## Statistical Hypothesis Testing

A statistical hypothesis test makes an assumption about the outcome, called the null hypothesis.

For example, the null hypothesis for the Pearson’s correlation test is that there is no relationship between two variables. The null hypothesis for the Student’s t test is that there is no difference between the means of two populations.

The test is often interpreted using a p-value, which is the probability of observing the result given that the null hypothesis is true, not the reverse, as is often the case with misinterpretations.

* **p-value (p)**: Probability of obtaining a result equal to or more extreme than was observed in the data.

## P-value

The p-value in a statistical test helps you determine whether to reject or support the null hypothesis. It's a metric that argues against the null hypothesis and relies on the alpha value, critical value and probability. Measuring a smaller p-value suggests the rejection of the null hypothesis, whereas a higher p-value indicates stronger evidence for supporting the null hypothesis.

The p-value is a probability measure and uses the degree of freedom and estimation based on the alpha value of a t-test. Taking the sample size n, subtract one to get the degree of freedom (n - 1). Comparing the result to a respective alpha level gives you the estimate for the p-value. It's important to note that p-values depend on the results t-tests give you and can change according to different t-statistics.

## How to interpret P-value?

In interpreting the p-value of a significance test, you must specify a significance level, often referred to as the Greek lower case letter alpha (a). A common value for the significance level is 5% written as 0.05.

The p-value is interested in the context of the chosen significance level. A result of a significance test is claimed to be “*statistically significant*” if the p-value is less than the significance level. This means that the null hypothesis (that there is no result) is rejected.

* **p <= alpha**: reject H0, different distribution.
* **p > alpha**: fail to reject H0, same distribution.

Where:

* **Significance level (alpha)**: Boundary for specifying a statistically significant finding when interpreting the p-value.

We can see that the p-value is just a probability and that in actuality the result may be different. The test could be wrong. Given the p-value, we could make an error in our interpretation.

## Types of Errors

* **Type I Error**. Reject the null hypothesis when there is in fact no significant effect (false positive). The p-value is optimistically small.  *α* = probability of a Type I error.
* **Type II Error**. Not reject the null hypothesis when there is a significant effect (false negative). The p-value is pessimistically large. *β* = probability of a Type II error.

## Significance level and confidence level

**Alpha is the significance level used to compute the confidence level**. The confidence level equals 100\*(1 - alpha)%, or in other words, an alpha of 0.05 indicates a 95 percent confidence level. Standard\_dev is the population standard deviation for the data range and is assumed to be known.

The**confidence level** in hypothesis testing is the probability of not rejecting the null hypothesis when the null hypothesis is True:

P(Not Rejecting H0|H0 is True) = 1 - P(Rejecting H0|H0 is True)

The default confidence level is set at 95%.

## What is the statistical power of a test?

*β* = probability of a Type II error, known as a "false negative"

1 − *β* = probability of a "true positive", i.e., correctly rejecting the null hypothesis. **"1 − *β*" is also known as the power of the test.**

*α* = probability of a Type I error, known as a "false positive"

1 − *α* = probability of a "true negative", i.e., correctly not rejecting the null hypothesis

Изображение выглядит как диаграмма, линия, График, дизайн

Автоматически созданное описание

The power of a test is the probability of rejecting the null hypothesis when it’s false. It’s also equal to 1 minus the beta.

It is generally accepted we should aim for a power of 0.8 or greater.

## What are two ways to increase the power of a test?

To increase the power of the test, you can do two things:

You can increase alpha, but it also increases the chance of a type 1 error

Increase the sample size, n. This maintains the type 1 error but reduces type 2.

## Effect Size

Effect size tries to answer the question of “Are these differences large enough to be meaningful despite being statistically significant?”.

Effect size addresses the concept of “minimal important difference” which states that at a certain point a significant difference (ie p≤ 0.05) is so small that it wouldn’t serve any benefits in the real world. *Keep in mind, by small we do not mean a small p-value.*

A different way to look at effect size is the quantitative measure of how much the IV affected the DV. A high effect size would indicate a very important result as the manipulation on the IV produced a large effect on the DV.

Effect size is typically expressed as Cohen’s d. Cohen described a small effect = 0.2, medium effect size = 0.5 and large effect size = 0.8

Smaller p-values (0.05 and below) don’t suggest the evidence of large or important effects, nor do high p-values (0.05+) imply insignificant importance and/or small effects. Given a large enough sample size, even very small effect sizes can produce significant p-values (0.05 and below). In other words, statistical significance explores the probability our results were due to chance and effect size explains the importance of our results.

## Cohen's *d*

Cohen’s d is an [effect size](https://en.wikiversity.org/wiki/Effect_size) used to indicate the standardized difference between two means. It can be used, for example, to accompany reporting of [*t*-test](https://en.wikiversity.org/wiki/T-test) and [ANOVA](https://en.wikiversity.org/wiki/ANOVA) results. It is also widely used in [meta-analysis](https://en.wikiversity.org/wiki/Meta-analysis).

**Cohen’s D**, or *standardized mean difference*, is one of the most common ways to measure effect size. An effect size is how large an effect is. For example, medication A has a larger effect than medication B. While a [p-value](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/) can tell you if there *is*an effect, it won’t tell you how large that effect is.

Cohen’s D specifically measures the effect size of the difference between two means.  
Изображение выглядит как текст, Шрифт, линия, белый

Автоматически созданное описание

* spooled =[pooled standard deviations](https://www.statisticshowto.com/pooled-standard-deviation/) for the two groups. The formula is: √[(s12+ s22) / 2]

## Power Analysis

Statistical power is one piece in a puzzle that has four related parts; they are:

* **Effect Size**. The quantified magnitude of a result present in the population. [Effect size](https://machinelearningmastery.com/effect-size-measures-in-python/) is calculated using a specific statistical measure, such as Pearson’s correlation coefficient for the relationship between variables or Cohen’s d for the difference between groups.
* **Sample Size**. The number of observations in the sample.
* **Significance**. The significance level used in the statistical test, e.g. alpha. Often set to 5% or 0.05.
* **Statistical Power**. The probability of accepting the alternative hypothesis if it is true.

All four variables are related. For example, a larger sample size can make an effect easier to detect, and the statistical power can be increased in a test by increasing the significance level.

A power analysis involves estimating one of these four parameters given values for three other parameters. This is a powerful tool in both the design and in the analysis of experiments that we wish to interpret using statistical hypothesis tests.

For example, the statistical power can be estimated given an effect size, sample size and significance level. Alternately, the sample size can be estimated given different desired levels of significance.

Perhaps the most common use of a power analysis is in the estimation of the minimum sample size required for an experiment.

As a practitioner, we can start with sensible defaults for some parameters, such as a significance level of 0.05 and a power level of 0.80. We can then estimate a desirable minimum effect size, specific to the experiment being performed. A power analysis can then be used to estimate the minimum sample size required.

In general, large effect sizes require smaller sample sizes because they are “obvious” for the analysis to see/find. As we decrease in effect size we required larger sample sizes as smaller effect sizes are harder to find. This works in our favor as the larger the effect size the more important our results and fewer participants we need to recruit for our study.

## What are the confidence intervals of the coefficients?

**Confidence interval (CI)** is a type of interval estimate (of a population parameter) that is computed from the observed data. The confidence level is the frequency (i.e., the proportion) of possible confidence intervals that contain the true value of their corresponding parameter. In other words, if confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. However, the interval computed from a particular sample does not necessarily include the true value of the parameter. Since the observed data are random samples from the true population, the confidence interval obtained from the data is also random. If a corresponding hypothesis test is performed, the confidence level is the complement of the level of significance, i.e. a 95% confidence interval reflects a significance level of 0.05. If it is hypothesized that a true parameter value is 0 but the 95% confidence interval does not contain 0, then the estimate is significantly different from zero at the 5% significance level.

The desired level of confidence is set by the researcher (not determined by data). Most commonly, the 95% confidence level is used. However, other confidence levels can be used, for example, 90% and 99%.

Factors affecting the width of the confidence interval include the size of the sample, the confidence level, and the variability in the sample. A larger sample size normally will lead to a better estimate of the population parameter. A Confidence Interval is a range of values we are fairly sure our true value lies in.

X ± Z\*s/√(n), X is the mean, Z is the chosen Z-value from the table, s is the standard deviation, n is the number of samples. The value after the ± is called the margin of error.

## What is A/B testing?

**A/B testing** is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

To understand what A/B testing is about, let’s consider two alternative designs: A and B. Visitors of a website are randomly served with one of the two. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine whether one of the two designs has better efficacy.

With the data we collected from the activity of users of our website, we can compare the efficacy of the two designs A and B. Simply comparing mean values wouldn’t be very meaningful, as we would fail to assess the **statistical significance** of our observations. It is indeed fundamental to determine how likely it is that the observed discrepancy between the two samples originates from chance.

In order to do that, we will use a [two-sample hypothesis test](https://en.wikipedia.org/wiki/Two-sample_hypothesis_testing). Our **null hypothesis H0** is that the two designs A and B have the same efficacy, i.e. that they produce an equivalent click-through rate, or average revenue per user, etc. The statistical significance is then measured by the **p-value**, i.e. the probability of observing a discrepancy between our samples at least as strong as the one that we actually observed.



Now, some care has to be applied to properly choose the **alternative hypothesis Ha**. This choice corresponds to the choice between [one- and two- tailed tests](https://en.wikipedia.org/wiki/One-_and_two-tailed_tests).

A **two-tailed test** is preferable in our case, since we have no reason to know a priori whether the discrepancy between the results of A and B will be in favor of A or B. This means that we consider the alternative hypothesis **Ha** the hypothesis that A and B have different efficacy.

The **p-value** is therefore computed as the area under the the two tails of the probability density function **p(x)** of a chosen test statistic on all **x’** s.t. **p(x’) <= p(our observation)**. The computation of such p-value clearly depends on the data distribution. So we will first see how to compute it for discrete metrics, and then for continuous metrics.

## What are Statistical Tests?

Statistical tests are a way of mathematically determining whether two sets of data are significantly different from each other. To do this, statistical tests use several statistical measures, such as the mean, standard deviation, and coefficient of variation. Once the statistical measures are calculated, the statistical test will then compare them to a set of predetermined criteria. If the data meet the criteria, the statistical test will conclude that there is a significant difference between the two sets of data.

There are various statistical tests that can be used, depending on the type of data being analyzed. However, some of the most common statistical tests are t-tests, chi-squared tests, and ANOVA tests.

## Types of Statistical Tests

When working with statistical data, several tools can be used to analyze the information.

**1. Parametric Statistical Tests**

Parametric statistical tests have precise requirements compared with non-parametric tests. Also, they make a strong inference from the data. Furthermore, they can only be conducted with data that adhere to common assumptions of statistical tests. Some common types of parametric tests are regression tests, comparison tests, and correlation tests.

**1.1. Regression Tests**

Regression tests determine cause-and-effect relationships. They can be used to estimate the effect of one or more continuous variables on another variable.

* **Simple linear regression** is a type of test that describes the relationship between a dependent and an independent variable using a straight line. This test determines the relationship between two quantitative variables.
* **Multiple linear regression** measures the relationship between a quantitative dependent variable and two or more independent variables, again using a straight line.
* **Logistic regression** predicts and classifies the research problem. Logistic regression helps identify data anomalies, which could be predictive fraud.

**1.2. Comparison Tests**

Comparison tests determine the differences among the group means. They can be used to test the effect of a categorical variable on the mean value of other characteristics.

* **T-test**

One of the most common statistical tests is the t-test, which is used to compare the means of two groups (e.g. the average heights of men and women). You can use the t-test when you are not aware of the population parameters (mean and standard deviation).

* **Paired T-test**

It tests the difference between two variables from the same population (pre-and post-test scores). For example, measuring the performance score of the trainee before and after the completion of the training program.

* **Independent T-test**

The independent t-test is also called the two-sample t-test. It is a statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. For example, comparing cancer patients and pregnant women in a population.

* **One Sample T-test**

In this test, the mean of a single group is compared with the given mean. For example, determining the increase and decrease in sales in the given average sales.

* **ANOVA**

ANOVA (Analysis of Variance) analyzes the difference between the means of more than two groups. One-way ANOVAs determine how one factor impacts another, whereas two-way analyses compare samples with [different variables](https://www.enago.com/academy/explanatory-response-variable-statistics/). It determines the impact of one or more factors by comparing the means of different samples.

* **MANOVA**

MANOVA, which stands for Multivariate Analysis of Variance, provides regression analysis and analysis of variance for multiple dependent variables by one or more factor variables or covariates. Also, it examines the statistical difference between one continuous dependent variable and an independent grouping variable.

* **Z-test**

It is a statistical test that determines whether two population means are different, provided the variances are known and the sample size is large.

**1.3. Correlation Tests**

Correlation tests check if the variables are related without hypothesizing a cause-and-effect relationship. These tests can be used to check if the two variables you want to use in a multiple regression test are correlated.

* **Pearson Correlation Coefficient**

It is a common way of measuring the linear correlation. The coefficient is a number between -1 and 1 and determines the strength and direction of the relationship between two variables. The change in one variable changes the course of  another variable change in the same direction.

**2. Non-parametric Statistical Tests**

Non-parametric tests do not make as many assumptions about the data  compared to parametric tests. They are useful when one or more of the common statistical assumptions are violated. However, these inferences are not as accurate as with parametric tests.

* **Chi-square test**

The chi-square test compares two categorical variables. Furthermore, calculating the chi-square statistic value and comparing it with a critical value from the chi-square distribution allows you to assess whether the observed frequency is significantly different from the expected frequency.

## How to choose statistical test?

Изображение выглядит как текст, диаграмма, линия, Шрифт

Автоматически созданное описание

**1. Research Question**

The decision for a statistical test depends on the [research question](https://www.enago.com/academy/how-to-develop-good-research-question-types-examples/) that needs to be answered. Additionally, the research questions will help you formulate the data structure and [research design](https://www.enago.com/academy/experimental-research-design/).

**2. Formulation of Null Hypothesis**

After defining the research question, you could develop a null hypothesis. A [null hypothesis](https://www.enago.com/academy/what-is-null-hypothesis-what-is-its-importance-in-research/) suggests that no statistical significance exists in the expected observations.

**3. Level of Significance in Study Protocol**

Before performing the study protocol, a level of significance is specified. The level of significance determines the statistical importance, which defines the acceptance or rejection of the null hypothesis.

**4. The Decision Between One-tailed and Two-tailed**

You must decide if your study should be a one-tailed or two-tailed test. If you have clear evidence where the statistics are leading in one direction,  you must perform one-tailed tests. However, if there is no particular direction of the expected difference, you must perform a two-tailed test.

**5. The Number of Variables to Be Analyzed**

Statistical tests and procedures are divided according to the number of variables that are designed to analyze. Therefore, while choosing the test , you must consider how many variables you want to analyze.

**6. Type of Data**

It is important to define whether your data is continuous, categorical, or binary. In the case of continuous data, you must also check if the data are normally distributed or skewed, to further define which statistical test to consider.

**7. Paired and Unpaired Study Designs**

A paired design includes comparison studies where the two population means are compared when the two samples depend on each other. In an unpaired [or independent study design](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996580/#:~:text=A%20statistical%20test%20is%20used,design%20is%20paired%20(dependent)), the results of the two samples are grouped and then compared.

## What is Pearson Correlation?

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows the [linear relationship](https://www.calculushowto.com/types-of-functions/linear-function/#relationships) between two sets of data. In simple terms, it answers the question, *Can I draw a line graph to represent the data?* Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.

ρ = cov(X, Y)/ **σX\* σY,**

where cov is the covariance,

**σ** is the standard deviation.

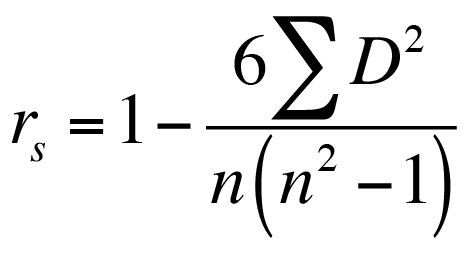
**Potential problems with Pearson correlation.**

The PPMC is not able to tell the difference between [dependent variables](https://www.statisticshowto.com/dependent-variable-definition/) and [independent variables](https://www.statisticshowto.com/independent-variable-definition/). For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, as a researcher you have to be aware of the data you are plugging in. In addition, the PPMC will not give you any information about the [slope of the line](https://calculushowto.com/what-is-a-slope/); it only tells you whether there is a relationship.

## What is Spearman Correlation?

The Spearman correlation between two variables is equal to the [Pearson correlation](https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient) between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) [rank](https://en.wikipedia.org/wiki/Ranking_(statistics)) (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.



## A comparison of the Pearson and Spearman correlation methods

Изображение выглядит как снимок экрана, линия, График

Автоматически созданное описание

Pearson = +0.851, Spearman = +1

A correlation coefficient measures the extent to which two variables tend to change together. The coefficient describes both the strength and the direction of the relationship.

The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

## Pearson’s chi-squared test

Fisher’s exact test has the important advantage of computing exact p-values. But if we have a large sample size, it may be computationally inefficient. In this case, we can use [Pearson’s chi-squared test](https://en.wikipedia.org/wiki/Pearson%27s_chi-square_test) to compute an approximate p-value.

Let us call **Oij** the observed value of the contingency table at row **i** and column **j**. Under the null hypothesis of independence of rows and columns, i.e. assuming that A and B have same efficacy, we can easily compute corresponding expected values **Eij**. Moreover, if the observations are normally distributed, then the χ2 statistic follows exactly a [chi-square distribution](https://en.wikipedia.org/wiki/Chi-square_distribution)with 1 degree of freedom.

Изображение выглядит как текст, Шрифт, линия, белый

Автоматически созданное описание

In fact, this test can also be used with non-normal observations if the sample size is large enough, thanks to the [central limit theorem](https://en.wikipedia.org/wiki/Central_limit_theorem).

In our example, using Pearson’s chi-square test we obtain **χ2 ≈ 3.825**, which gives **p-value ≈ 5.1%**.

## Z-test (Z-value)

The [Z-test](https://en.wikipedia.org/wiki/Z-test) can be applied under the following assumptions.

* The observations are normally distributed (or the sample size is large).
* The sampling distributions have known variance **σX** and **σY**.

Under the above assumptions, the Z-test exploits the fact that the following **Z statistic** has a standard normal distribution.

The z-value is another common test statistic where the null hypothesis suggests the means of two populations are equal. This metric goes beyond the t-value, which tests only a sample of the population. The z-score is also important for calculating the probability of a data value appearing within the normal distribution for a specific standard. This allows for the comparison of two z-values from different sample groups with varying standard deviations and mean values. To get the z-value, you can use the formula:

**z = (X - μ) / σ**, where *X* represents the raw data or score, *μ* is the mean of the population and *σ* is the standard deviation for the population. Изображение выглядит как текст, Шрифт, линия, белый

Автоматически созданное описание

## Student’s t-test (T-value)

In most cases, the variances of the sampling distributions are unknown, so that we need to estimate them. [Student’s t-test](https://en.wikipedia.org/wiki/Student%27s_t-test) can then be applied under the following assumptions.

* The observations are normally distributed (or the sample size is large).
* The sampling distributions have “similar” variances **σX ≈ σY**.

Under the above assumptions, Student’s t-test relies on the observation that the following **t statistic** has a Student’s t distribution.

The t-value is one type of test statistic that results from performing either t-tests or regression tests. Evaluating the t-value requires testing a null hypothesis where the means of both test samples are equal. If you perform a t-test or regression rest and find the means are not equal, you reject the null hypothesis for the alternative hypothesis. You can calculate a t-value using a common t-test with the formula:

**t = (X‾ - μ0) / (s / √n)**, where *X‾* is the sample mean, *μ*0 represents the population mean, *s* is the standard deviation of the sample and *n* stands for the size of the sample.

Изображение выглядит как текст, Шрифт, белый, линия

Автоматически созданное описание

## F-value

An f-value is a test statistic that you can get from an analysis of variance (ANOVA). This statistical test measures the difference in means for two or more independent samples. The f-value shows the significance of the mean differences, indicating whether the variance between the groups forms a relationship.

If the f-value is greater than or equal to the variation between the groups, the null hypothesis holds true. If the f-value is less than the variation between the sample groups, it rejects the null hypothesis. Calculating the f-value relies on sophisticated computations, which many data scientists perform with computer software.

## X2 value

The X2 value comes from non-parametric correlation tests that measure whether there is a causal relationship between variables. This value can also tell you whether the two variables you want to use in a statistical analysis already display a relationship. This test statistic becomes useful when preparing variables for testing in regression analysis, as the null hypothesis for the X2 value indicates independent samples.

## Welch’s t-test

In most cases Student’s t test can be effectively applied with good results. However, it may rarely happen that its second assumption (similar variance of the sampling distributions) is violated. In that case, we cannot compute a pooled variance and rather than Student’s t test we should use [Welch’s t-test](https://en.wikipedia.org/wiki/Welch%27s_t-test).

This test operates under the same assumptions of Student’s t-test but removes the requirement on the similar variances. Then, we can use a slightly different **t statistic**, which also has a Student’s t distribution, but with a different number of degrees of freedom **ν**.

Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание

Welch’s t-test

The complex formula for **ν** comes from [Welch–Satterthwaite equation](https://en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite_equation).

## Continuous non-normal metrics

In the previous section on continuous metrics, we assumed that our observations came from normal distributions. But non-normal distributions are extremely common when dealing with per-user monthly revenues etc. There are several ways in which normality is often violated:

* [zero-inflated distributions](https://en.wikipedia.org/wiki/Zero-inflated_model)— most user don’t buy anything at all, so lots of zero observations;
* [multimodal distributions](https://en.wikipedia.org/wiki/Multimodal_distribution) — a market segment tends purchases cheap products, while another segment purchases more expensive products.

However, if we have enough samples, tests derived under normality assumptions like Z-test, Student’s t-test, and Welch’s t-test can still be applied for observations that signficantly deviate from normality. Indeed, thanks to the [central limit theorem](https://en.wikipedia.org/wiki/Central_limit_theorem), the distribution of the test statistics tends to normality as the sample size increases. In the zero-inflated and multimodal example we are considering, even a sample size of 40 produces a distribution that is well approximated by a normal distribution.

But if the sample size is still too small to assume normality, we have no other choice than using a non-parametric approach such as the Mann-Whitney U test.

## Mann–Whitney U test

This test makes no assumption on the nature of the sampling distributions, so it is fully nonparametric. The idea of [Mann-Whitney U test](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test) is to compute the following **U statistic**.



Mann-Whitney U test (image by author)

The values of this test statistic are tabulated, as the distribution can be computed under the null hypothesis that, for random samples **X** and **Y** from the two populations, the probability **P(X < Y)** is the same as **P(X > Y)**.