# Course Notes: Introduction to Statistics for HCI Using Jamovi

A course for HCI practitioners and researchers on inferential statistics and Jamovi, an open-source statistics software, comparable to IBM SPSS

**Jurek Breuninger, IU International University of Applied Sciences, Germany**

When developing and improving products in a human-centered way or testing a hypothesis in your HCI research, you will very likely want to collect some quantitative data like user satisfaction or user performance and make sense of it. Is your design faster than the competitor? Does this new technology really lead to fewer errors? But alas, testing hypotheses... don't you need to understand statistics for that? And isn't statistics just boring mathematics, but harder? Do you want to learn how to do basic statistical tests on HCI-related data with a free open-source tool and have little to no affection for math? Then this course is for you!

https://j.breunis.de/CHI2023/

# Statistics in HCI

We use statistics in HCI to make sense of quantitative data. There are two kinds of statistics that are helpful for this purpose: Descriptive statistics and inferential statistics.

We use descriptive statistics to describe a dataset. If we want to report our findings of an experiment, descriptive statistics are an easy way of giving a summative conclusion of the dataset. Descriptive statistics mainly compute to metrics: 1. Measures of central tendency, i.e. what is the average location of the date, e.g. mean, median, mode. 2. Measures of dispersion, i.e. how varied are the datapoints in the set, how far are they spread, e.g. variance, standard deviation.

But many problems in HCI cannot be easily observed and analyzed by an experiment that considers all possible datapoints. If you develop a business software for internal use, you can conduct a usability test with all people in the company that will use the software. But if you want to sell a product, the number of potential users is usually far larger than what you can observe in an experiment. For those cases you need inferential statistics.

Inferential statistics help you to draw general conclusions if you can only observe a part of the datapoints that are involved in your question. In HCI we usually invent or change a solution of a technical system that will have an effect on very many people that use it. But since we cannot test these changes with all potential users, we will have to infer from an experiment with some users to the effect on all users. So inferential statistics offers tools to assess the observations in an experiment and draw conclusions for a broader user audience.

Three methods of inferential statistics are common in HCI: Hypothesis testing, confidence intervals, and Bayesian statistics.

Hypothesis testing is the most common in HCI and other research disciplines and will be the focus of this course. One can use several statistical tests to determine based on the dataset of an experiment, if a claim (e.g., “Reading is faster on large displays than on small displays”) is likely true in general. Given a measure of certainty, the test will either accept or reject the claim.

Although hypothesis testing established the term “statistically significant” for well observed effects, it gives little information about the importance of the claim. A design improvement might be clearly measurable and accepted by a hypothesis test, but so small that it has no practical relevance to most users in real-world scenarios. Confidence intervals are a way of predicting the average effect as a range, e.g. the improved system is very likely between 20 and 40 seconds faster than the old one. This helps interpreting the relevance of the findings. Given a 60 second task, this would be an important improvement, given a 60 minute task, not as much.

Bayesian statistics is a branch of statistics that interprets stochastic probability not as the relative frequency of repeated random experiments, but derives it based on prior knowledge and conditional probabilities. This can have advantages for the sensitivity of the tests and the interpretation of their results. But gaining and applying prior knowledge to those tests can be challenging. Jamovi contains also Bayesian statistical tests, but they are out of the scope of this course.

## Types of Studies

In order to identify a link between design decisions in product design and the impact for the users, according to Sauro and Lewis (2016, p. 12) three types of studies are available:

- The experiment: Randomly selected test persons interact with the product. As many factors as possible influencing the interaction other than the product itself are avoided by the study design. This is also referred to as a controlled experiment, since an attempt is made to control all factors that could influence the result. This allows such a study to have high internal validity. Excluding interfering factors is usually quite costly. External validity can also be very high, depending on how representative the participants are. However, a laboratory environment can itself become an influencing factor, as subjects may behave differently there than under realistic conditions. This reduces the validity.
- The quasi-experiment: Here, too, an attempt is made to exclude interfering influencing factors. However, the assignment of the test persons is not random, but is usually determined by external circumstances, e.g., active users of product A are compared with active users of product B. Observed differences can thus be due to the characteristics of the users, which led to the choice of product, and not due to the product itself. The internal validity is therefore quite weak. The extent of the external validity depends on the representativeness of the test persons; in the case of natural groupings, as in this example, it is quite strong.
- The correlation study: The correlation between two metrics, e.g. the results of a usability questionnaire with the shopping cart value, is analyzed. However, since correlation does not imply causality, internal validity is very low. External validity again depends on representativeness of the participants considered. Correlation studies are often much easier to implement than experiments if the relevant metrics are easy to collect or already available. However, they are only useful if there is a very reasonable assumption that causality exists in addition to correlation, as this kind of study cannot prove this. Correlation can also result from two apparently dependent variables being influenced by a third, unobserved factor. For example, a correlation study may show a correlation between the training time with a software and the quality of the work results with this software. However, if the subjects were not randomly selected, causality need not be present. It could also be the case that exactly those subjects who have great fun with the work under consideration have both done more training and produced better work results. The individuals who produced worse work results also did less training, but the reason for both was their low motivation. Thus, motivation was confounded with training time and work outcome, creating an appearance of causality.

Due to the described advantages and disadvantages of the different study types, in practice the UX researcher will mostly aim to conduct a controlled experiment if well-supported results are needed. However, to save time and resources, quasi-experiment and correlational study are also frequently used. However, such studies tend to have lower explanatory power and the higher uncertainty in the interpretation of the data must be taken into account.

# Hypotheses Testing, Variables, Scale of Measure

In a summative usability study, a controlled experiment is usually conducted in which one or more factors are varied and one or more metrics are observed that are expected to be affacted. The manipulated factors are also called the independent variables since they can be freely selected in the experimental design. The metrics on which an influence is to be detected are the dependent variables.

Before choosing the variables, planning the procedure and data collection, a hypothesis is formulated based on the given question to be proven by the experiment. An example of such a hypothesis is:

$H_1$: With the new touchscreen interface, user satisfaction changes compared to the previous mouse control.

A hypothesis cannot be proven based on a sample because not all data points are known. Any of the unknown data points could disprove the hypothesis. However, it is quite easy to disprove a hypothesis; all that is needed is at least one data point that is inconsistent with the hypothesis. In terms of statistical tests, this means that a hypothesis is proven by disproving its opposite. The opposite of a hypothesis is the so-called null hypothesis:

$H_0$: There is no difference between the average user satisfaction with the touchscreen and the mouse control.

The above hypothesis is undirected, it only describes whether there is a difference between the two variants or not. In many UX studies, however, no difference is relevant, but an improvement by the newer variant. A directed hypothesis is formulated for this purpose:

$H_1$: With the new input assistant, users creating a data record are faster than with the old input mask.

The appropriate null hypothesis for this is:

$H_0$: With the new input assistant, users creating a data record are slower than or as fast as with the old input mask.

The directed hypothesis test has the advantage that it has greater discriminatory power than the undirected test. It therefore requires fewer subjects to detect an effect or can detect a smaller effect with the same number of subjects. The problem is that such an experimental design partly anticipates the result before any data have been collected and analyzed. Until the 1950s, therefore, only undirected tests were usually used, since a researcher cannot know for sure the direction of the results before data collection, and only the undirected hypothesis was perceived as unbiased (Sauro & Lewis, 2016, p. 257). Since then, controversy arose as to whether a directed hypothesis was not warranted in some experiments. Especially in the UX field, it is very common to compare variants, one of which was designed using proven techniques to outperform the other. Degradation seems very unlikely under these conditions. Sauro and Lewis (2016, p. 257) recommend using directed hypothesis tests only when comparing to a known benchmark, such as the predecessor product, and preferring undirected tests otherwise. Especially in usability studies in the product development process, if the new variant is worse than the predecessor product, it is not a bad thing that this cannot be proven. If the proof of improvement is lacking, there is still a need for action.

To prove a hypothesis, the associated null hypothesis must be refuted. In the example of touchscreen and mouse control, the average operating speed of the test subjects is compared. The goal is to show that the observed difference does not occur by chance in this sample, but that if the experiment were repeated with any number of other samples, a difference would also be observed in most cases.

Various statistical tests can be used to determine this probability. Which test is appropriate for hypothesis testing depends first on the nature of the values. Depending on the study design, both independent and dependent variables can take on different classes of values. These different levels of measurement are called the scale of measure. The simplest form of scale is the nominal scale (from Latin nomen = name). Its values correspond to different categories that do not have a ranking order with respect to each other. In the example above, the values for input are nominal scale, it can take the value "touch screen" or "mouse control". Other examples of nominal data are gender or occupation. Values that can be ranked are called ordinal data (Latin ordinalis = ordered). However, they cannot be put into fixed ratios, so it is not useful to calculate with them, even if they are sometimes coded in numerical form. Ordinal scales are, for example, school grades, educational attainment, placements in rankings. Data, which stand in fixed distance to each other, are called cardinally scaled. The basic arithmetic operations can also be applied to them and, for example, a mean value can be calculated. They can be further differentiated into discrete and continuous, depending on whether they can only assume certain values (usually integer, e.g. number of errors), or also all intermediate values (e.g. time for task completion).

## Errors, Significance Level

When a UX study is conducted with a sample, it is important to check whether differences in the dependent variables were caused by the independent variables. If differences, or effects, are observed, it is important to ensure that they do not occur by chance only in this sample. Statistical hypothesis tests are used for this purpose. They can determine whether an effect would be observed again with a given probability if repeated any number of times with other samples. This probability is called the significance level. It can be chosen arbitrarily for a hypothesis test. In UX research, however, as in most branches of research, a significance level of $\alpha = 5\%$ has become the quasi-standard. This means that the hypothesis is confirmed by the test if the calculated probability that the observed effect also occurs in the population is greater than 95 percent.

Conversely, this means that in 5 percent of the cases in which the hypothesis test confirms the hypothesis, the observed effect only occurs by chance in this sample and does not occur in the population. Accordingly, every twentieth hypothesis test produces a false result. It proves an effect that does not exist in reality. This is a type I error or false positive. Correspondingly, there is also a type II error or false negative: when the test rejects the hypothesis although an effect exists in reality. This occurs when the test has too little power, i.e. it is not sensitive enough to detect the effect. The power of a test depends on the significance level, effect size and sample size (Cohen, 1992). In theory, it would be desirable to also keep the probability of type II errors ($\beta$) as small as possible, say also at 5 percent. Then the test would have a power of 0.95 (power = $1 - \beta$).

However, such a high power in combination with a high significance level either leads to only very strong effects being detected or the necessity for very large sample sizes. To avoid this, lower power is accepted in practice, usually 0.8 (Sauro & Lewis, 2016, S. 262). The common practice of mainly wanting to avoid type I errors and rather accepting type II errors comes from scientific research, since there the erroneous publication of an effect that does not exist is seen more critically than the inconclusive conduct of a study (which is then usually not published).

In applied UX research, this practice should be questioned, and test strength and significance level should be more closely aligned with real-world conditions. If a new variant is compared with an existing product and this is wrongly assessed as better (type I error), it is likely to be introduced although it does not bring any advantages or even disadvantages. If the new variant is wrongly assessed as not better than the existing product (type II error), it will hopefully be further improved, so that in a later test the existing advantage is even more likely to show up. However, if development is discontinued, the inferior product is used for the time being and the development costs are wasted. With regular and frequent evaluation of a product, where small improvements have far-reaching positive consequences, an occasional incorrect test result is quickly corrected by subsequent iterations.

However, if testing is continuously done with low power, much potential for improvement is wasted because it is not recognized. It can be argued that high test power should have a similar value to significance in UX studies, but preferring significance is also useful in some cases. Sauro und Lewis (2016, S. 259–261) refer to reducing the significance level to 90 percent and even 80 percent in some cases as tenable in a product development context. This allows for an increase in test power (such as to 0.9 or higher) or a reduction in sample size.

More important than achieving a certain level of significance in a test is a broad understanding of the capabilities and limitations of statistical tests. Since the test of a sample can never give absolute certainty, a superior strategy in product development is to compensate for uncertainty through continuous development and frequent evaluation. Frequent evaluations (with a new sample or

modified experimental design) also reduce the tendency of some researchers to bias the results of statistical tests in a desired direction by making subsequent changes. If the result of the intended test does not meet expectations, some try to influence this by adjusting the hypothesis or test parameters. However, such changes are test repetitions that bring with them the known probabilities of error. These error probabilities are cumulative. Thus, a subsequent increase in the significance level from 5 to 10 percent increases the error probability not to 10 percent but to 14.5 percent ($1 - (0.95 * 0.9) = 0.145$).

## Effect Size

To design an experiment, the effect size plays an important role (it affects e.g. sample size). Since the effect size is unknown in advance, it must be estimated. The effect size is the amount of the difference that the change in the independent variable causes in the dependent variable. To be able to specify it independently of the size and unit of the measured variable, there are several dimensionless measures to describe the effect size. They are suitable for different scale levels and the associated statistical tests. A common measure of effect size is Cohen's d, it describes the difference in means between two groups. Cohen (1988, S. 25) calls effects with $d > 0.2$ small, $d > 0.5$ medium, and $d > 0.8$ large. This is based on his studies and experiences in behavioral research and he emphasizes that this classification is not arbitrarily transferable to other disciplines and should be adapted to the research method even within behavioral research.

In usability and UX research and evaluation, according to this definition, mainly large effects are investigated and observed, medium ones rarely, and small ones hardly ever. This is because the proof of small effects is in most cases disproportionate to the effort. Whereas a small effect in the efficacy of a drug can save many lives, justifying studies with over a thousand participants, a small effect in product improvement is unlikely to be noticed, and is especially likely to be overshadowed by the many other random influencing factors in the interaction with a product. If design adjustments make the operating speed of software a few seconds faster for a typical task that takes several minutes, this is unlikely to be very helpful, since even a pause in thinking or the operator's physical condition will have a greater impact on operating speed. But for the detection of such a small effect an elaborate experiment with a large sample is needed. Therefore, the detection of small effects in the UX area is usually not economical and the required resources are better used elsewhere.

In A/B testing, on the other hand, uncovering small effects can be valuable. Large samples are very easy to implement in A/B testing on the web. In the case of products or services with a very large number of users, for example, a change that prevents the few users who would otherwise abort from continuing at one point can relevantly increase customer satisfaction and profitability. If the effect size is estimated during experiment design in product development and no prior experience is available, $d = 0.8$ is a reasonable starting point. A preliminary experiment with a small sample can also be helpful for effect size estimation. In the medium term, however, UX researchers should be able to use the effect size of completed studies as basis for similar experiments.

The (estimated) effect size is not only important for the design of the study. The observed effect size is also very important for the interpretation of the results. In many UX studies (as well as in other disciplines), the primary goal is to reach the significance level (the level of certainty that the effect is not reported by error), and a detailed interpretation of a significant test result is often left out. It is quite easy to reach the significance level for many hypotheses in real use cases. Either in a trivial way by setting the significance level quite low or by choosing a very large sample or by observing many variables. Some researchers also lower the significance level or increase the sample size subsequently after they have performed the hypothesis test once and are dissatisfied with the lack of significance. This is strictly a test replication, which increases the error probabilities accordingly. This can be compensated for by adjusting the test design, but there is no consensus on the extent to which this should be done and whether it is acceptable to do so (Di Gennaro, 2019; Wason et al., 2014). To interpret a result in a purposeful way, the effect size should be considered in addition to reaching the significance level. For many decisions in the product development process, it may make more sense to pursue design decisions whose large effect is subject to some uncertainty, rather than integrating very well-documented small effects whose relevance to the overall user experience and product success is beside the point.

## Population and Sample

The participants of an empirical study are called test persons or subjects. In principle, empirical UX studies should recruit test persons from the target group for the product. However, it is rarely possible to study the entire target group. In statistics, the entire target group is called the population. Instead, in practice, the study can only be conducted with a sample, i.e. a subset of the population. There are two questions to answer when selecting a sample: How can we ensure that the sample is representative of the population? And how large must the sample be to be able to draw conclusions about the population from its results?

When selecting a sample, the question arises as to whether it behaves on average like the population or whether precisely those subjects were selected who differ on average from the population, so that their test results are not representative of the population. This must be avoided, as it can distort the results, i.e. reduce validity. The theoretically best method to prevent this is random sampling. Even then, it is still possible that the sample is not representative of the population, but the probability of this happening is minimized. Almost all statistical tests rely on the assumption of a random sample when assessing the probability of an observed effect occurring by chance. In practice, however, it is impossible to take a truly random sample in studies with people as subjects. For this, one would have to have a list of all people in the target group, which does not exist for almost all products. Then people randomly selected from this list would have to participate in the study and produce valid data.

Real samples are always more homogeneous than would be theoretically desirable, i.e. they have a unifying property. For example, in many UX studies, all subjects come from the city in which the study is conducted. If the study is conducted at a university, the subjects are mostly young students. If a study is conducted by online questionnaire, the subjects are mostly frequent Internet users and tech-savvy. Such non-representativeness cannot be avoided in practice. Nevertheless, this does not necessarily lead to a bias in the results. This is because the non-representative nature of the sample does not always have a measurable impact on the target metric under investigation. However, the greater the homogeneity of the sample, the higher the probability of bias.

Such common characteristics should only be allowed in the sample if they do not affect the target metric. In order to estimate the homogeneity of the sample, demographic data is collected, e.g. age, gender, place of residence, level of education and others. For UX studies, data that might impact UX metrics are most relevant, e.g., experience with certain products, affinity for technology, eyesight, handedness, and others. Before collecting demographic data, consider whether it might be relevant to the study. Although it is tempting to collect as much data as possible to be sure, in the interest of data minimization, only those that can be well justified should be chosen. Some demographic data is necessary to ensure membership in the target group. For example, when evaluating a navigation app for motorcycling, it should be found out whether the test subjects ride a motorcycle. Other data are necessary to avoid bias. For example, it could make sense to record the size of the hand in the case of the motorcycle app in order to avoid a study that was only carried out with test subjects with small hands causing usability problems for users with large hands to remain undetected.

The required sample size depends heavily on the experimental design. However, in practice, the choice of sample size is often based more on habit and ease of recruitment than on careful planning of the experimental design. In particular, the number of 30 subjects has been established here as a rule of thumb (Sauro & Lewis, 2016, S. 254). However, for valid results, fewer subjects may be sufficient or significantly more may be necessary. This can be calculated with an a priori test power analysis. Software such as G*Power can be used for this purpose (Faul et al., 2007). It calculates the necessary sample size given the statistical test used, the significance level, the test power and the

expected effect size. The following figure shows as an example that under certain assumptions a sample of twelve test persons can be sufficient for a comparison of two variants. A one-sided dependent-sample t-test with a significance level of 5 percent and a required test strength of 0.80 achieves a test strength of 0.83 with this number of test persons.

## t-Test

For the very frequent question in UX/usability studies as to whether cardinally scaled values differ between two groups, the t-test is usually used as a hypothesis test. There are several types of t-tests. The three most important in the UX field are:

- One-sample t-test: This test is used to check whether the mean value of a sample differs from a given value. This allows a comparison with a benchmark, e.g. a predecessor or competitor product whose target value is already known, for example from a past test.
- Paired samples t-test: Tests whether the means of two samples with the same subjects differ (repeated measures). In other words, it is a sample in which the independent variable can take on two nominal scaled values. This method is used in many study designs to compare two design variants using quantitative metrics.
- independent samples t-test: Tests whether the mean values of two samples with different subjects differ (between-subjects design). It is used where repeated measurements are not possible or undesirable due to interference.

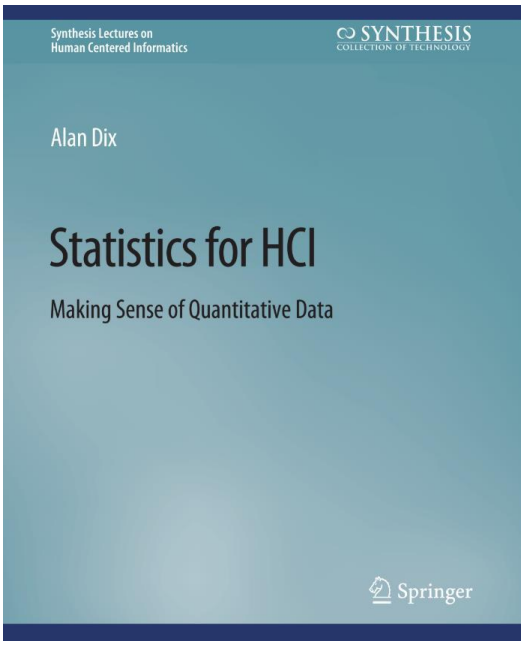
A t-test is used to calculate a value p from the data points of the sample, which describes the probability that the rejection of the null hypothesis occurs by chance (type I error). If p is smaller than the previously determined significance level, the null hypothesis $H_0$ is rejected and the alternative hypothesis $H_1$ is considered proven.

## Analysis of Variance

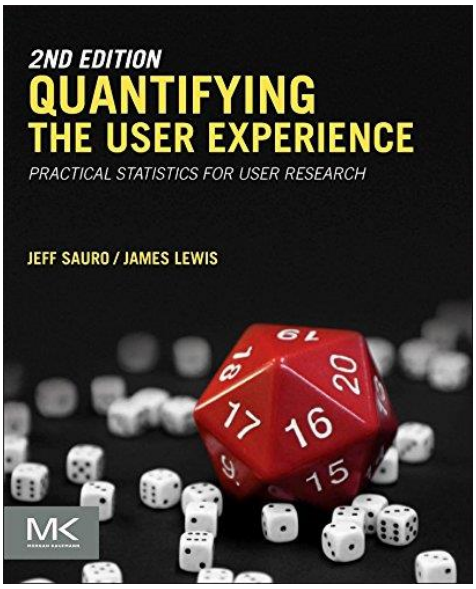
The t-test has the disadvantage that it can only be applied to a maximum of two samples. However, if the independent variable takes on more than two values, for example because more than two product concepts/design variants are being compared, an analysis of variance (ANOVA) is used. It works similarly to a t-test but avoids the problem of error accumulation that would occur when repeatedly applying the t-test for pairwise individual comparisons. As with the t-test, the result of an ANOVA is a p-value that represents the probability of a type I error. If this is smaller than the previously determined significance level, the hypothesis that there is a difference between the variants is accepted. However, the ANOVA itself only tests whether a difference exists between any pair of the variants. There is now no information as to which variants differ from each other and to what extent. For this purpose, so-called post-hoc tests are used, which compare the variants pairwise and calculate the error probability for this. There are several types of post-hoc tests that differ in their assumptions about how much the post-hoc test increases the probability of error. A common, conservative post-hoc test is the Bonferroni post-hoc test.

There are also several variants of ANOVA that are designed for different experimental designs. There is the repeated measures ANOVA variant, as well as variants for tests with several independent variables (Two-way ANOVA, Factorial ANOVA) and for tests with several dependent variables (Multivariate ANOVA = MANOVA).

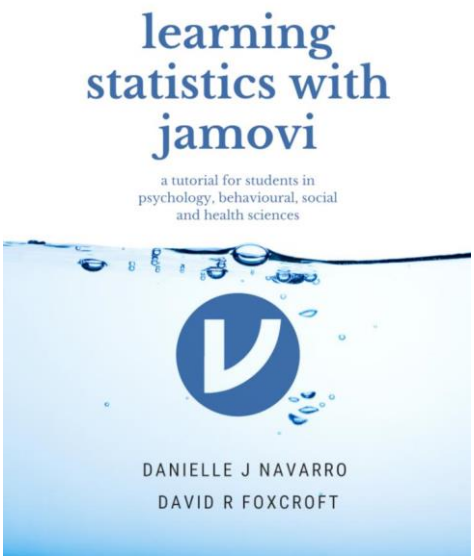
# Literature



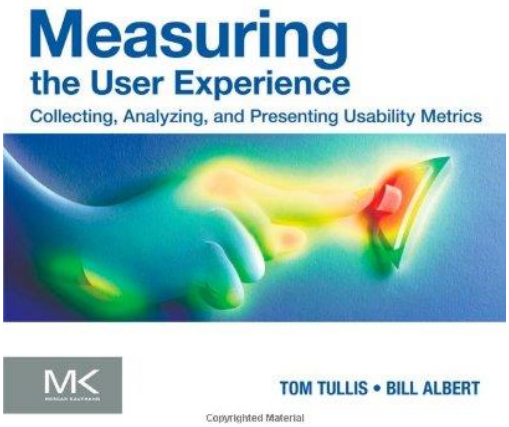

Dix, A. (2020). *Statistics for HCI: Making Sense of Quantitative Data* (1. Aufl.). *Synthesis lectures on human-centered informatics*. Springer International Publishing; Imprint: Springer.



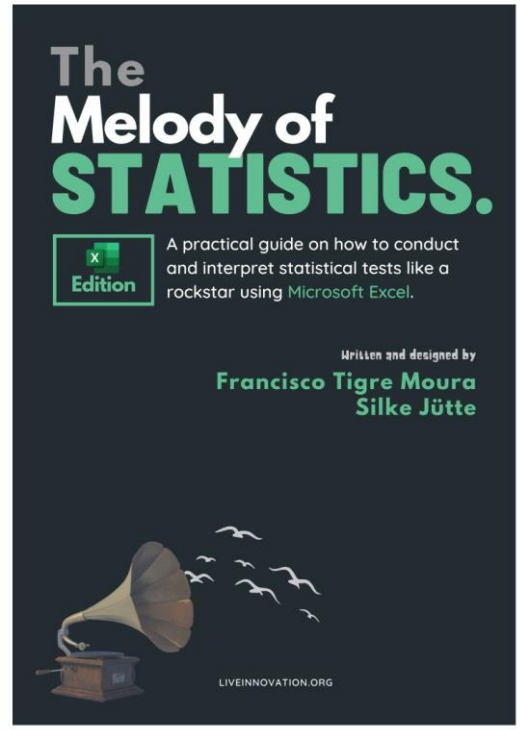

Sauro, J. & Lewis, J. R. (2016). *Quantifying the User Experience: Practical Statistics for User Research* (2nd ed.). Elsevier Science. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4592083




Navarro, D. J. & Foxcroft, D. R. (2022). Learning Statistics with Jamovi: a tutorial for students in psychology, behavioural, social and health sciences. Vorab-Onlinepublikation. https://doi.org/10.24384/hgc3-7p15




Tullis, T. & Albert, B. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. *Interactive Technologies*. Elsevier/Morgan Kaufmann.




Tigre Moura, F. & Jütte, S. (2022). *The Melody of Statistics: A practical guide on how to conduct and interpret statistical tests like a rockstar using Microsoft Excel*. Excel Edition. https://liveinnovation.org

## References


Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Erlbaum.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Di Gennaro, G. (2019). *Multiple testing: when should we adjust for multiplicity?* StatsImprove. https://www.statsimprove.com/en/multiple-testing-when-should-we-adjust-for-multiplicity/

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,* 39(2), S. 175–191.

Nielsen, J. (2000). *Why You Only Need to Test with 5 Users.* Nielsen Norman Group. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Sauro, J. & Lewis, J. R. (2016). *Quantifying the User Experience: Practical Statistics for User Research* (2nd ed.). Elsevier Science. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4592083

Wason, J. M. S., Stecher, L. & Mander, A. P. (2014). Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*, *15*, 364. https://doi.org/10.1186/1745-6215-15-364