EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2017)
K. Lawonn, N. Smit, and D. Cunningham (Editors)

From a user study to a valid claim: how to test your hypothesis and avoid common pitfalls

N.H.L.C. de Hoon and E. Eisemann and A. Vilanova

Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands

Abstract

The evaluation of visualization methods or designs often relies on user studies. Apart from the difficulties involved in the design of the study itself, the existing mechanisms to obtain sound conclusions are often unclear. In this work, we review and summarize some of the common statistical techniques that can be used to validate a claim in the scenarios that are commonly present in user studies in visualization, i.e., hypothesis testing. Usually, the number of participants is small and the mean and variance of the distribution are not known. Therefore, we will focus on the techniques that are adequate within these limitations. Our aim for this paper is to clarify the goals and limitations of hypothesis testing from a user study perspective, that can be interesting for the visualization community. We provide an overview of the most common mistakes made when testing a hypothesis that can lead to erroneous claims. We also present strategies to avoid those.

Categories and Subject Descriptors (according to ACM CCS): G.3 [Mathematics of Computing]: Probability and Statistics—Experimental Design

1. Introduction

In visualization, we typically work with user-based measures of quality. We often have to ask users or measure users performance to evaluate how efficient or effective the visualization method is to achieve its expected goal. Objective measurements of efficiency and effectiveness are often not possible. The users have their own opinion and certainty on the questions asked, and as such we will have multiple measurements of the underlying true answer. Moreover, often the acquisition of the experimental data is not ideal, e.g., the users have various backgrounds, measurements contain noise and are a discretization of the underlying continuous variable, or they are simply incomplete. This is because a user study is merely a sample of the full population, e.g., we cannot involve every potential user, and often users in a specific domain are scarce. As such, the true value for the population (ground truth) is uncertain. Therefore, when we develop user studies we typically want to test a hypothesis, i.e., we want to show that a claim we make has validity.

In this paper, we revise common literature on hypothesis testing, and we provide a summary of the basics for such scenarios. Often, we can assume the distribution is normal, however, the mean and variance of this distribution are unknown and the number of samples is often small. Therefore, the t-distribution is discussed, as it can be used in this case to provide an estimate of the mean of the underlying normal distribution of the population. Once we have an estimate of the underlying distribution, one can derive a confidence interval, for example, for the mean, which provides us with some confidence that the true value lies in a certain interval. With these concepts in place, hypothesis testing and the pitfalls of hypothesis testing are then explored. In this paper, we present knowledge that is available in books and other sources, however, our goal is to summarize it, such that is more easily accessible.

2. Related work

The amount of evaluations based on experiments with users is increasing in the visualization field [TM04, IIC*13]. Munzer et al. [Mun09] describe the various validation options for visualizations based on the characteristics of the visualization and underlying data. Furthermore, the work by Smit et al. [SL16] provides guidelines on the different contribution types and visualization scenarios and when to apply the different evaluation types. Naturally, not every validation is as valuable or suitable for every task. However, often user studies are useful to determine the usability for the target users of the visualization. How to conduct a good user study is demonstrated with an example by Glaβer et al. [GSB*16]. In their work, they provide a practical example and describe the user evaluation process in detail. Moreover, they provide a decision tree, that can help to determine which statistical tools should be used per situation. More general guidelines on conducting a good user study are given by Carpendale [Car08].

In this work instead of focusing on conducting a user study, we mainly focus on the statistics that help to validate claims made based on user studies. While some of the statistical methods are covered, only the techniques most common in our field are dis-

cussed in this paper. For a more in-depth, yet accessible explanation of statistical evaluation, among others the book by Montgomery and Runger [MR06] provides a good reference.

3. t-Distribution

While crowd sourcing could be used to gather many participants, often domain experts are required for our user studies. Usually the tasks are not easily generalizable such that layman participants can be used. Therefore, the number of participants is usually small and both the mean and the variance of the distribution are not known, and as such an assumption must be made on the underlying distribution. In many cases, a reasonable assumption would be that the distribution is normal. This can be tested using a *normality test*, e.g. the Shapiro-Wilk test. The t-distribution can be used under such assumptions to estimate the mean of the underlying normal distribution of the full population. Since the t-distribution is developed to describe the samples drawn from a full population with a normal distribution the number of samples is taken into account. A t-distribution represents the probability of estimating a value as mean given a number of samples. The more samples are included, the more the t-distribution represents a normal distribution. If we have N participants that provide us with N independent samples $X_1, X_2, ..., X_N$ with mean $\bar{X}$ of the underlying normal distribution $N(\mu, \sigma^2)$ the t-distribution of the random variable T with $N-1$ degrees of freedom is defined as:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{N}}, \tag{1}$$

where μ is unknown and the sample variance, S, is given by:

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{2}$$

Figure 1 shows a comparison between the normal distribution and the t-distribution for different sample sizes. By increasing the number of samples, the t-distribution approximates a normal distribution. On the other hand, for lower number of samples the probabilities are lower and the spread is wider.

Note, that when no assumptions on the normality of the distribution can be made a so-called *non-parametric* or *distribution-free* test should be applied. An example with real world data is provided by Glaβer et al. [GSB*16].

4. Confidence intervals

Given the distribution of the data it is often useful to estimate an interval in which an interesting population parameter lies, for example, the mean, μ, of an underlying normal distribution. Although we cannot be sure the true (unknown) parameter actually lies within this interval, we can have a certain *confidence*. For example, we would like to compute the probability that with 95% confidence, the expected μ is situated within an interval of the t-distribution. That means we would like to compute the interval such that the probability that the true μ falls in this interval equals 0.95. For the t-distribution T we would compute this as follows:

$$P\left(-t_{\alpha/2,N-1} \leq T \leq +t_{\alpha/2,N-1}\right) = 1-\alpha, \tag{3}$$

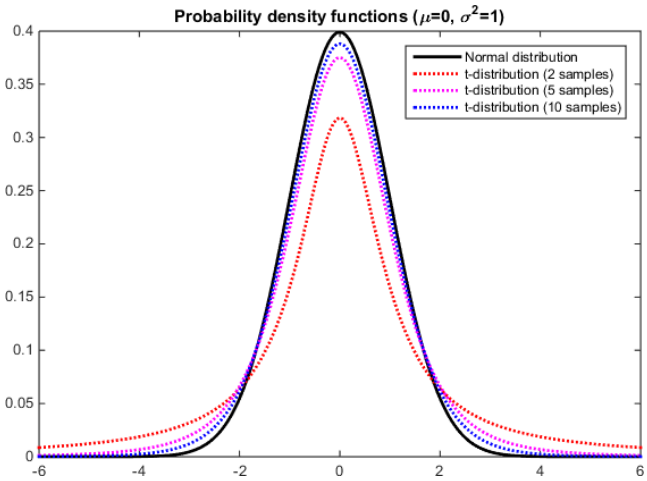

Figure 1: *A comparison between the normal distribution (black) and the t-distribution for different sample sizes.*

where $t_{\alpha/2,N-1}$ is value for which the integral over t-distribution with $N-1$ degrees of freedom covers $100(\alpha/2)\%$ of the total probability, so for a 95% confidence interval, $\alpha = 0.05$. Note that, the t-distribution is symmetric. Now if we fill in the definition of T from Equation 1 and rewrite Equation 3 a bit we get the following *two-sided* confidence interval for μ:

$$P\left(\bar{X} - t_{\alpha/2,N-1} S/\sqrt{N} \leq \mu \leq \bar{X} + t_{\alpha/2,N-1} S/\sqrt{N}\right) = 1-\alpha \tag{4}$$

To compute a *one-sided* confidence interval, one of the boundaries can be dropped and $t_{\alpha,N-1}$ should be used instead to obtain the same confidence. An example of the three possible 95% confidence intervals is given by Figure 2, here $N = 5$, $N = 3$ and $\alpha = 0.05$.

Note that, when we have to use the t-distribution the confidence interval will be bigger when less samples are available. This due to the tails being wider and longer compared to the normal distribution, as is illustrated by Figure 1.

5. Hypothesis testing

Once we have an idea of the distribution of our data, most often we want to show that our data supports a claim, i.e., test a hypothesis on our data. To do so, a *null hypothesis* H_0 is to be defined, typically an equality, as well as an *alternative hypothesis* H_1 which represents our claim. Examples of null hypotheses could be "the model yields results equal to the measurements", or "the users think system A performs as good as system B". The corresponding alternative hypotheses could be "the model yields results worse than the measurements", or "the users think system A performs better than system B". Note that, in these examples, we have a one-sided alternative hypothesis, that is, here we want to show system A performs better, not just different from system B. The goal of hypothesis testing is to determine the likeliness of the null hypothesis to be true; the less likely the null hypothesis is true, the more likely it is our alternative hypothesis is a good representation. The alternative hypothesis is shown to be valid by showing the unlikeliness of the

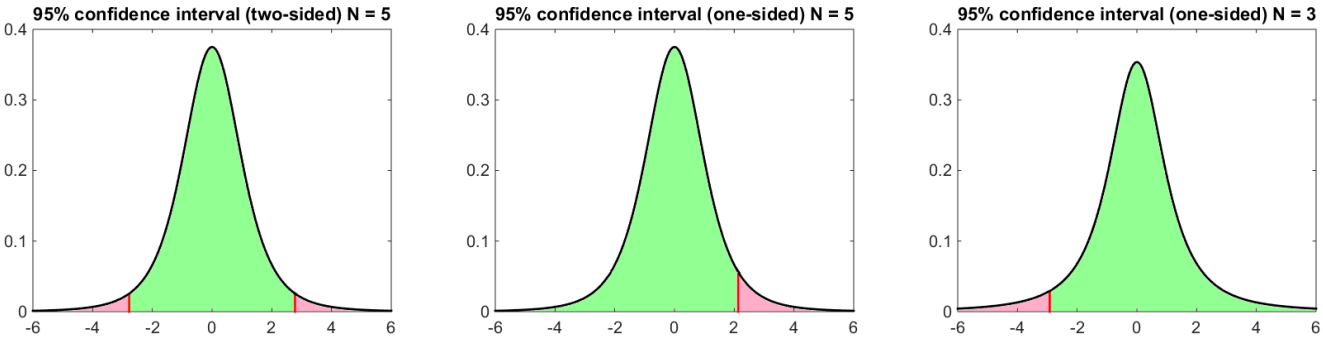

Figure 2: *The 95% confidence interval highlighted in green for the t-distribution with 4 degrees of freedom ($N = 5$, $N = 3$ and $\alpha = 0.05$). The left most image shows the two-sided confidence interval, while the middle and right image show the one-sided confidence intervals. The areas highlighted in red fall outside the interval.*

consequence of assuming the null hypothesis to be true. More precisely, we expect the alternative hypothesis to be true if the null hypothesis is false. This implies that the alternative hypothesis is much more plausible than the null hypothesis given the data. Hence for the examples above, rejecting the null hypothesis would mean respectively that it is likely that that system B performs better than system A (according to the users).

Again, we assume the variance and mean to be unknown but the population distribution to be approximately normal. We can write for the null hypothesis $H_0 : \mu = \mu_0$, that is, the real μ is represented properly by our estimate μ_0. The alternative hypotheses H_1 can be

$$\begin{aligned} &\mu \neq \mu_0, \\ &\mu < \mu_0 \text{ or} \\ &\mu > \mu_0. \end{aligned} \tag{5}$$

Rejecting the null hypothesis would mean for the examples in Figure 2 that the observed value falls in the highlighted areas, the so-called *critical regions*. If this is the case it means that the probability that the true mean is estimated by μ_0 is actually equal or below 0.05, i.e. the probability that the null hypothesis is correct is only 5%. This is where the *p-value* is used. The *p*-value is the probability of finding a value in an independently obtained data set, that is at least as extreme as what was measured, under the assumption that null hypothesis is true. That is the *p*-value gives us the probability that the true mean is estimated by μ_0. Note that, in order to obtain the strongest conclusion we want to reject the null hypothesis. Furthermore, we typically want to fix a low value for the probability of rejecting the null hypothesis when in fact this hypothesis is true. Hence, we want to reduce the probability of a *false positive*. A false positive occurs when the null hypothesis is true, but will be rejected based on the data, i.e., your p-value is below 0.05, due to the specific samples that were tested.

In the following we will focus on the alternative hypothesis $H_1 : \mu \neq \mu_0$. Since we consider the *t*-distribution, our test statistic is given by:

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{N}}. \tag{6}$$

If the null hypothesis is assumed to be true, we cannot reject that T_0 is indeed a *t*-distribution with $N-1$ degrees of freedom with $\mu = \mu_0$. We would reject the null hypothesis if our observed value t_0 of T_0 falls in the critical region, that is for our two-sided interval $t_0 < -t_{\alpha/2,N-1}$ or $t_0 > t_{\alpha/2,N-1}$.

When we cannot reject the null hypothesis, this does not necessarily mean that the null hypothesis is true. It just means that there is a high probability that the results are obtained chance, so no conclusion can be drawn. Hence instead of accepting the null hypothesis, we say to *fail to reject* the hypothesis. This does not mean the null hypothesis is true, this simply means there is a lack of evidence to reject it. This conclusion is weak, *failing to reject the null hypothesis* is not the same as saying that it was proven, since it was only not disproved. This is also demonstrated by the justice system, a suspect is assumed to be innocent until proven otherwise, i.e. the null hypothesis is that the suspect is innocent. The verdict "not guilty" does not necessarily mean the suspect is innocent, the evidence is simply not sufficient to reject null hypothesis and thus the suspects innocence. Therefore, the interpretation of non-significant results should be done with care [AB95].

Another quantity that is useful to report is the *power*, the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. Hence, the power of a statistical test is defined as:

$$\text{Power} = 1 - P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) \tag{7}$$

Naturally, one wants the power to be as close to 1 as possible. Note that, since the power depends the probability of failing to reject the null hypothesis when it is false, the power is based on the variance in the data, the sample size, and the magnitude of the effect, *how well does the alternative hypothesis describe the system*. A good practice is to compute the power before the experiment is developed, but this requires knowledge that it is often not available.

6. Hypothesis testing pitfalls

While hypothesis testing may seem easy, one should be aware of the mistakes that are commonly made. Often, hypothesis testing is applied incorrectly and as such the resulting conclusion is wrong, providing a false sense of validity. In this section, some of the more

common pitfalls of hypotheses testing are outlined, as well as how to avoid them.

While it may seem trivial, one must define their hypotheses before conducting the user study. This makes it possible to link every question to a single hypothesis, making the analysis easier and thus, the resulting conclusion clearer. Furthermore, by doing this, one can state the expected result, that is set a threshold beforehand for when the hypothesis should be rejected. Doing so also makes you more aware of the possible impact of the questions asked on the outcome of the experiment. Formulating the hypothesis after looking at the results and tuning your values after the experiment is bad practice. In this case one would have to redo the experiments with the hypothesis to be able to claim its validity.

Know your data; measuring performance provides a different type of data compared to user studies, and as such the resulting conclusion will be different. For example, a measurement of the performance can provide an idea of which component is working well, and which does not. On the other hand, a user study provides a more high level overview, and thus conclusions on the underlying details can be hard to derive.

Make sure the hypothesis is clear and testable. That means that the hypothesis should be based on something that can be measured, such as the task completion. Hypotheses like "our tool encourages user to explore the data in more depth" will be difficult to measure and therefore, difficult to test. A better hypothesis would be "our tool increases the productivity", as this can be evaluated using specific questions.

The hypothesis should be backed up by reasoning, that is, why you would expect to find a certain result. By having knowledge on why a result is expected, the probability of the result being a false positive is reduced. Numerous examples of trends that are statistically significant exist yet are, quite obviously, not actually correlated [Vig17].

On a similar note, testing too many hypotheses increases the possibility of finding a false positive by chance. This can be avoided by having a reasoning as to why the hypothesis makes sense to test and by defining the hypothesis beforehand.

Another crucial point is using the right participants for the user study, the opinion or experience of a layman does not state any usefulness regarding a highly specialized tool for domain experts. This seems obvious but is often difficult to fulfill. On the other hand, the full user population should be sampled. The more users that can be included the more the actual population is being sampled.

However, one should not keep including users until the results are significant. This is an obvious way to end up in false positives, i.e. the more tests one does, the higher the probability one finds a false positive. Ideally a power study determines beforehand the amount of users required sense to get sensible results, however, as mentioned before, a power study is often not possible without assumptions.

7. Conclusion

Visualizations are often evaluated with the aid of user studies. Such evaluations can be utilized to support a claim and provide a sense of validity. In this work, the common statistical techniques that can be used to validate such a claim, thus to test a certain hypothesis, are discussed. For this purpose, we focus on the techniques that work under the assumption that the number of participants is small and the mean and variance of the underlying Gaussian distribution are not known. This work is meant to clarify the goals and limitations of hypothesis testing from the perspective of a user study. Furthermore, an overview of the most common mistakes made when hypothesis testing is provided. For each of these mistakes recommendations are presented on how to avoid them. Our goal is to make this knowledge more easily accessible to newcomers.

References

[AB95] ALTMAN D. G., BLAND J. M.: Statistics notes: Absence of evidence is not evidence of absence. *BMJ 311*, 7003 (1995), 485. URL: http://www.bmj.com/content/311/7003/485, arXiv:http://www.bmj.com/content/311/7003/485.full.pdf, doi:10.1136/bmj.311.7003.485. 3

[Car08] CARPENDALE S.: *Evaluating Information Visualizations*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 19–45. URL: http://dx.doi.org/10.1007/978-3-540-70956-5_2, doi:10.1007/978-3-540-70956-5_2. 1

[GSB*16] GLAβER S., SAALFELD P., BERG P., MERTEN N., PREIM B.: How to evaluate medical visualizations on the example of 3d aneurysm surfaces. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (2016), Bruckner S., Preim B., Vilanova A., Hauser H., Hennemuth A., Lundervold A., (Eds.), The Eurographics Association. doi:10.2312/vcbm.20161283. 1, 2

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MÖLLER T.: A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2818–2827. URL: http://dx.doi.org/10.1109/TVCG.2013.126, doi:10.1109/TVCG.2013.126. 1

[MR06] MONTGOMERY D. C., RUNGER G. C.: *Applied Statistics and Probability for Engineers*, 4 ed. Wiley, 2006. 2

[Mun09] MUNZNER T.: A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (Nov 2009), 921–928. doi:10.1109/TVCG.2009.111. 1

[SL16] SMIT N. N., LAWONN K.: An introduction to evaluation in medical visualization. In *Proceedings of EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization* (2016), Eurographics Digital Library. URL: http://graphics.tudelft.nl/Publications-new/2016/SL16. 1

[TM04] TORY M., MÖLLER T.: Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics 10*, 1 (Jan. 2004), 72–84. URL: http://dx.doi.org/10.1109/TVCG.2004.1260759, doi:10.1109/TVCG.2004.1260759. 1

[Vig17] VIGEN T.: Spurious correlations. http://www.tylervigen.com/spurious-correlations, 2017. [Online; accessed 27-Februari-20178]. 4