform time series data into frequency domain data. The Fourier transformation can be helpful

in identifying periodic patterns in data.

8. Principal Component Analysis (PCA): PCA is a technique used to transform high-dimensional data into a lower-dimensional space. The transformation involves finding the principal components of the data that capture the most significant variation in the data.

These are some of the most commonly used transformation techniques in statistics, but there are many others that can be used depending on the nature of the data and the analysis being performed.

## Confidence Interval In statistics

A confidence interval is a range of values that is likely to contain the true value of a population parameter, such as the mean or proportion, based on a sample of data. It is a common statistical tool used in inferential statistics to estimate the precision or accuracy of a sample statistic and to make inferences about the population.

A confidence interval is calculated using a point estimate of the population parameter, such as the sample mean or proportion, and a margin of error that reflects the variability of the estimate. The level of confidence is typically set at 90%, 95%, or 99%, and represents the probability that the true population parameter falls within the confidence interval.

For example, a 95% confidence interval for the mean height of a population might be calculated as follows:

• Take a random sample of n individuals from the population and measure their heights

• Calculate the sample mean height, $\bar{x}$, and the sample standard deviation

• Use a t-distribution (or z-distribution, depending on the sample size and assumptions) to calculate the margin of error, which is based on the sample size, standard deviation, and level of confidence

• Construct the confidence interval as $\bar{x} \pm$ margin of error

If we repeated this process many times, we would expect 95% of the resulting confidence intervals to contain the true population mean.

It is important to note that a confidence interval provides information about the precision or accuracy of an estimate, but does not guarantee that the true population parameter falls within the interval. It is also influenced by the sample size, level of confidence, and assumptions about the population distribution.

## Type 1 And Type 2 error

Type 1 and Type 2 errors are terms used in statistical hypothesis testing to describe the possible errors that can occur when making a decision based on the results of a statistical test.

A Type 1 error, also known as a false positive, occurs when a hypothesis is rejected even though it is actually true. In other words, a Type 1 error is the probability of concluding that there is a statistically significant effect or relationship when in fact there is none. The probability of making a Type 1 error is denoted by the symbol alpha (α), and it is usually set at 0.05 or 0.01.

A Type 2 error, also known as a false negative, occurs when a hypothesis is not rejected even though it is actually false. In other words, a Type 2 error is the probability of failing to conclude that there is a statistically significant effect or relationship when in fact there is one. The probability of making a Type 2 error is denoted by the symbol beta (β), and it is usually set at 0.2 or 0.1.

The probability of making one type of error is related to the probability of making the other type of error. In general, as the probability of making a Type 1 error decreases, the probability of making a Type 2 error increases, and vice versa. Therefore, the choice of significance level (alpha) and sample size in a statistical test should be carefully considered in order to minimize the likelihood of both types of errors.

## One-Tailed And 2 Tailed Tests

One-tailed and two-tailed tests are types of statistical hypothesis tests that are used to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

A one-tailed test, also known as a directional test, is a statistical test that only examines one direction of an effect, either positive or negative. In other words, it tests whether a specific relationship or effect exists in a particular direction. For example, a one-tailed test might be used to determine whether a new drug is better than an existing drug, with the hypothesis being that the

new drug is better. The alternative hypothesis in this case would be that the new drug is better than the existing drug, and the null hypothesis would be that there is no difference between the two drugs.

A two-tailed test, also known as a non-directional test, is a statistical test that examines both directions of an effect, both positive and negative. It tests whether a specific relationship or effect exists, without specifying a particular direction. For example, a two-tailed test might be used to determine whether a new drug is different from an existing drug, without specifying whether the new drug is better or worse. The alternative hypothesis in this case would be that the new drug is different from the existing drug, and the null hypothesis would be that there is no difference between the two drugs.

The choice between a one-tailed and a two-tailed test depends on the research question and the directionality of the hypothesis. One-tailed tests are appropriate when the researcher has a clear directional prediction about the outcome, whereas two-tailed tests are appropriate when the researcher does not have a directional prediction.

## Hypothesis Testing

Hypothesis testing is a statistical method used to determine whether a hypothesis about a population is likely to be true or false based on a sample of data. It involves formulating two hypotheses: the null hypothesis (H0) and the alternative hypothesis (Ha).

The null hypothesis is the default assumption that there is no significant difference or relationship between the population parameters being studied. The alternative hypothesis, on the other hand, is the statement that contradicts the null hypothesis and suggests that there is a significant difference or relationship between the parameters being studied.

Hypothesis testing involves collecting sample data, calculating a test statistic, and determining the probability of obtaining the observed test statistic assuming the null hypothesis is true. This probability is called the p-value, and it represents the likelihood of observing the test statistic or a more extreme value if the null hypothesis is true.

If the p-value is below a predetermined significance level (usually 0.05), the null hypothesis is rejected, and the alternative hypothesis is accepted. If the p-value is above the significance level, the null hypothesis is not rejected, and no significant difference or relationship is inferred.

Hypothesis testing is commonly used in scientific research, quality control, and decision-making in various fields such as business, finance, and healthcare.

## P-value

The p-value is a statistical measure that helps to determine the significance of an observed effect in a hypothesis test. In hypothesis testing, the p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed result, assuming the null hypothesis is true.

In other words, the p-value tells us the likelihood of observing the results that we have if the null hypothesis is true. The smaller the p-value, the less likely it is that the observed results occurred by chance, and the more likely it is that there is a significant difference or relationship between the variables being studied.

A p-value of 0.05 (5%) is often used as a cutoff value to determine statistical significance. If the calculated p-value is less than 0.05, it is considered statistically significant, and we reject the null hypothesis in favor of the alternative hypothesis. On the other hand, if the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is not enough evidence to support the alternative hypothesis.

It is important to note that the p-value does not tell us the size or practical significance of the effect, but only the statistical significance. Therefore, it should be interpreted in conjunction with other measures of effect size and practical significance.

## Steps For Hypothesis Testing

1. Formulate the research question and the null hypothesis: The first step is to clearly define the research question and the null hypothesis. The null hypothesis typically represents the default position or the assumption that there is no difference or relationship between the variables being studied.

2. Formulate the alternative hypothesis: The alternative hypothesis represents the opposite of the null hypothesis and states that there is a significant difference or relationship between the variables.

3. Determine the level of significance: The level of significance is the threshold at which we reject the null hypothesis. Typically, a level of significance of 0.05 (or 5%) is used.

4. Select the appropriate test statistic: The choice of test statistic depends on the research question, the type of data, and the distribution of the data.

5. Collect and analyze the data: Collect the data and calculate the test statistic, which measures the difference or relationship between the variables.

6. Calculate the p-value: The p-value is the probability of observing the test statistic or a more extreme value, assuming the null hypothesis is true.

7. Interpret the results: If the p-value is less than the level of significance, the null hypothesis is rejected, and the alternative hypothesis is accepted. If the p-value is greater than the level of significance, the null hypothesis is not rejected.

8. Draw conclusions and report the results: Based on the results, draw conclusions and report the findings, including the p-value and the effect size, if applicable. It is important to interpret the results in the context of the research question and to discuss any limitations or assumptions made in the analysis.

## T-test

A t-test is a statistical hypothesis test used to determine whether there is a significant difference between the means of two groups. It is commonly used in scientific research and data analysis to compare two groups of data and to determine whether any observed differences are due to chance or whether they are statistically significant.

There are two main types of t-tests: the independent samples t-test and the paired samples t-test. The independent samples t-test is used when the two groups being compared are independent, meaning that the data in one group has no relation to the data in the other group. The paired samples t-test is used when the two groups being compared are dependent, meaning that the data in one group is related to the data in the other group.

The t-test calculates a t-value, which is a measure of the difference between the means of the two groups, relative to the variation within each group. The t-value is then compared to a critical value based on the sample size and desired level of significance, typically 0.05. If the t-value exceeds the critical value, the null hypothesis (i.e., that there is no significant difference between the means of the two groups) is rejected in favor of the alternative

hypothesis (i.e., that there is a significant difference between the means of the two groups).

## Z- test

A z-test is a statistical test used to compare a sample mean to a population mean, when the population standard deviation is known. The z-test is based on the standard normal distribution, which is a bell-shaped distribution that has a mean of 0 and a standard deviation of 1.

In a z-test, the sample mean is standardized to the standard normal distribution by subtracting the population mean and dividing by the standard deviation of the population. The resulting value is called the z-score. The z-score is then compared to a critical value based on the desired level of significance, typically 0.05.

If the z-score exceeds the critical value, the null hypothesis (i.e., that there is no significant difference between the sample mean and the population mean) is rejected in favor of the alternative hypothesis (i.e., that there is a significant difference between the sample mean and the population mean).

Z-tests are often used when the sample size is large and the population standard deviation is known. When the population standard deviation is not known, a t-test is often used instead.

## Annova test

The ANOVA (Analysis of Variance) test is a statistical method used to analyze whether there are significant differences between the means of two or more groups. It determines whether the variation in the data is due to the differences between groups or due to random chance.

The ANOVA test involves comparing the variance between groups to the variance within groups. If the variance between groups is larger than the variance within groups, it suggests that there are significant differences between the groups being compared. In contrast, if the variance within groups is larger than the variance between groups, it suggests that there are no significant differences between the groups being compared.

The ANOVA test is commonly used in experimental and observational studies to compare means of multiple groups, such as comparing the effectiveness of different treatments on a particular disease. It can be performed using software programs such as SPSS, SAS, and R.

## Chi-square test

The Chi-square test is a statistical method used to determine whether there is a significant association or relationship between two categorical variables. It is typically used to test the hypothesis that there is no difference between the expected and observed frequencies of two or more categories.

The Chi-square test involves calculating the difference between the expected frequencies of each category and the observed frequencies, and then comparing these differences to determine if they are statistically significant. The test results in a Chi-square statistic and a p-value, which indicates the level of significance of the relationship between the variables.

The Chi-square test can be used to analyze data from a variety of studies, including surveys, experiments, and observational studies. It is commonly used in social science research, epidemiology, and genetics.

There are different types of Chi-square tests, including the Pearson's Chi-square test, which is used to analyze two or more nominal variables, and the Mantel-Haenszel Chi-square test, which is used to analyze the relationship between two nominal variables while controlling for a third variable. The Chi-square test can be performed using software programs such as SPSS, SAS, and R.

## Thanks for Reading!

*If you enjoyed this, follow me to never miss another article on data science guides, tricks and tips, life lessons, and more!*

Statistics | Data Science | Machine Learning | Mathematics | Artificial Intelligence



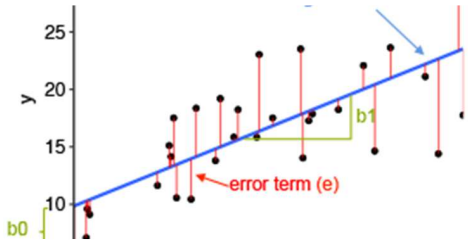
**Written by Ritesh Gupta**

1.2K Followers

Follow

Data Scientist, I write Article on Machine Learning| Deep Learning| NLP | Open CV | AI Lover ❤️

## More from Ritesh Gupta



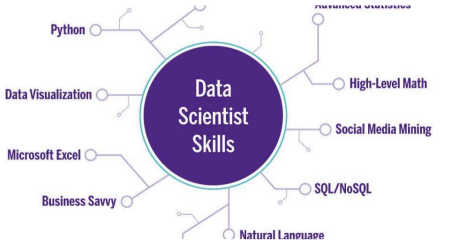

Ritesh Gupta

### 10 Most Common Machine Learning Algorithms Explained...

1. Linear Regression

20 min read · Jan 19

1.1K 21



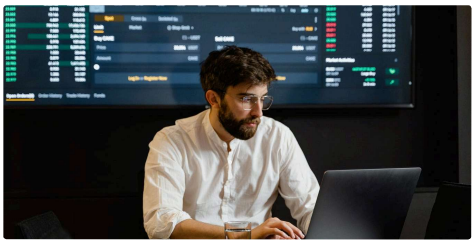

Ritesh Gup... in Artificial Intelligence in Plain Engli...

### Master Data Science with This Comprehensive Cheat Sheet

Comprehensive Cheat Sheet for Data Science: Numpy, Pandas, Python, R, ML, DL,...

6 min read · Jan 30

809 13



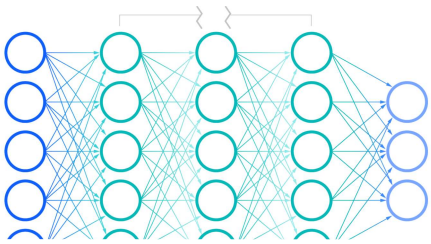
Ritesh Gupta

### Mastering the Fundamentals of Statistics for Data Science -Basic...

Statistics:

11 min read · Jan 28

435 6



Ritesh Gupta

### Mastering the Deep Learning Interview: Top 35 Questions and...

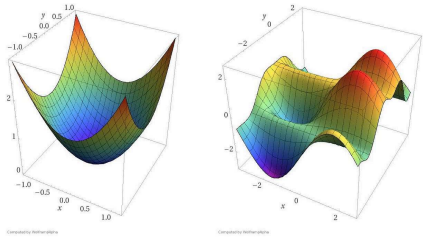
What is Deep Learning?

13 min read · Jan 23

448 9

See all from Ritesh Gupta

## Recommended from Medium



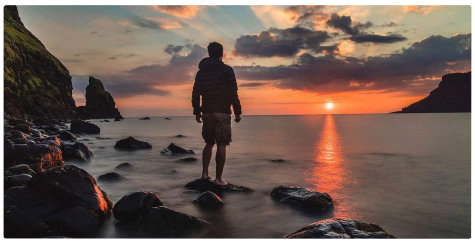
Purushottam Mitra

## Gradient Descent

All that you wanna know about the most commonly used Optimization Algorithm

4 min read · May 31

28 1



Madison Hunter in Towards Data Science

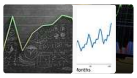
## 6 Habits to Include in Your Daily Routine for a Long, Happy Career...

All of these habits take less than 10 minutes out of your day

9 min read · Oct 31, 2022

966 10

### Lists



**Predictive Modeling w/ Python**

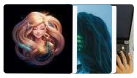
18 stories · 176 saves

**Practical Guides to Machine Learning**

10 stories · 194 saves

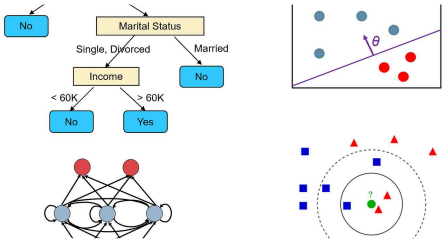
**Natural Language Processing**

444 stories · 82 saves



**ChatGPT prompts**

22 stories · 170 saves




Dr. Roi Yehoshua in Towards AI

## Which ML Algorithm to Choose?

One of the key decisions you need to make when solving a data science problem is whic...

6 min read · Mar 3

245 3

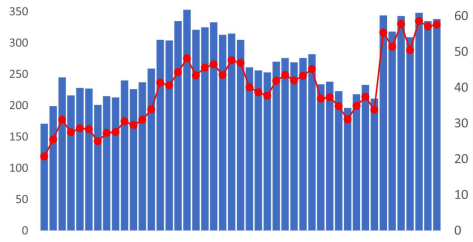
Matt Chapman in Towards Data Science

## The Portfolio that Got Me a Data Scientist Job

Spoiler alert: It was surprisingly easy (and free) to make

10 min read · Mar 24

4K 69



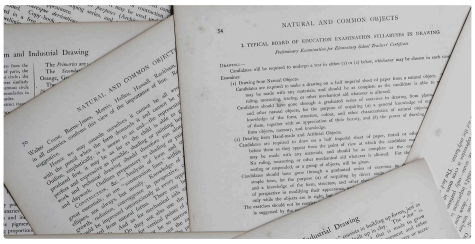

René F. Najera, MPH, DrPH

## Tell a Story Instead of Just Showing the Data

Be a storyteller, not a number cruncher.

4 min read · May 18, 2022

113



Youssef Hosni in Towards AI

## How to Read Machine Learning Papers Effectively

The field of machine and deep learning is evolving very fast, and there are new researc...

10 min read · Oct 9, 2022

1.1K 8

See more recommendations