Statistical Testing - Understanding Which Testing Methods to Use

Author, Abhishek Mishra

A Data Science Foundation White Paper

May 2020

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

Data Science, Machine Learning, Artificial Intelligence, Deep Learning - You need to learn the basics before you become a good Data Scientist. Math and Statistics are the building blocks of Algorithms for Machine Learning. Knowing the techniques behind different Machine Learning Algorithms is fundamental to knowing how and when to use them. In this paper, we will look at statistics as a concept. What are the different tests and most important, 'When to use Which?' so let us begin by learning more about statistics.

What is Statistics? As per the Oxford Dictionary definition:

"*Statistics is the practice or science of collecting and analyzing numerical data in large quantities, especially to infer proportions in a whole from those in a representative sample.*"

Statistics are used to interpret data and solve complicated real-world issues. Data scientists and analysts use statistics to search for concrete patterns and data changes. To put it simply, Statistics can be used to carry out mathematical computations to extract useful insights from data.

Key terminology in statistics

1. The Population is the collection of sources from which to gather data.

2. A Variable is any characteristic, number or quantity observable or countable.

3. A Sample is a Population subset

4. A Statistical Parameter or Population Parameter is a quantity indexing a family of distributions of probabilities

Understanding '**Statistical Hypothesis'** is quite important. The Wikipedia definition, which I think best explains it, is as follows:

A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is testable based on observing a process that is modeled via a set of random variables.

A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized

model. An alternative hypothesis is proposed for the statistical-relationship between the two datasets and is compared to an idealized null hypothesis that proposes no relationship between these two datasets. This comparison is deemed statistically significant if the relationship between the datasets would be an unlikely realization of the null hypothesis according to a threshold probability—the significance level. Hypothesis tests are used when determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

Or in other words, Hypothesis testing is a statistical approach that is used with experimental data to make statistical decisions. This is used to determine whether an experiment performed offers ample evidence to reject a proposal. Before we start to distinguish between various tests or experiments, we need to gain a clear understanding of what a null hypothesis is.

A **Null Hypothesis** implies that there is no strong difference in a given set of observations.

The basic assumption of a statistical test is called the null hypothesis, and we can quantify and interpret statistical measurements to determine if the null hypothesis should be accepted or not. We are interested to learn whether there is an actual or statistically meaningful difference between the two models when choosing models based on their estimated skills.

The comparison of machine learning models through statistical significance tests imposes some requirements which will in effect the types of statistical tests that can be used.

To make a statement on whether to deny the null hypothesis a test statistic will be determined. The decision is taken based on the test statistical quantitative value. There are two methods of how this decision can be derived:

1. Critical Value

2. p-Value

Critical Value Method

In this the idea is to find out whether or not the test statistic observed is more extreme than a given critical value. As stated in Wikipedia,

In statistical hypothesis testing, the critical values of a statistical test are the boundaries of the acceptance region of the test. The acceptance region is the set of values of the test statistic for which the null hypothesis is not rejected. Depending on the shape of the acceptance region, there can be one or more than one critical value.

The critical value separates the area in the rejection region(s) and not-rejection region within the probability distribution curve. Determination of the form of distribution is needed to determine the critical value and test to be selected to validate any hypothesis

- In a two-sided test, if the test numbers are either too small or too high the null hypothesis is rejected.
- In a left-tailed test, if the test results are too low the null hypothesis is rejected.
- In a right-tailed test, if the test numbers are too high the null hypothesis is rejected.

P-Value

In the p-value method, the likelihood of the test statistic's numerical value is compared to the hypothesis test's defined significance point. The p-value refers to the probability that sample data were observed at least as extreme as the test statistics obtained.

The lower the p-value (closer to 0), the stronger is proof against the null hypothesis.

The table below provides instructions for using the p-value:

P-Value

p-value	Evidence against H0
p>0.10	Weak or no evidence
0.05<p≤0.10	Moderate evidence
0.01<p≤0.05	Strong evidence
p≤0.01	Very strong evidence

Relation of the p-value, critical value, and test statistics - Can be defined as below:

The critical value is the point or border beyond which we dismiss the null hypothesis. On the other hand, P-value is defined as the probability of the right to the respective statistics like T, Z, or Chi. The benefit of using p-value is that it estimates a likelihood estimate, by comparing this likelihood directly with significance level, we can evaluate at any desired level of significance.

Statistical tests can be used to:

1. Determine if a predictor variable has a statistically important outcome variable.
2. Estimate disparity between groups of two or more.

Statistical tests assume a null hypothesis that there is no connection or difference between groups. They assess if the data observed falls beyond the range of values foreseen by the null hypothesis.

Statistical tests operate by measuring a test statistic – a number that explains how different the relationship between variables in your test is, from the null hypothesis of no relationship.

If the test statistics value is more severe than the statistics determined from the null hypothesis, then you can assume a statistically significant relationship between the predictor and the variables of the outcome also if the test statistic value is less extreme than the one determined from the null hypothesis, then you cannot conclude any statistically significant association between the predictor and the variables of the outcome. You may conduct statistical tests on data obtained in a statistically accurate manner – either through an experiment or observations using methods of probability sampling.

The main two criteria that need to be taken into consideration before determining which statistical test to use are:

1. If the data meets those defined assumptions
2. Type of Variable you have to do the test

On the first point, here are some common assumptions:

- Homogenous: The variation is identical across all groups within each group being compared. When one group has much more variability than others, the efficacy of the test will be reduced.
- No autocorrelation: Independence of observations. There is no relation between the observations / variables you use in your test (for example, multiple measurements of a single test subject are not independent, whereas measurements of several different test subjects are independent).
- Normality: The data follows a normal distribution, the famous bell curve.

If your data does not meet the normality or homogeneity of variance assumptions, you may be able to perform a non-parametric statistical test which enables you to make comparisons without any assumptions about the distribution of data. If your data does not fulfill the presumption that results are independent, you might be able to use a test that accounts for consistency in your data (repeated tests or tests that involve blocking variables).

Before we pe into the different tests and their scenarios. Let us also investigate a certain problem in choosing a Hypothesis test.

Evaluating Classification Methods

Evaluating classification methods using classification accuracy is standard practice, evaluation of the model is done using 10-fold cross-validation, using a Gaussian distribution. We may request that any

classifier evaluated using this procedure be evaluated through 10-fold cross-validation on the same splits of the dataset.

We should then pick and use the Student's paired t-test to verify that the difference in mean accuracy between the two models is statistically important, e.g. dismissing the null hypothesis that implies the two samples have the same distribution.

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. A t-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. Wikipedia

The problem is, a primary assumption was violated by the paired Student's t-test. This means that the predicted skill scores are biased, not independent, and in effect, the measurement of the t-statistics in the test along with any interpretations of the statistics and p-value would be misleadingly incorrect.

Unfortunately, it is more difficult to find a suitable statistical hypothesis check for model selection in applied machine learning than it first appears

The statistical testing process

Below is an extract from Wikipedia and I found it quite useable here as it clearly explains the entire process neatly:

https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

1. There is an initial research hypothesis of which the truth is unknown.
2. The first step is to state the relevant null and alternative hypotheses. This is important, as misstating the hypotheses will muddy the rest of the process.
3. The second step is to consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.
4. Decide which test is appropriate, and state the relevant test statistic T.
5. Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases, this will be a well-known result. For example, the test statistic might follow a Student's t distribution or normal distribution.
6. Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%
7. The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected—the so-called critical region—and those for which it is not. The probability of the critical region is α.
8. Compute from the observations the observed value t of the test statistic T.

9. Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H0 if the observed value t is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

Ok, now let's see what is the right test to use and under which scenario.

The most common types of parametric tests include comparison tests, regression tests, and correlation tests.

Comparison tests look for group mean differences and can be used to measure the effect on the mean value of any other attribute of a categorical variable.

T-tests are used when comparing the results for two classes in particular. ANOVA and MANOVA measures are used to equate the means of the two classes.

T-Tests

	Predictor Variable: Categorical	**Outcome Variable: Quantitative**
Paired t-test	1 predictor	Groups come from the same population
Independent t-test	1 predictor	Groups come from different populations
ANOVA	1 or more predictor	1 outcome
MANOVA	1 or more predictor	2 or more outcomes

Regression experiments are used to assess the relationships between cause and effect. They are looking for influence on another variable of one or more continuous variables.

Regression Experiments

	Predictor Variable	**Outcome Variable**
Simple Linear Regression	Continuous	Continuous
	1 predictor	1 outcome
Multiple Linear Regression	Continuous	Continuous

	Predictor Variable	Outcome Variable
	2 or more predictors	1 outcome
Logistic Regression	Continuous	Binary

Correlation tests check whether two variables are related without suggesting associations between cause and effect. This can be used to check if there is autocorrelation between two variables you choose to use in (for example) a multiple regression check.

	Predictor Variable	Outcome Variable
Pearson	Continuous	Continuous
Chi-Square	Categorical	Categorical

In a **Z-test**, the distribution of the sample is presumed to be natural. A z-score is determined using population parameters such as "population mean" and "population standard deviation" and is used to verify an inference that the sample being drawn belongs to the same population.

If the test statistics are lower than the critical value, consider the hypothesis or otherwise deny the statement

The mean of two given samples is compared with a t-test. If parameters of the population (mean and standard deviation) are not specified, a t-test is used.

There are three T-Test models:

1. Independent t-test analyses comparing mean for two groups
2. Paired sample t-test which compares means at different times from the same group
3. One sample t-test compares a single group's average against a known mean.

ANOVA also known as variance analysis, is used to equate a single study with several (three or more) samples. One-way ANOVA is used to compare the difference between a single independent variable's three or more samples/classes.

MANOVA helps one to check two or more dependent variables for the influence of one or more independent variables.

For evaluating categorical variables, the **Chi-square** method is used. It has 2 types.

1. The goodness of fit test determines whether a sample matches the population.
2. With two independent variables, a chi-square fit test is used to evaluate two variables in a contingency table to determine if the data matches.

I hope you enjoyed reading my paper and that you found it useful. In the tests we have mentioned a statistical significance to accept or deny a hypothesis is compared. The test and methods to calculate it depend on the end objective, along with the kind of variable and sample size. Based on the objective and type of variable an appropriate test is chosen along with the null hypothesis. Do let me know in comments about any specific point you would like me to discuss in detail.

Statistics is a vast topic and the statistical test is one of the key skills that most of the Data Science, Artificial Intelligence or Machine Learning Professional keeps developing throughout their career. Almost all data science projects use statistical tests and the success of many projects is dependent on the correct test being chosen.

In a future paper, I will draft a practical example-driven approach. This will help you gain some real-world experience of statistical tests and usages on the different types of variables / data. Hit Like if you liked the paper. Below given are some of the useful references. Do share your experience of using various types of statistical tests under a different scenario.

Reference:

http://www.statisticshowto.com/probability-and-statistics/chi-square/

http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP

https://www.investopedia.com/terms/n/null_hypothesis.asp

https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation