

The User Is in the Numbers

Jeff Sauro | Oracle Corporation | jeff@measuringusability.com

Measuring usability is an important topic that very much needs to be discussed. Being part of the conversation usually requires reading peer-reviewed journal articles in HCI. This, in turn, requires knowledge of the techniques and the jargon of statistics and experimental design (the foundation of numerical precision in empirical disciplines). To help move the discussions along, I've attempted to provide some clarification and explanations on statistical concepts you'll encounter: p-values, power, and confidence intervals. The examples come right from papers published in HCI proceedings. I've kept the discussion high-level and left out formulae, as it is more important to understand the concepts than to plug values into a formula.

What's That *P-Value* All About?

It's hard to get far in a journal article without encountering a p-value (probability value). For example:

“The effect of time of completion depending on the instructional medium is statistically significant, $F(3, 71) = 3.75, p = 0.015$ ” [7].

The p-value is like the punch line to a long joke: If you don't hear the whole story, the ending won't make a lot of sense. So here's a short version of the long story.

The p-value is one result derived from a technique called Null Hypothesis Significance Testing (NHST). NHST is like a judge who must decide the fate of an individual based on some evidence. There are essentially two choices: guilty or innocent. From this decision, you can be correct in your judgment or make one of two errors. You convict an innocent person or let a criminal get away. With NHST, just like in court, you're innocent until proven guilty. That is, we assume nothing has happened, there is no effect, there is no difference until proven otherwise (that's the "null" in NHST—the difference is null).

A researcher is faced with the same decision and set of potential errors as a judge. These are conveniently referred to as Type I and Type II errors. A Type I error is akin to convicting an innocent person. In effect, it's saying there is a difference when one doesn't really exist. This would be like concluding product B has a faster task-completion time than product A when in fact there isn't a difference. A Type II error is letting the criminal go free or saying there is no difference when one really does exist. Using the same example, it's like concluding both products have the same task time, when product B's is actually faster. Unlike the accused in court, who knows if he committed the crime, there is no way to force the truth out of an interface. We must rely on sampling and testing to help inform our decisions.

Back to the punch line. The p-value quantifies the probability of making a Type I error. So when you see "these differences are statistically significant ($p < .05$)," you can read that as: "The probability that the difference you see (which is real in the data) could have resulted from a scenario in which there was truly no difference is less than five percent."

Using the example above, the data in that research paper showed that there was a difference in task times among three instructional media. If the researchers were to test every user on each instructional medium, the probability that the difference between versions is actually smaller (or non-existent) than their data, is less than five percent.

Naturally, you'd want the probability of making the wrong decision to be zero. In statistics (as in life), nothing is 100 percent certain, so we have to be willing to accept some level of risk. By convention, that level (called alpha) is usually five percent. That is, if you conduct the same experiment 100 times, in five instances you *would not* see a difference as big, or bigger, than what you observed simply due to chance variation. When being 95 percent sure isn't good enough (such as situations with a risk of fatality), this level can be raised to 99 percent or even 99.9 percent (alpha lowered to 0.01 and 0.001).

Power and Type II Errors

So what about the Type II error? Here's the catch. The p-value doesn't provide the probability that you're making a Type II error. Take this example:

The mean movement time for the joystick was 1.544 seconds (sd = .305). For the touchpad, the mean movement time was 1.563 seconds (sd =.285). These differences are not statistically significant ($F_{1,22} = .024, p > .05$) [4].

You would probably conclude there is no difference between the joystick and touchpad since the probability of a Type I error is too high ($p > .05$). But in saying the difference hasn't been proved, we now are susceptible to making a Type II error and should immediately ask, if there really is a difference, what is the likelihood the researchers would be able to detect it?

To know the probability of a Type II error requires a power calculation. Power is 1—the probability of a Type II error (called beta). While you may be familiar with the Type I error convention ($\alpha = .05$), you're probably not as familiar with the Type II error convention ($\beta = .20$), making the Power convention .80. Why is beta .20 and not .05 like alpha? Consider this. Which is a worse offense, letting a criminal go free or sending an innocent person to jail? Most would argue we shouldn't falsely convict an innocent person—better to let the criminal go (perhaps they will be caught in another nefarious act). By setting beta to .20, you're willing to accept four times the risk of making a Type II error. In many research situations this is acceptable (especially since the alternative of using five percent usually requires very large sample sizes.)

While the calculation of power is beyond the scope of this article, you should know that it involves the sample size, alpha, the variability of the sample (in standard deviations), and the difference between groups (see [2, 5] for calculations). Using these variables from the example above (12 users per group, SD of .3, alpha .05 and difference of .02), the probability of a Type II error is greater than 90 percent, so the power is less than ten percent—not very powerful (remember, the goal is 80 percent). The best way to increase

power is increasing the sample size, as the other variables are largely outside the control of the practitioner.

Even though there was an observed difference (1.562 seconds vs. 1.544 seconds), we can't conclude it's greater than a chance difference since the Type I error rate is too high ($p > .05$). But if we conclude there is no difference, there's a greater than 90 percent chance we're making a Type II error. The Catch 22 with NHST is that without enough power, it's difficult to make any decision.

There is not enough statistical power to draw any real conclusions here [6].

Power is one of the more-neglected areas in usability research and behavioral/psychological research in general; much has been written about the paucity of power in publications [2, 3]. Considering the cost and effort needed to conduct a usability analysis, power should be a high priority. If you know ahead of time that the probability you'll be able to detect a difference (even a large difference) is very small, then you're wasting a lot of intellectual and financial capital. A summary of the relationships in NHST is in Table 1.

Table 1: NHST Relationships

	The Truth	
Your Decision	No difference (person is innocent)	There is a difference (person is guilty)
There is a difference (convict)	Type I Error (alpha)	Correct Decision (Power, which is 1-beta)
No difference (acquit)	Correct Decision (1-alpha)	Type II Error (beta)

So What's Wrong with NHST?

Even with sufficient power, a major criticism with NHST lies in the fact that it's easy to interpret statistically significant events as practically significant. That is, just because you get a statistically significant p-value doesn't mean the difference you observe has major consequences. For example, take the following situation from actual usability data. Users completed the same two tasks on two versions of a product. Version 2 had a statistically faster task-completion time on both tasks ($p < .05$, *log transformation*). Great. Go with Version 2, right? But when the product is released to the public, how much faster should we expect users to complete each of the tasks? It's not evident from the p-value.

A Picture Tells a Thousand Words: *Confidence Intervals*

Taking the same data and graphing it using confidence intervals reveals the wide variability and imprecision found in the comparison of versions (see Figures 1 & 2).

While the difference in sample means is statistically significant in both tasks, for Task 1 (Figure 1) the difference may only be a few seconds (where the intervals overlap) to more than a minute (where they aren't overlapping). These are very different results on a two-minute task. In other words, you can expect anywhere from a barely detectable difference

to a very noticeable one. Contrast this with the results of Task 2 (Figure 2), where there is no overlap in the confidence intervals. Version 2 is much less variable (narrower interval), allowing for a more precise prediction. While context will dictate whether either of these tasks has practical significance to users, even the top of the confidence interval on Version 2 for Task 2 suggests a noticeable reduction in task time.

The ability of confidence intervals to communicate precision and location often makes them better than p-values alone. Next time you see a significant p-value, construct a confidence interval from the data and see what evidence there is for practical significance.

Figure 1: Task 1, 95 percent confidence intervals for Versions 1 and 2 ($p < .05$)

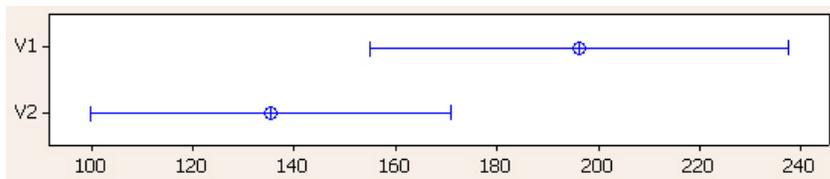
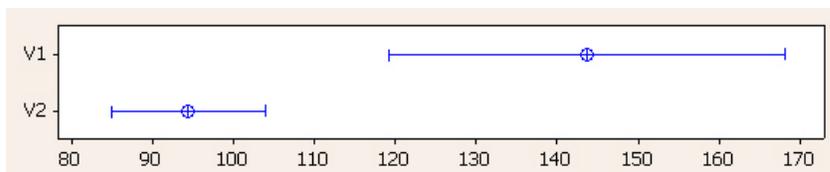


Figure 2: Task 2, 95 percent confidence intervals for Versions 1 and 2 ($p < .05$)



Conclusion

So what do you do after you generate the confidence intervals? If you say there's no difference, you're likely committing a Type II error. If you say there is a difference, you're most likely not committing a Type I error, but the resulting difference is just not that significant. Hypothesis testing, like statistics in general, is a careful balancing act in drawing conclusions from data. Larger differences from data allow for stronger claims. Perfunctory number crunching without regard to consequences is a bad idea.

This complication should not lead you to conclude, as some unfortunately do, that statistics should be categorically dismissed with such phrases as "you can show anything you want with statistics." If someone intends to deceive an audience, they are likely to get away with it (with or without statistics) only for so long. If the intent is to dispassionately uncover the truth (by eliminating possibilities), statistics provides a method for systematically verifying ideas. It also allows others to systematically verify the verification process [1]. The alternative to quantitative precision is intuition, experience, or guessing. Anecdotes and good intentions don't come with p-values, power calculations, and confidence intervals. Relying on the former to make important decisions leaves little room to verify that the right decision was made—much less detect deception.

Whenever you use Hypothesis Testing (NHST), you either convict or acquit. When you convict you are subject to a making a Type I error (alpha). When you acquit (technically fail to convict), you are subject to a Type II error (beta). Even with that in mind, statistical significance is unrelated to practical significance. A one-second difference

could either be statistically significant or not, and practically significant or not. The first will depend on your data; the second depends on the context.

References

1. Abelson, Robert P (1995) "Statistics as Principled Argument," Laurence Erlbaum Associates
2. Cohen, J. (1992) *A Power Primer*. Psychological Bulletin 112, 155-159
3. Kirakowski, J, (2005)"Summative Usability Testing: Measurement and Sample Size" in R.G. Bias and D.J. Mayhew (Eds): "Cost Justifying Usability: An Update for the Internet Age." Morgan Kauffman Publishers, CA, 2005.
4. Douglas, Sarah A., Kirkpatrick, Arthur E.; MacKenzie, Scott (1999) Testing pointing device performance and user assessment with the ISO 9241, Part 9 standard environments in *Proceedings of the SIGCHI conference on Human factors in computing systems*.
5. Lewis, J. R. (2006). Usability testing. In G. Salvendy (ed.), *Handbook of Human Factors and Ergonomics* (pp. 1275-1316). New York, NY: John Wiley.
6. Polys, Nicholas F., Kim, Seonho ; Bowman, Doug A. (2005) Effects of information layout, screen size, and field of view on user performance in information-rich virtual environments in *Proceedings of the ACM symposium on Virtual reality software and technology VRST '05*

7. Tang, Arthur ; Owen, Charles ; Biocca, Frank; Mou Weimin (2003) “Comparative effectiveness of augmented reality in object assembly” in *Proceedings of the SIGCHI conference on Human factors in computing systems*