HCI Statistics without *p*-values

Pierre Dragicevic

HCI Statistics without *p*-values

Pierre Dragicevic*

Project-Team AVIZ

Research Report n° 8738 — June 2015 — 32 pages

Abstract: Statistics are tools to help end users (here, researchers) accomplish their task (advance scientific knowledge). Science is a collective and cumulative enterprise, so to be qualified as usable, statistical tools should support and promote clear thinking as well as clear and truthful communication. Yet areas such as human-computer interaction (HCI) have adopted tools – i.e., *p*-values and statistical significance testing – that have proven to be quite poor at supporting these tasks. The use and misuse of *p*-values and significance testing has been severely criticized in a range of disciplines for several decades, suggesting that tools should be blamed, not end users. This article explains why it would be beneficial for HCI to switch from statistical significance testing to *estimation*, i.e., reporting informative charts with effect sizes and confidence intervals, and offering nuanced interpretations of our results. Advice is offered on how to communicate our empirical results in a clear, accurate, and transparent way without using any *p*-value.

Key-words: Human-computer interaction, information visualization, statistics, user studies, p-value, statistical testing, confidence intervals, estimation

* Pierre Dragicevic is with Inria, France; email: pierre.dragicevic@inria.fr

Statistiques pour l'IHM sans valeur-*p*

Résumé : Les statistiques sont des outils dont l'objectif est d'aider des utilisateurs (ici, des chercheurs) à accomplir leur tâche (faire avancer le savoir scientifique). La science étant un processus collectif et cumulatif, pour être qualifiés d'utilisables, les outils statistiques doivent faciliter et promouvoir une pensée cohérente ainsi qu'une communication claire et honnête. Pourtant des domaines tels que l'IHM (l'interaction homme-machine) ont adopté des outils (les valeurs-*p* et les tests de signification statistique) qui se sont montrés clairement insuffisants pour ces tâches. L'utilisation abusive des valeurs *p* et des tests statistiques a été violemment critiquée dans de nombreuses disciplines depuis plusieurs décennies, suggérant que la faute en revient aux outils, et non aux utilisateurs. Cet article explique pourquoi il serait bénéfique pour l'IHM de passer des tests statistiques à l'*estimation*, c'est à dire à la production de graphiques informatifs avec des tailles d'effet et des intervalles de confiance, et à des interprétations nuancées de nos résultats. Cet article explique comment communiquer nos résultats empiriques de façon claire, précise, et transparente sans utiliser aucune valeur-*p*.

Mots-clés : Interaction homme-machine, visualisation d'information, statistiques, études utilisateur, valeur-p, tests statistiques, intervalles de confiance, estimation

This report is a draft version of a book chapter currently in preparation.

1   Introduction

A common analogy for statistics is the toolbox. As it turns out, researchers in human-computer interaction (HCI) study computer tools. A fairly uncontroversial position among them is that tools should be targeted at end users, and that we should judge them based on how well they support users' tasks. This applies to any tool. Also uncontroversial is the idea that the ultimate task of a scientific researcher is to contribute useful scientific knowledge by building on already accumulated knowledge. Science is a collective enterprise that heavily relies on the effective communication of empirical findings. Effective means clear, accurate, and open to peer scrutiny. Yet the vast majority of HCI researchers (including myself in the past, as well as researchers from many other disciplines) fully endorse the use of statistical procedures whose usability has proven to be poor, and that are able to guarantee neither clarity, nor accuracy, nor verifiability in scientific communication.

One striking aspect of orthodox statistical procedures is their mechanical nature: data is fed to a magic black box, called "statistics", and a binary answer is produced: the effect exists or (perhaps) not. The idea is that for the work to qualify as scientific, human judgment should be put aside and the procedure should be as objective as possible. Few HCI researchers see the contradiction between this idea and the values they have been promoting – in particular, the notion that humans+algorithms are more powerful than algorithms alone (Beaudouin-Lafon, 2008). Similarly, researchers in information visualization (infovis) went to great lengths to explain why data analysis cannot be fully delegated to algorithms (Fekete et al, 2008): computing should be used to augment human cognition, not act as substitutes for human judgment. Every year infovis researchers contribute new interactive data analysis tools for augmenting human cognition. Yet when analyzing data from their own user studies, they strangely resort to mechanical decision procedures.

Do HCI and infovis researchers suffer from multiple personality disorder? A commonly offered explanation for this contradiction is that there are two worlds in data analysis: *i)* exploratory analysis, meant to generate hypotheses, and where human judgment is crucial and *ii)* confirmatory analysis, meant to test hypotheses, and where human judgment is detrimental. This article focuses on confirmatory analysis, and challenges the view that human judgment can be left out in the process.

By "orthodox statistical ritual" and "mechanical decision procedures" I refer to a family of statistical procedures termed *null hypothesis significance testing (NHST)*. I will compare NHST with *estimation*, an alternative that consists in reporting effect sizes with confidence intervals and offering nuanced interpretations. I will skip many technical details, widely available elsewhere. The key difference between the two approaches lies in their usability, and it can be summarized as follows:
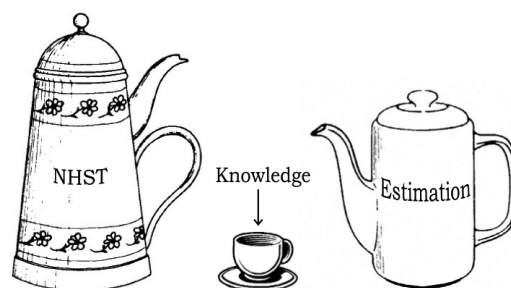

Figure 1: A picture summarizing the practical difference between null hypothesis significance testing (NHST) and estimation. Drawing inspired from Norman (2002).

NHST proceeds in several steps among which *i)* computing quantities called *p*-values and *ii)* applying a cut-off to these *p*-values to determine "statistical significance". Discussions about NHST often conflate the two, but it is better to address them separately. Section 2 focuses on the notion of *p*-value divorced from the notion of statistical testing. Confidence intervals will be used both to explain *p*-values informally, and as a baseline of comparison. Section 3 discusses the notion of statistical significance and compares it with estimation. Section 4 offers practical advice on how to achieve clear and truthful statistical communication through estimation.

2 Choosing a Pill: p-values, Effect Sizes and Confidence Intervals

The following problem is adapted from Ziliak and McCloskey (2008, p.2303). Suppose your best friend wants to buy a pill to loose weight, but she cannot make up her mind. As a proponent of evidence-based decision making, you search for publications online, find four different studies testing different pills, write down the results and compile them into a single chart:
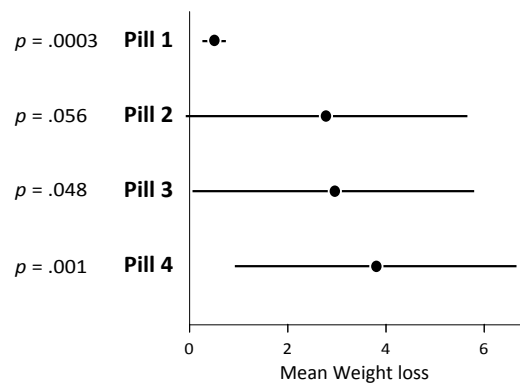

Figure 2: Chart showing the results from four (imaginary) studies on the effectiveness of different weight-loss pills. Error bars are 95% confidence intervals and *p*-values assume a null hypothesis of no effect.

Let us pause and make sure we understand what the figure shows[1]. In any clinical trial, there is uncertainty as to the true efficacy of a drug, partly because the drug has only been tested on a few people. Ideally, these people constitute a random *sample* from a *population* of interest (e.g., all overweight US citizens). The mean weight loss only informs us about the *sample*, but a better measure would be the weight loss averaged across the entire *population*. Neither measure will tell us for sure what will happen to your friend, but the population average (to which we assume your friend belongs) would be much more reliable. Thus we decide it is really our quantity of interest, even though it can only be guessed.

Each black dot in the chart shows our *best bet* for this population mean. It corresponds to the average weight loss for the study participants, i.e., the sample means[2]. The bars are 95% confidence intervals. Roughly speaking, they show other plausible values for the population-wise weight loss (Cumming, 2012). The numbers to the left are the *p*-values for a null hypothesis of no effect (zero weight loss). Briefly, *p*-values capture the strength of evidence against this null hypothesis[3]. The closer the value is to 0, the more evidence that the pill has at least some effect on the population overall (since *p* is between 0 and 1, the zero from the integer part is typically omitted).

---

[1]What follows is the most widespread interpretation of statistical inference, but alternatives exist (Frick, 1998). In this article we stick to the classical interpretation for convenience.

[2]Other slightly different methods for estimating the population mean will be briefly mentioned in Section 4.3.

[3]This is consistent with a Fisherian interpretation of *p*-values (Goodman, 1999, p.997; Gigerenzer, 2004, p.593). It is not the same as the fallacy of the transposed conditional, which will be discussed in Section 2.2.2.

Anyone familiar with stats will have immediately noticed the enormous amount of uncertainty in the data (except maybe for pill 1 – perhaps the sample was much larger?) and should not feel compelled to make any definitive statement. But here you need to make a decision. Given the data you have, which pill would you recommend?

I have shown this problem to different audiences and most people would choose pill 4. And it is indeed a sensible answer: it is reasonable to favor a pill that yields the maximum expected utility – here, weight loss. Each black dot, i.e., sample mean, shows your friend's most likely weight loss with that pill – a black dot is actually about *seven times more plausible* than the confidence interval's endpoints (Cumming, 2013, p.17). For your friend, pill 4 is the best bet, and it is certainly a much better bet than pill 1.

Now assume pill 4 does not exist. Which pill would you pick among the remaining ones? Look at Figure 2 on the previous page carefully. With some hesitation, most people reply pill 3. Then, with a bit more hesitation, pill 2. Some people would pick pill 1 on the way, but the most reasonable choice is really pill 4, then 3, then 2, then 1. The expected weight loss with pill 1 is way lower than with any other. Unless your friend had bet her life that she will lose at least some weight (even one gram), there is no logical reason to favor pill 1 over any other.

2.1   How Useful is the Information Conveyed by p?

Not very much. When presented with the pill problem, many researchers will ignore *p*-values, despite using them in their papers[4]. This stems from a correct intuition: the *p*-values are not only largely irrelevant to the decision, but also redundant. If needed, a *p*-value can be roughly inferred from a confidence interval by looking at how far it is from zero (Cumming and Finch, 2005; Cumming, 2012, pp.106–108).

But suppose we get rid of all confidence intervals and only report the *p*-values:

Pill 1	$p = .0003$
Pill 2	$p = .056$
Pill 3	$p = .048$
Pill 4	$p = .001$

Table 1: The *p*-value for each pill

Ranking the pills based on this information yields a quite different outcome: pill 1 appears to give the most impressive results, with a hugely "significant" effect of $p = .0003$. Then comes pill 4 ($p = .001$), then pills 3 and 2, both close to .05. Such a ranking assumes that losing *some* weight (even a single gram) is the *only* thing that matters, which is absurd, both in research and in real-world decision making (Gelman, 2013b). We should obviously acount for the black dots in Figure 2 (i.e., our best bets) at the very least.

2.1.1   The Importance of Conveying Effect Sizes

The black dots are what methodologists generally mean by "effect sizes". Many researchers are intimidated by this term, thinking it involves some fancy statistical concept. But the most general meaning is simply *anything that might be of interest* (Cumming, 2012, p.34), and these are often means, or differences between means. More elaborate measures of effect size do exist (Coe, 2002) – among which *standardized effect sizes* – but they are not always necessary, nor are they necessarily recommended (Baguley, 2009; Wilkinson, 1999, p.599).

---

[4]Several HCI researchers find it hard to relate to this example because they are used to compute *p*-values to compare techniques, not to assess each technique individually. If that can help, think of the reported weight losses as differences between pills and a common baseline (e.g., placebo). Regardless, our decision maker cannot run more analyses to compare results. The point of the scenario is to capture how results from *already published studies* are communicated and used to inform decisions.

p-values capture what is traditionally termed *statistical significance*, while effect sizes capture *practical significance* (Kirk, 2001). Results for pill 1, for example, can be said to exhibit a high statistical significance (with zero as the null hypothesis) but only a moderate practical significance compared to others. The term *statistical significance* is regularly and misleadingly contracted to *significance*, leading some to recommend dropping the word *significance* (Tryon, 2001, pp.373–374; Kline, 2004, p.86).

Practical significance is our focus, both in research and in real-world decision making. But as mentioned previously, we are really interested in *population effect sizes*. A black dot only conveys our best guess, and it is thus crucial to also convey information on uncertainty. Here are two ways of doing this:
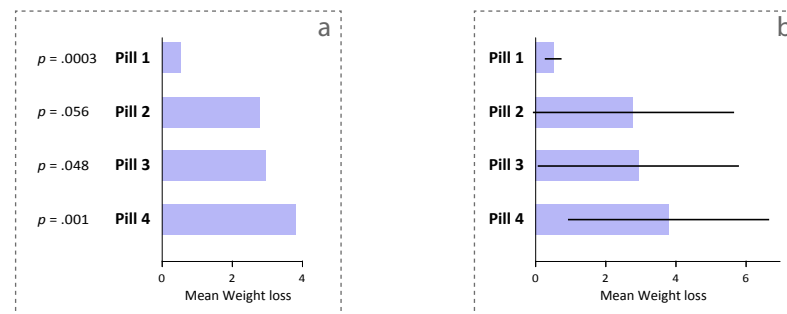

Figure 3: Showing the most plausible effect sizes and their associated uncertainty using a) *p*-values with sample means (here shown as bar charts); b) confidence intervals around sample means.

Each black dot has been replaced by a purple bar, but this is really only a matter of presentation (see Figure 8 in Section 4). The option *a* (left) follows the orthodoxy and the common recommendation to report *p*-values with effect sizes (Thompson, 1998). The option *b* (right) follows an *estimation* approach (Cumming, 2012) that consists in only reporting effect sizes with confidence intervals. In the simplest cases assuming symmetric confidence intervals, *a* and *b* are theoretically equivalent (Cumming, 2012). However, it is easier for a reader to convert from *b* to *a* (as discussed previously) than from *a* to *b*. Conversion is especially hard when confidence intervals are asymmetrical (e.g., confidence intervals on proportions, correlations, transformed data, or bootstrap confidence intervals). But most importantly, which of the two figures above is the clearest and the most informative?

Methodologists who remain attached to NHST (APA, 2010; Abelson, 1995; Levine et al, 2008b) suggest to report everything: *p*-values, effect sizes, and confidence intervals. No clear explanation has been offered on why *p*-values are needed, as the same information is already provided by confidence intervals. The recommendation to "complement" *p*-values with effect sizes and 95% confidence intervals also misleadingly suggests that effect sizes and their associated uncertainty are secondary information.

Some may still find it useful to complement a confidence interval with a *p*-value that captures precisely how far it is from zero. Later I offer arguments against this idea, which can be summarized as follows:

Figure 4: The approach consisting in reporting everything just in case.

2.1.2   The Importance of Conveying Effect Similarity

The following (imaginary) chart shows the differences between three interactive information visualization techniques in terms of average number of insights. We can safely say that *B* outperforms *A*. We can also say that *A* and *C* are similar in that they may yield a different number of average insights across the population, but the difference is likely less than 0.5. We have less information on *A* versus *D*, but we can be reasonably confident that the mean population difference is less than 2.
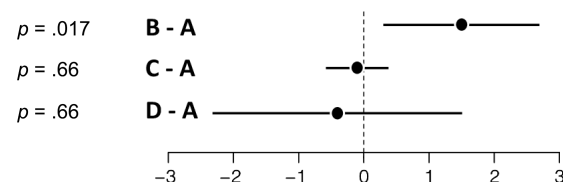
Figure 5: 95% confidence intervals showing differences between conditions.

Since the confidence interval for $C - A$ is roughly centered at zero, its *p*-value is quite high ($p = .66$). It is common knowledge that we cannot conclude anything from a high *p*-value: it only tells us that zero is plausible, and says nothing about other plausible values – the confidence interval could be of any size. In fact, the *p*-value for $D - A$ is exactly the same: $p = .66$. Knowing the sample mean in addition to the *p*-value does not help, unless it is used to reconstruct the confidence interval (assuming it is possible). Had you only focused on *p*-values and effect sizes in your study, you could have thrown almost all of your data away. Had you not tested technique *B*, you probably wouldn't have submitted anything.

Knowing that two conditions are similar is very useful. In medicine, it is important to know when a drug is indistinguishable from a placebo. In HCI, if two techniques perform similarly for all practical purposes, we want to know it. Medicine has developed equivalence testing procedures, but confidence intervals also support formal (Dienes, 2014, p.3; Tryon, 2001) as well as informal (see above) inferences on equivalence.

We can often conclude something from a confidence interval. Arguably, if an experiment does not have enough participants and/or the effect is small (i.e., the experiment has *low power*), confidence intervals can be disappointingly wide as with $D - A$, making it hard to conclude anything really useful. Confidence intervals just faithfully show the uncertainty in the experimental data. This is crucial information.

2.2   Usability Problems with p-values

So far we have mostly focused on the amount of exploitable information conveyed by *p* (i.e., on its *usefulness*), but a lot has been also written on how ineffective *p* is at conveying that information (i.e., on its poor *usability*). Recall that the task is to communicate empirical findings clearly and accurately.

2.2.1   General Interpretation Difficulties

It is safe to assume that the general public can grasp confidence intervals more easily than *p*-values. Confidence intervals simply convey the uncertainty around a numerical estimate, and they are used by the media, for example to report opinion polls (Cumming and Williams, 2011). Another important difference is that confidence intervals can be represented visually while *p*-values cannot.

One issue specific to confidence intervals is their lack of standardization. They are typically shown by the way of error bars, which can stand for many things, including standard errors (typically half the

size of 95% CIs) and standard deviations. Researchers simply need to get used to clearly indicating what error bars refer to, and if possible, consistently use 95% confidence intervals (Cumming et al, 2007).

As evidenced by numerous studies on statistical cognition (Kline, 2004; Beyth-Marom et al, 2008), even trained scientists have a hard time interpreting *p*-values, which frequently leads to misleading or wrong conclusions. Decades spent educating researchers have had little or no influence on beliefs and practice (Schmidt and Hunter, 1997, pp.20–22). Below we review common misinterpretations and fallacies. Confidence intervals being theoretically connected with *p*-values, they can also be misinterpreted and misused (Fidler and Cumming, 2005). We will discuss these issues as well, and why they may be less prevalent.

2.2.2 Misinterpretations Regarding Probabilities

p is the probability of seeing results as extreme (or more extreme) as those actually observed if the null hypothesis were true. So *p* is computed under the assumption that the null hypothesis is true. Yet it is common for researchers, teachers and even textbooks to instead interpret *p* as the probability of the null hypothesis being true (or equivalently, of the results being due to chance), an error called the "fallacy of the transposed conditional" (Haller and Krauss, 2002; Cohen, 1994, p.999).

Confidence intervals being theoretically connected with *p*-values, they are subject to the same fallacy: stating that a 95% confidence interval has a 0.95 probability of capturing the population mean is committing the fallacy of the transposed conditional. But confidence intervals do not convey probabilities as explicitly as *p*-values, and thus they do not encourage statements involving precise numerical probabilities that give a misleading impression of scientific rigor despite being factually wrong. Shown visually, confidence intervals look less rigorous. They appropriately make the researcher feel uncomfortable when making inferences about data.

A lot has been written on the fallacy of the transposed conditional, but a widespread and equally worrisome fallacy consists in ascribing magical qualities to *p* by insisting on computing and reporting *p*-values as rigorously as possible, as if they conveyed some objective truth about probabilities. This is despite the fact that the probability conveyed by *p* is only a theoretical construct that does not correspond to anything real. Again, *p* is computed with the assumption that the null hypothesis is true – i.e, that the population effect size takes a precise numerical value (typically zero) – which is almost always false (Cohen, 1994; Gelman, 2013a).

Reasoning with probabilities is possible, using Bayesian statistical methods (Downey, 2013; Barry, 2011). In particular, tools exist for computing confidence intervals that convey probabilities. This article does not address Bayesian methods and focuses on classical confidence intervals due to their simplicity, and to the fact that they need less contextual information to be interpreted. Nevertheless, several of this article's recommendations about truthful statistical communication can be combined with Bayesian data analysis.

2.2.3 Misinterpretation of High p-values

Although strictly speaking, *p*-values do not capture any practically meaningful probability, we can use them an informal measure of strength of evidence against the null hypothesis (Goodman, 1999, p.997; Gigerenzer, 2004, p.593). The closer a *p*-value is to 0, the stronger the evidence that the null hypothesis is false. If the null hypothesis is the hypothesis of zero effect and *p* is very low, we can be reasonably confident that there is an effect. But unfortunately, the closer *p* is to 1 the less we know. As seen before (see Figure 5), we cannot conclude anything from a high *p*-value, because it tells us that zero is plausible, but says nothing about other plausible values. Despite this, few researchers can resist the temptation to conclude that there is no effect, a common fallacy called "accepting the null" which had frequently led to misleading or wrong scientific conclusions (Dienes, 2014, p.1). Plotting confidence intervals such as in see Figure 5 eliminates the problem.

2.2.4 Misinterpretations Regarding Reliability

Many researchers fail to appreciate that *p*-values are unreliable and vary wildly across replications. This can be shown with simple simulations such as in the *dance of* p-*values* video (Cumming, 2009a):
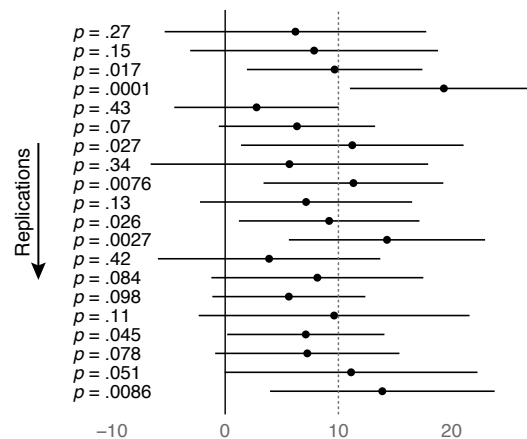

Figure 6: *p*-values and confidence intervals obtained by simulating replications of an experiment whose population effect size is 10, with a power of 0.56 (after Cumming (2009a) and Dienes (2014)).

Running an experiment amounts to closing your eyes and picking one of the *p*-values (and confidence interval) above. With a statistical power of about 0.5 common in psychology (Rossi, 1990) and HCI (Kaptein and Robertson, 2012), about any *p*-value can be obtained. The behavior of *p*-values across replications is well understood (Cumming, 2008). If an experiment yields *p* = .05 and is repeated with different participants, it will have 80% chances of yielding a *p*-value between .00008 and .44 (Cumming, 2008). This means *a 1/5 chance the new* p-*value is outside*. Even if the initial experiment yielded an impressive *p* = .001, the 80% *p*-interval would still be (.000006, .22). Sure *p* will remain appropriately low most of the time, but with such a wide range of possible values, reporting and interpreting *p* values with up to three decimal places should strike us as a futile exercise.

Many find it hard to believe that "real" *p*-values can exhibit such a chaotic behavior. Suppose you run a real study and get a set of observations, e.g., differences in completion times. You compute a mean difference, a standard deviation, and do a one-sample *t*-test. Now suppose you decide to re-run the same study with different participants, again for real. Would you expect the mean and standard deviation to come up identical? Hopefully not. Yet *p* is nothing else but a function of the mean and the standard deviation (and sample size, if not held constant). Thus *p* would come up different for the exact same reasons: sampling variability.

Any statistical calculation is subject to sampling variability. This is also true for confidence intervals, which "jump around" across replications (see Figure 6). Being fully aware of this is certainly an important prerequisite for the correct use and interpretation of confidence intervals. Watching simulated replications (e.g., from Cumming (2009a)) is enough to get a good intuition, and one can hardly claim to understand sampling variability (and thus inferential statistics) without being equipped with such an intuition. *p*-values add another layer of complexity. It is easier to remember and picture a typical dance of confidence intervals (they are all alike) than to recall all possible replication *p*-intervals. Any single confidence interval gives useful information about its whole dance, in particular where a replication is likely to land (Cumming, 2008; Cumming, 2012, Chap. 5). Any single *p*-value gives virtually no such information. There are also perceptual and cognitive differences: confidence intervals, especially shown graphically, do not give the same illusion of certainty and truth as *p*-values reported with high numerical precision.

2.3 Conclusion

There seems to be no reason to report *p*-values in a research paper, since confidence intervals can present the same information and much more, and in a much clearer (usable) manner. Perhaps the only remaining argument in favor of *p*-values is that they are useful to formally establish statistical significance. But as we will now see, the notion of statistical significance testing is a terrible idea for those who care about good statistical communication.

3 Statistical Significance Testing vs. Estimation

We previously mentioned that *statistical significance* can be quantified with *p*-values. Roughly speaking, *p*-values tell us how confident we can be that the population effect size differs from some specific value of interest – typically zero. We also explained why this notion is less useful than the orthodoxy suggests. As if the current overreliance on *p*-values was not enough, a vast majority of researchers see fit to apply a conventional (but nonetheless arbitrary) cut-off of $\alpha = .05$ on *p*-values. If *p* is less than .05, then the "results" are declared significant, otherwise they are declared non-significant (the term "statistically" is typically omitted). This is, in essence, null hypothesis significance testing (NHST).

The insights yielded by NHST procedures can be assessed by returning to our first scenario and considering again the respective merits of our four pills:
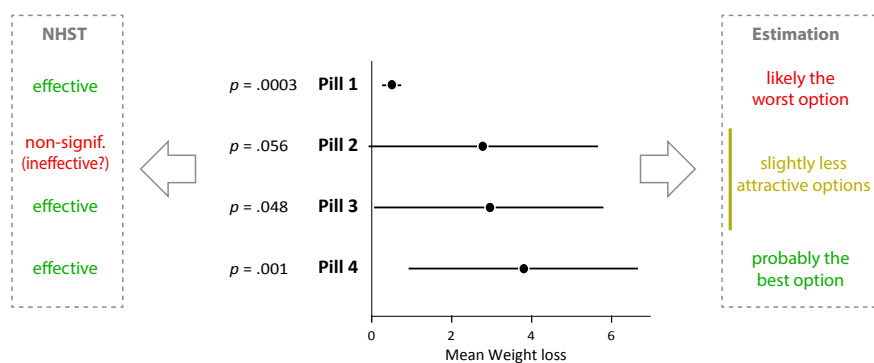

Figure 7: The same four pills, ranked based on the outcome of statistical tests (left), and based on an examination of effect sizes and confidence intervals (right).

As we saw previously, a sensible ranking scheme (shown to the right) would give a preference to pill 4, then pills 2–3 (almost identical), then pill 1. Nothing is for certain and we may well be wrong, especially about pills 2, 3 and 4 for which very little data is available. But since we need to decide on a pill we are forced to rank. In a scientific paper one would be typically way more conservative and would perhaps only comment on the likely superiority of pill 4 over pill 1. Regardless, doing statistical inference is always betting. There are good bets and bad bets.

Good bets require relevant information. The left part of the Figure shows how our decision maker would have summarized the studies had the authors focused on NHST analyses: pills 1, 3 and 4 had a statistically significant effect on weight loss ($p < .05$): they would have been presented as effective. Pill 2, on the other hand, would have been presented as having a non-significant effect[5] and despite textbook warnings against "accepting the null", the message would have almost certainly become that the pill may not be effective at all.

---

[5]This may seem caricatural but it is not: many traditionalists consider it is a serious fault to even point out that a non-significant *p*-value is very close to .05. More on this in Section 3.2.3.

A large amount of information is thrown away by the use of a cut-off. Statistical significance in its continuous form – i.e., reporting exact *p*-values – already did not carry much useful information (compared to confidence intervals). It is only logical to assume that statistical significance in its binary form cannot carry more.

3.1 Note on Doing NHST using Confidence Intervals

Importantly, confidence intervals can be used to carry out statistical tests without the need for *p*-values: examining whether or not a 95% CI contains zero is equivalent to doing a statistical test of the null hypothesis of no effect at the .05 level. This can be verified in all previous figures[6]. Thus one could perform such tests visually and then focus all discussions on statistical significance. This approach being essentially NHST, it inherits all of NHST's drawbacks. Using confidence intervals to carry out binary tests does not achieve much, and several methodologists strongly recommend nuanced interpretations that do not emphasize confidence limits (Schmidt and Hunter, 1997, p.13; Cumming, 2012). This is what estimation really refers to (Cumming, 2012), and more details will be provided in Section 4. For now, keep in mind that doing NHST using confidence intervals is *not* doing estimation.

Now, we know that both *p*-values and confidence intervals can be interpreted in a nuanced way or in a crude, binary fashion. It is a major limitation of confidence intervals that they are commonly interpreted in a binary fashion (Lai et al, 2010). This may partly stem from a design issue (an interval is black-or-white), and partly from NHST being deeply ingrained in researchers' minds (Lai et al, 2010). This is another topic on which more researcher education is needed, e.g., through exposure to alternative visual representations that show how plausibility smoothly decreases with the distance to the point estimate (Cumming, 2013, p.17; Correll and Gleicher, 2014). With this issue in mind, we now turn to usability problems with NHST as it is traditionally carried out (i.e., by computing *p*-values), and contrast it with estimation.

3.2 Usability Problems with Statistical Significance Testing

Statistical testing is based on *p*-values and therefore inherit their usability problems. The use of a binary decision rule based on a cut-off also brings a range of additional usability problems that we discuss here.

3.2.1 Misjudgments of Uncertainty

p-values give a seducing illusion of certainty and truth (Cumming, 2012, Chap.1). The sacred $\alpha = .05$ criterion amplifies this illusion, since results end up being either "significant" or "non-significant". In a researcher's mind, significant results have passed the rigorous test of statistics and are declared "valid" – uncertainty almost ceases to exist, and sample means often end up being interpreted as being exact (Vicente and Torenvliet, 2000, pp.252–258). This amounts to say that in Figure 3a, for example, each bar with $p < .05$ should be trusted fully. On the other hand, non-significant results are interpreted either as no effect or no information whatsoever, both of which are incorrect. Potential misjudgments abound and are easily dispelled by plotting confidence intervals, as in Figure 3b.

The use of a cut-off on *p* is especially problematic in studies with low statistical power, given how widely *p*-values vary across replications (see Section 2.2.4). Thus many HCI experiments really amount to tossing a coin (Dragicevic et al, 2014).

---

[6]Any point null hypothesis (not only zero) can be visually tested this way. One way to see confidence intervals is that they show the outcomes of all possible null hypothesis tests.

3.2.2   Misinterpretations Regarding Comparisons

The use of a cut-off yields disturbing paradoxes regarding comparisons of findings. The results for pills 2 and 3, for example, appear very different despite being virtually identical (Figure 7). In fact, pill 2 has close to 50% chances of coming up better than pill 3 on the next replication (remember the dance in Figure 6). This paradox causes erroneous inferences both within studies and across studies. Within studies, two conditions can be wrongly interpreted as being different, simply because one happened to pass a test while the other one did not (Gelman and Stern, 2006; Abelson, 1995, p.111). Across studies, research can appear inconsistent or controversial for the same reasons (Cumming, 2012, Chap.1). Although it has been recognized that statistical significance cannot be used as a criterion of comparison (Gelman and Stern, 2006), refraining from comparing entities that are given very different labels goes against the most basic human intuitions. The problem clearly lies not in researchers' mind, but in the tool design. With estimation, results are labeled with confidence intervals, whose comparison is not always trivial (Cumming, 2012, Chap. 6) but is certainly much less problematic[7].

3.2.3   Misinterpretations Regarding Type I Error Rates

Due to sampling error, any statistical analysis is error-prone. The key theoretical underpinning of NHST is the idea of *controlling for type I error rates*. NHST guarantees that when a null hypothesis is true (e.g., if a pill has no effect), its test will come up statistically significant – thus causing a type I error – in only 1 out of 20 experiments (with α = .05). The idea that we can take control over the likelihood of coming to wrong conclusions is appealing, but just like p and for the same reasons (see Section 2.2.2), the type I error rate is only a theoretical convenience. Contrary to common interpretations, it rarely corresponds to a real frequency or probability (Pollard and Richardson, 1987).

Some methodologists – among whom Fisher (Gigerenzer, 2004, p.593) – also reject the rationale for α on epistemological grounds, arguing that it is relevant for repeated decision making (e.g., doing quality controls in a production line), but not for assessing evidence in individual experiments (Perneger, 1998, p.1237). The notion of type I error is certainly a useful theoretical construct that can help understand key aspects of statistical inference (see Section 4.5), but there is no reason why research practice and statistical communication should be based on this theoretical framework.

3.2.4   Multiple Levels of Significance

A practice that became less popular in HCI though still advocated is the use of multiple levels of significance by the way of "post-hoc" α values (.001, .01, .05, .1), stars (***, **, *), or codified significance terminology ("highly significant", "marginally significant", etc.). This categorical approach suffers from the same problems as binary approaches, and is inconsistent with both Fisher's approach of using exact p-values as a measure of strength of evidence, and the Type I error rate approach seen above (Gigerenzer, 2004). No serious methodologist seems to endorse the use of multiple levels of significance.

3.2.5   Issues Regarding Publication Bias

Since statistical significance is a major criterion for publishing study papers, conference proceedings and journals can give a very distorted image of reality. This issue, termed *publication bias* or the *file drawer problem*, is harming science's credibility (The Economist, 2013; Goldacre, 2012). In HCI, publication bias can hamper scientific progress because results on ineffective techniques never get published, and

---

[7]For example, instead of simply saying *"we were not able to replicate previous work by Schmidt (2010) and John (2012) who found a significant improvement on task completion time"*, a conscientious researcher could say *"we found a mean improvement of 1.9 s, 95% CI [-0.7, 4.4], consistent with the improvement of 3.1 s, 95% CI [1.7, 4.7] reported by Schmidt (2010) but lower than the improvement of 5.2 s, 95% CI [4.1, 6.6] reported by John (2012)"*.

those that get published because of statistical luck or flawed analyses never get disproved. By legitimizing negative and (to some extent) inconclusive results and making publication criteria more flexible (Anderson, 2012), estimation can reduce publication bias, advance our knowledge of what does *not* work, and encourage replication (Hornbæk et al, 2014) and meta-analysis (Cumming, 2012).

3.2.6 Issues Regarding p-Hacking

Another damaging consequence of the NHST ritual is the widespread use of *"statistical convolutions [...] to reach the magic significance number"* (Giner-Sorolla, 2012). These include selectively removing outliers, and trying different testing procedures until results come up significant (Abelson, 1995, p.55). They go by various names such as *p-hacking*, *torturing data*, *data dredging*, or *researcher degrees of freedom* (Nuzzo, 2014; Lakens et al, 2014; Simmons et al, 2011; Brodeur et al, 2012). Information obfuscation can also occur after *p*-values have been computed, e.g., by selectively reporting results, using post-hoc α cut-offs (Gigerenzer, 2004), or elaborating evasive narratives (Abelson, 1995, p.55). Such practices differ from the legitimate practice of exploratory data analysis (Tukey, 1980) because their goal is to obtain the results one wishes for, not to learn or to inform. They are unscientific despite their appearance of objectivity and rigor. Humans excel at taking advantage of the fuzzy line between honest and dishonest behavior without being fully aware of it (Mazar et al, 2008). Thus merely promoting scientific integrity is likely counter-productive. To be usable, statistical tools should be designed so that they do not leave too much space for questionable practices fueled by self-deception. Estimation approaches do not draw a sharp line between interesting and uninteresting results, and thus make "torturing" data much less useful. As will be later discussed, planned analyses are another very effective safeguard.

3.2.7 Dichotomous Thinking

Humans like to think in categories. Categorical thinking is a useful heuristic in many situations, but can be intellectually unproductive when researchers seek to understand continuous phenomena (Dawkins, 2011). A specific form of categorical thinking is dichotomous thinking, i.e., thinking in two categories. Some dichotomies are real (e.g, pregnant vs. non-pregnant), some are good approximations (e.g., male vs. female, dead vs. alive), and some are convenient decision making tools (e.g., guilty vs. not guilty, legal vs. illegal drug). Many dichotomies are however clearly *false dichotomies*, and statistical thinking is replete with these. For example:

1. there is an effect or not.
2. there is evidence or not.
3. an analysis is either correct or wrong[8].

Statistical testing promotes the second dichotomy by mapping statistical significance to conclusive evidence, and non-significance to no evidence. This dichotomy is false because the degree of evidence provided by experimental data is inherently continuous. NHST procedures also promote the first dichotomy by forcing researchers to ask questions such as "is there an effect?". This dichotomy is false because with human subjects, almost any manipulation has an effect (Cohen, 1994).

There is a more insidious form of false dichotomy concerning effects. In HCI, researchers generally do not test for the mere presence of an effect, but instead ask directional questions such as "is *A* faster than *B*?". Since there is likely a difference, *A* can only be faster than *B* or the other way around. Thus the dichotomy is formally correct, but it conceals the importance of magnitude. For example, if *A* takes one second on average and *B* takes two, *A* is clearly better than *B*. But the situation is very different if *B* takes only a millisecond longer. To deal with such cases, some recommend the use of equivalence testing procedures (e.g., Dienes, 2014, p.3; Tryon, 2001). However, this does little more than turn an

---

[8]Space is lacking for addressing this particular dichotomy, but for elements of discussion see Stewart-Oaten (1995); Norman (2010); Velleman and Wilkinson (1993); Wierdsma (2013); Abelson (1995, Chap. 1) and Gigerenzer (2004, pp. 587–588).

uninformative dichotomy into a false trichotomy, as there is rarely a sharp boundary between negligible and non-negligible effects.

Thinking is fundamental to research. A usable research tool should support and promote clear thinking. Statistical significance tests encourage irrational beliefs in false dichotomies that hamper research progress – notably regarding strength of evidence and effect sizes – and their usability is therefore low. Estimation is more likely to promote clear statistical thinking.

3.2.8   Misinterpretations of the Notion of Hypothesis

Although the term *hypothesis testing* may sound impressive, there is some confusion about the meaning of a hypothesis in research. Most methodologists insist on distinguishing between *research* (or *substantive*) *hypotheses* and *statistical hypotheses* (Meehl, 1967; Hager, 2002). Briefly, research hypotheses are general statements that follow from a theory, i.e., a model with some explanatory or predictive power. Statistical hypotheses are experiment-specific statements derived from research hypotheses in order to assess the plausibility of a theory. Juggling between theories and statistical hypotheses is a difficult task that requires considerable research expertise (Meehl, 1967; Vicente and Torenvliet, 2000, pp.252–258).

Many research hypotheses are dichotomous: the acceleration of a falling object is either a function of its mass or it is not. An input device either follows Fitts' Law or some other (say, Schmidt's) law. Such dichotomies are understandable: although there is the possibility that a device follows a mix of Fitts' and Schmidt's laws, it is sensible to give more weight to the simplest models. In such situations, asking dichotomous questions and seeking yes/no answers can be sensible, and Bayesian approaches (rather than NHST) can be considered (Downey, 2013; Barry, 2011). That said, in many cases choosing a theory or a model is a decision that is partly based on data and partly based on extraneous information, so estimation methods (e.g., estimating goodness of fit) may still be beneficial in this context.

Regardless, the majority of HCI studies are *not* conducted to test real research hypotheses. That technique *A* outperforms technique *B* on task *X* may have practical implications, but this information is far from having the predictive or explanatory power of a theory. Using the term "hypothesis" in such situations is misleading: it presents a mere hunch (or hope) as something it is not, a scientific theory that needs to be tested. And as we have seen before, the respective merits of two techniques cannot be meaningfully classified into sharp categories. Estimation is enough and appropriate to assess and communicate this information, in a way that does not require the concept of hypothesis.

3.2.9   End User Dissatisfaction

NHST has been severely criticized for more than 50 years by end users to whom truthful statistical communication matters. Levine et al (2008a) offer a few quotes: *"[NHST] is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research (Rozeboom, 1960)"*; *"Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published (Likken, 1968)"*. Some go as far as saying that *"statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution"* (Schmidt and Hunter, 1997). Ten years ago, Kline (2004) reviewed more than 300 articles criticizing the indiscriminate use of NHST and concluded that it should be minimized or eliminated. Even Fisher – who coined the terms "significance testing" and "null hypothesis" – soon rejected mindless testing. In 1956 he wrote that *"no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."* (Gigerenzer, 2004). The damaging side effects of NHST use (publication bias and *p*-hacking in particular) have even led some researchers to conclude that *"most published research findings are false"* (Ioannidis, 2005).

3.3 Conclusion

The notions of *p*-value and of statistical significance are interesting theoretical constructs that can help reflect on difficult notions in statistics, such as the issues of statistical power and of multiple comparisons (both briefly covered in the next Section). However, it is widely understood that they are not good tools for scientific investigation. I – as many others before – have pointed out a range of usability problems with such tools. HCI researchers may think they can ignore these issues for the moment, because they are currently being debated. In reality, the debate mostly opposes strong reformists who think NHST should be banned (e.g., Schmidt and Hunter, 1997; Cumming, 2013) with weak reformists who think it should be *i)* de-emphasized and *ii)* properly taught and used (e.g., Abelson, 1995; Levine et al, 2008b). I already gave arguments against *i)* by explaining that *p*-values (and therefore NHST) are redundant with confidence intervals (Section 2). Concerning *ii)*, I suggested that the problem lies in the tool's usability, not in end users. This view is consistent with decades of observational data (Schmidt and Hunter, 1997, pp.3-20) and empirical evidence (Beyth-Marom et al, 2008; Haller and Krauss, 2002; Fidler and Cumming, 2005). There is no excuse for HCI to stay out of the debate. Ultimately, everyone is free to choose a side, but hopefully HCI researchers will find the usability argument compelling.

4 Truthful Statistical Communication Through Estimation

What do we do now? There are many ways data can be analyzed without resorting to NHST or *p*-values. Two frequently advocated alternatives are estimation and Bayesian methods, although the two actually address different issues and are not incompatible. There is a Bayesian version of estimation, based on credible intervals (Downey, 2013), and much of the justification and discussion of interpretation of CIs can be transferred to these methods. Here we focus on a non-Bayesian (or *frequentist*) estimation approach, because it is simple and accessible to a wide audience of investigators and readers (thus it emphasizes clarity as discussed next). Keep in mind, however, that some Bayesians strongly reject any kind of frequentist statistics, including confidence intervals (Lee, 2014; Trafimow and Marks, 2015).

From a theoretical standpoint, confidence intervals have been studied extensively, and statistical packages like *R* offer extensive libraries for computing them. But there is a lack of pedagogical material that brings all methods together in a coherent fashion. Currently there is also a lack of guidance on how to use estimation in practice, from the experiment design stage to the final scientific communication. Cumming (2012) is a good place to start for those already familiar with NHST, and hopefully we will see intro textbooks in the near future. Since the topic is extremely vast, in this section we will only discuss a few principles and pitfalls of estimation.

4.1 General Principles

Adopting better tools is only part of the solution: we also need to change the way we think about our task. Many researchers' tasks require expertise, judgment, and creativity. The analysis and communication of empirical data is no exception. This task is necessarily subjective (Thompson, 1999), but it is our job.

While we cannot be fully objective when writing a study report, we can give our readers the freedom to decide whether or not they should trust our interpretations. To quote Fisher, *"we have the duty of [...] communicating our conclusions in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."* (Fisher, 1955). This is the essence of truthful statistical communication. From this general principle can be derived a set of more basic principles:

Clarity. Statistical analyses should be as easy to understand as possible, because as implied by Fisher, one cannot judge without understanding. The more accessible an analysis is, the more the readers who can judge it, and the more free minds. Thus a study report should be an exercise of pedagogy as much as an exercise of rhetoric.

Transparency. All decisions made when carrying out a statistical analysis should be made as explicit as possible, because the results of an analysis cannot be fairly assessed if many decisions remain concealed (see *p*-hacking in Section 3.2.6).

Simplicity. When choosing between two analysis options, the simplest one should be preferred even if it is slightly inferior in other respects. This follows from the principle of clarity. In other words, the KISS principle (Keep It Simple, Stupid) is as relevant in statistical communication as in any other domain.

Robustness. A statistical analysis should be robust to sampling variability, i.e., it should be designed so that similar experimental outcomes yield similar results and conclusions[9]. This is a corollary of the principle of clarity, as any analysis that departs from this principle is misleading about the data.

Unconditionality. Ideally, no decision subtending an analysis should be conditional on experimental data, e.g., *"if the data turns out like this, compute this, or report that"*. This principle may seem less trivial than the previous ones but it follows from all of them. It is a corollary of the principles of clarity, transparency and simplicity, because conditional procedures are hard to explain and easy to leave unexplained. It is also a corollary of the principle of robustness because any dichotomous decision decreases a procedure's robustness to sampling variability.

Precision. Even if all the above precautions are taken, a study report where nothing conclusive can be said would be a waste of readers' time, and may prompt them to seek inexistent patterns. High statistical *power* (Cohen, 1990), which in the estimation world translates to statistical *precision* (Cumming, 2012, Chap.5), should also be a goal to pursue.

4.2 Before Analyzing Data

Experiment design and statistical analysis are tightly coupled. A bad experiment cannot be saved by a good analysis (Drummond and Vowler, 2011, p.130). Many textbooks provide extensive advice on how to conduct research and design experiments, and most of it is relevant to estimation research. Here are a few tips that are particularly relevant to estimation methods and can help ensure truthful statistical communication.

Tip 1: Ask focused research questions. Ask clear and focused research questions, ideally only one or a few, and design an experiment that specifically answers them (Cumming, 2012). This should result in a simple experiment design (see Tip 2), and make the necessary analyses straightforward at the outset (see Tip 5).

Tip 2: Prefer simple designs. Except perhaps in purely exploratory studies and when building quantitative models, complex experiment designs – i.e., many factors or many conditions per factor – should be avoided. These are hard to analyze, grasp and interpret appropriately (Cohen, 1990). There is no perfect method for analyzing complex designs using estimation (Franz and Loftus, 2012; Baguley, 2012), and even NHST procedures like ANOVA that have been specifically developed for such designs are not without issues (Smith et al, 2002; Baguley, 2012; Kirby and Gerlanc, 2013, p.28; Rosnow and Rosenthal, 1989, p.1281; Cumming, 2012, p.420). Faithfully communicating results from complex designs is simply hard, no matter which method is used. Best is to break down studies in separate experiments, each answering a specific question. Ideally, experiments should be designed sequentially, so that each experiment addresses the questions and issues raised by the previous one.

Tip 3: Prefer within-subjects designs. While not always feasible, within-subjects designs yield more statistical precision, and also facilitate many confidence interval calculations (see Tip 10).

Tip 4: Prefer continuous measurement scales. Categorical and ordinal data can be hard to analyze and communicate, with the exception of binary data for which estimation is routinely used (Newcombe, 1998a,b). Binary data, however, does not carry much information and thus suffers from low statistical

---

[9]The meaning of *robust* here differs from its use in *robust statistics*, where it refers to robustness to outliers and to departures from statistical assumptions.

precision (Rawls, 1998). For measurements such as task errors or age, prefer continuous metrics to binary or categorical scales[10].

Tip 5: Plan all analyses using pilot data. It is very useful to collect initial data, e.g, from co-authors and family, and analyze it. This makes it possible to debug the experiment, refine predictions, and most importantly, plan the final analysis (Cumming, 2013). Planned analyses meet the unconditionality principle and are way more convincing than post-hoc analyses because they leave less room for self-deception and prevent questionable practices such as "cherry picking" (see Section 3.2.6). An excellent way to achieve this is to write scripts for generating all confidence intervals and plots, then collect experimental data and re-run the same scripts. Pilot data should be naturally thrown away. If all goes well, the researcher can announce in her article that all analyses were planned. Additional post-hoc analyses can still be conducted and reported in a separate "Exploratory Analysis" subsection.

Tip 6: Twenty participants is a safe bet. The idea that there is a "right" number of participants in HCI is part of the folklore and has no theoretical basis. If we focus on meeting statistical assumptions and ignore statistical precision (discussed next), about *twenty participants* put the researcher in a safe zone for analyzing numerical data with about any distribution (see Tip 13). If all scales are known to be approximately normal (e.g., logged times, see Tip 12), exact confidence intervals can be used and the lower limit falls to *two participants* (Forum, 2015).

Tip 7: Anticipate precision. High precision means small confidence intervals (Cumming, 2012, Chap.5). When deciding on an appropriate number of participants, the most rudimentary precision analysis is preferable to wishful thinking. One approach consists in duplicating participants from pilot data (see Tip 5) until confidence intervals get small enough. How small is small enough? At the planning stage, considering whether or not an interval is far enough from zero is a good rule of thumb. This criterion has so much psychological influence on reviewers that it is not unreasonable to try to meet it. But cursing the fate in case it fails is not worthy of an estimation researcher.

Tip 8: Hypotheses are optional. Hypotheses have their place in research, when they are informed by a theory or by a careful review of the past literature. However, it is often sufficient to simply *ask* questions. Reporting investigators' initial expectations does benefit transparency, especially since these can bias results (Rosenthal and Fode, 1963; Rosenthal, 2009). But they do not need to be called hypotheses. Also, expectations can change, for example after a pilot study (see Tip 5). This is part of the research process and does not need to be concealed. Finally, having no hypothesis or theory to defend departs from typical narratives such as used in psychology (Abelson, 1995), but admitting one's ignorance and taking a neutral stance seems much more acceptable than fabricating hypotheses after the fact (Kerr, 1998).

4.3 Calculating Confidence Intervals

About any statistical test can be replaced with the calculation of a confidence interval. The counterpart of a classic *t*-test is termed (a bit misleadingly) an *exact confidence interval* (Cumming, 2012). There is not much to say about calculation procedures, as they are extensively covered by textbooks and on the Web. Here are a few tips that are not systematically covered by theoretical material on confidence intervals. Some of them are in contradiction with orthodox practices as popularized by textbooks, but they are in better accordance with truthful statistical communication and are supported by compelling arguments from the methodology literature. I have also tried to include common pitfalls I have observed while working with students.

---

[10]There is considerable debate on how to best collect and analyze questionnaire data, and I have not gone through enough of the literature to provide definitive recommendations at this point. Likert scales are easy to analyze if they are measured as they should, i.e., by combining responses from multiple items (Carifio and Perla, 2007). If responses to individual items are of interest, it is often sufficient to report all responses visually (see Tip 22). Visual analogue scales seem to be a promising option to consider if inferences need to be made on individual items (Reips and Funke, 2008). However, analyzing many items individually is not recommended (see Tips 1, 5 and 30).

Tip 9: Always as many observations as participants. Perhaps the only serious mistake that can be made when computing confidence intervals is by using more observations than participants. Often HCI experiments involve multiple successive observations, such as completion times across many trials. These need to be aggregated (e.g., averaged) so that each participant ends up with a single value (Lazic, 2010). This is because – except in rare cases – the purpose of statistical inference in HCI is to generalize data to a population of people, not of trials[11] (see Section 2). NHST has developed notations that make it possible for statistics-savvy readers to spot such mistakes, but estimation has not. Thus it is good practice to mention the number of observations involved in the computation of confidence intervals, either in the text or in figure captions (e.g., *n=20*).

Tip 10: Feel free to process data. As long as Tip 9 is observed, it does not matter how the per-participant values were obtained. Raw observations can be converted into different units (see Tip 12) and be aggregated in any way: arithmetic means, geometric means, sums, or percentages. With within-subject designs, new data columns can be added to capture averages across several conditions, differences between conditions, differences between differences (i.e., interactions), or even regression coefficients for learning effects. There is nothing sacred about raw measurements (Velleman and Wilkinson, 1993, pp.8-9), and these can be processed in any way as long as the numbers reflect something *meaningful* about participants' performance, answer a relevant research question (Tip 1), and all calculations have been planned (Tip 5).

Tip 11: Avoid throwing data away. Data can be discarded for good reasons, e.g., when a researcher ignores certain effects to achieve a more focused analysis (Tip 1). But data can also be discarded pointlessly, e.g., by turning continuous measurements into discrete or binary values through binning (see Tip 4). This results in a loss of information, and therefore of statistical precision, and possibly biased results (Rawls, 1998; MacCallum et al, 2002). Discarding observations beyond a certain range (*truncation*, see Ulrich and Miller (1994)) or based on standard deviation (*restriction*, see Miller (1991)) can help eliminate spurious observations, but can also result in a loss of precision or in bias (Miller, 1991; Ulrich and Miller, 1994). Discarding observations based on rank (*trimming*, see Wilcox (1998), of which the *median* is a special case) can in some cases increase precision (Wilcox, 1998), but for approximately normal distributions nothing beats the mean (Wilcox, 1998). In general there is considerable disagreement on how to discard observations, and whether this should be done at all (see Osborne and Overbay (2004) for a favorable stance), but the simplicity principle would suggest to simply ignore such procedures.

Tip 12: Consider the log transform. The log transform corrects for positive skewness in completion time data and gives less weight to extreme measurements, thus rendering outlier removal unnecessary (Sauro and Lewis, 2010). Another nice consequence is that it yields asymmetric confidence intervals, which better convey the underlying distributions and prevents the embarrassing situation where a confidence interval extends to negative values. The procedure consists in turning all raw measurements into logs, doing all analyses as usual, then converting back (*antilogging*) the means and interval limits at the very end, when they need to be presented numerically or graphically (Gardner and Altman, 1986, p.749). All means will indicate geometric (instead of arithmetic) means, and differences between means become ratios between geometric means (Gardner and Altman, 1986, p.750). As it turns out, ratios between completion times are often more meaningful than differences (Dragicevic, 2012). No justification or test is needed for using a log transform on time measurements (Keene, 1995) (see also Tip 14).

Tip 13: Consider bootstrapping. One extremely useful method that has not received enough attention is *bootstrapping*, which produces *bootstrap confidence intervals* (Kirby and Gerlanc, 2013; Wood, 2005, 2004). This method is recent in the history of statistics because it requires computers, but it is very general and works for all kinds of distributions (Kirby and Gerlanc, 2013). Most importantly, bootstrapping has many benefits in terms of truthful communication, as the basic method is algorithmically trivial. Thus it promotes computational transparency in statistical reports (Kirby and Gerlanc, 2013, p.6;

---

[11]Both types of inference can be combined by using hierarchical (or multi-level) models (Gelman and Hill, 2006, Chap. 11). Thanks to Maurits Kaptein for pointing this out.

Wood, 2005, p.455), and also shows remarkable promise in statistics teaching (Ricketts and Berry, 1994; Duckworth and Stephenson, 2003). The focus on simulation rather than analytical calculation should be particularly compelling to the computer scientists in HCI. Bootstrap confidence intervals are generally accurate with about 20 observations or more (Kirby and Gerlanc, 2013, p.8), but tend to be a bit narrow with 10 or less (Wood, 2005, p.467).

Tip 14: Do not test for normality. The world is not sharply divided into normal and non-normal distributions. This false dichotomy has been largely promoted by normality testing procedures, which violate the unconditionality principle and are both logically and practically unsound (Wierdsma, 2013; Stewart-Oaten, 1995, p.2002). It is enough to keep in mind that *i) normality is not such a big deal*: like the *t*-test, exact confidence intervals are robust against departures from normality (Norman, 2010), and methods for non-normal distributions are obviously able to accommodate normal distributions; *ii) distributions are rarely entirely unknown*: measurements that are strictly positive (e.g., time) or bounded (e.g, percents) cannot be normally distributed. Strictly positive scales are almost always positively skewed and approximate a normal distribution once logged (Tip 12); *iii) when in doubt*, use bootstrapping (Tip 13).

Tip 15: Report confidence intervals for about everything. Any statistic is subject to sampling variability, not only sample means. A report should complement *all* statistics from which inferences are made – including standard deviations, correlation coefficients, and linear regression coefficients – with confidence intervals that convey the numerical uncertainty around those estimates. Many sources are available in textbooks and online on how to compute such confidence intervals.

4.4   Plotting Confidence Intervals

Confidence intervals can be conveyed numerically, or graphically by the way of error bars. There exists a standard numerical notation (APA, 2010, p.117), but no well-established standard for representing confidence intervals graphically. The tips I include here emphasize truthful statistical communication and most of them are, I believe, based on common sense. As before, I have tried to include common pitfalls.

Tip 16: Prefer pictures. Graphic formats for confidence intervals effectively convey magnitudes and patterns (Fidler and Loftus, 2009). Some would consider this as a disadvantage as many such patterns can be spurious, but in reality, plotting confidence intervals greatly facilitates judgment. For example, how different is 2.9 kg, 95% CI [0.02, 5.8] from 2.8 kg, 95% CI [-0.08, 5.7]? Or from 4.5 kg, 95% CI [2.5, 6.5]? The answers are in Figure 2. Graphical representations are not inherently more misleading, they just conceal less. Moreover, visual inference appropriately suggests imprecision.

Tip 17: Use numbers wisely. If plots are already provided, numerical values are not very useful. However, the numerical format being more compact, it can be useful for reporting secondary results. Key effect sizes can also be restated numerically to facilitate comparison with future studies. Similarly, a complete list of numerical confidence intervals can be included in the accompanying material to facilitate replication and meta-analysis. But in the article itself, refrain from reporting an absurdly high number of significant digits, e.g., 2.789 kg, 95% CI [-0.0791, 5.658].

Tip 18: Stick to classical CIs. Avoid showing standard errors (SEs) in your plots. As Cumming and Finch (2005, p.177) pointed out, *"if researchers prefer to publish SE bars merely because they are shorter, they are capitalizing on their readers' presumed lack of understanding of SE bars, 95% CIs, and the relation between the two."*. When plotting confidence intervals, aim for simplicity and stick to the few existing conventions. Figures should be interpretable with as little contextual information as possible. Changing the level of confidence to 50% or 99% does not help. The same can be said about procedures that "adjust" or "correct" the length of confidence intervals. Several such procedures have been described to facilitate visual inference or reinforce the equivalence with classical NHST procedures (Baguley, 2012; Tryon, 2001; Bender and Lange, 2001), but their downside is that they change the meaning of confidence intervals and increase the amount of contextual information required to interpret them.
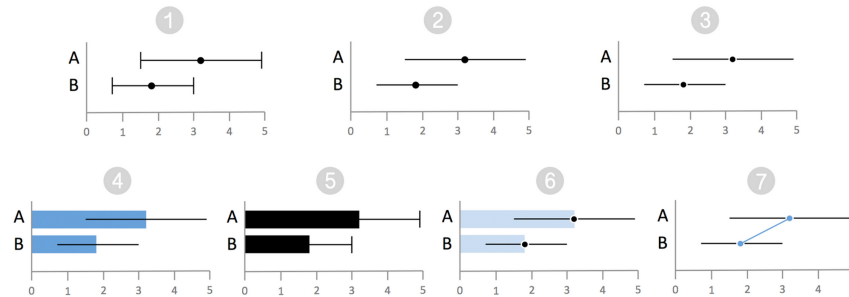
Figure 8: Seven ways of plotting effect sizes with confidence intervals.

Tip 19: Be creative. The absence of well-established visual standards should be taken as an opportunity to explore alternative representations, within the limits suggested by Tip 18. Figure 8 shows a number of options for displaying error bars: the error bars in (1) are perhaps the most widely used. (2) is a common alternative that has the advantage of de-emphasizing confidence limits (and has been adopted by Geoff Cumming). (3) is a variant that increases the legibility of point estimates. Confidence intervals can also be combined with bar charts (4). Bars help compare magnitudes on a ratio scale (Zacks and Tversky, 1999) but introduce visual asymmetry (Newman and Scholl, 2012; Correll and Gleicher, 2014) and tend to de-emphasize confidence intervals. The option (5) called "dynamite plot" certainly does not help. The design (6) supports ratio comparison while maintaining emphasis on confidence intervals. Finally, confidence intervals can be combined with line charts (7) to show temporal data (Zacks and Tversky, 1999), but line connectors are also a convention for conveying within-subject data (Cumming, 2012, p.172).
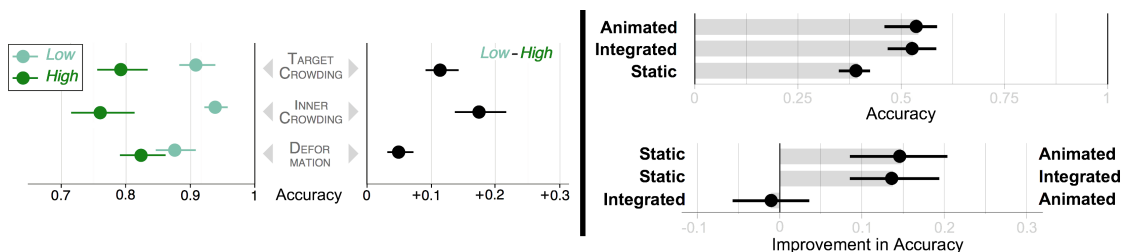
Figure 9: *Left:* the effects of three animation complexity metrics (one per row) on visual tracking accuracy. The left plot shows mean subject accuracy depending on whether animations are low or high on that metric, while the right plot shows mean within-subject improvement when switching from high to low (after Chevalier et al (2014)). *Right:* map reading accuracy for three terrain visualization techniques. Below, each row shows mean within-subject improvement when switching from the left technique to the right one (the scale has been magnified) (after Willett et al (2015)). All error bars are 95% CIs, *n*=20.

Tip 20: Emphasize effects of interest. When choosing what to plot, focus on the effects that answer your research questions (Cumming, 2012). These are typically differences between means, e.g., differences in average task average completion times. In within-subject designs, differences are typically computed on a per-participant basis, as with the paired *t*-test (Cumming, 2012, pp.168–175) (also see Tip 10). For multiple comparisons (see Tips 30 and 31), stick to the most informative pairs. For a nice example of informative pairwise comparisons, see the second experiment in (Jansen, 2014, Chap. 5). Even though individual means are rarely the researcher's focus, it can be very informative to show them alongside the differences (Franz and Loftus, 2012). Doing so also has an explanatory value, and thus contributes to clarity. See Figure 9 for two examples.

Tip 21: Aim for visual robustness. By *visually robust*, I refer to visual representations that are not overly affected by small changes in data[12], and are thus resistant to sampling variability. While it is hard to make a plot visually more robust without discarding information, there are many ways to make it less robust without adding any new information. One is by sorting conditions or pairwise comparisons by effect size. If effects are comparable, every replication would lead to a totally different order. Thus the figure would be misleading, and best is to choose a sensible ordering ahead of time. For similar reasons, boxplots (Wickham and Stryjewski, 2011) lack visual robustness because they embed dichotomous decisions as to whether an observation should be considered as an "outlier". The resulting dots draw unnecessary attention to observations that just happen to be at the tails of the population distribution (Wickham and Stryjewski, 2011, pp.3–5). When designing a figure, always try to imagine how it would "dance" across replications.

Tip 22: Consider showing all the data. Some empirical data – especially categorical or ordinal data like Likert item responses – is hard to faithfully capture with a single aggregated measure. Distributions and differences between participants, which can be very informative (Vicente and Torenvliet, 2000, pp.250–253), also get lost through the process of statistical inference. Data from questionnaire results can be conveyed without any loss of information using compact visualization methods such as matrix displays (Perin et al, 2014). Showing individual observations next to confidence intervals can also be informative as well as pedagogical (Drummond and Vowler, 2011; Ecklund, 2012).

4.5   Interpreting Confidence Intervals

Interpreting confidence intervals is a key aspect of estimation. It is not an easy task, and it could almost be called an "art". Despite this, not much has been written on the topic. Here is a list of principles that can help interpret plots with confidence intervals, and since this is both an important and an error-prone task, some of the tips here will be developed more extensively than the previous ones.

Tip 23: Build strong intuitions. The key to a correct interpretation of confidence intervals is a deep understanding of their relationship to sampling variability, and of sampling variability itself. Simulated replications like shown by Cumming (2009a) are a powerful tool for building such intuitions, perhaps more so than mathematical formulas. As with language learning, improvement comes with repeated exposure: watch simulations over and over again, and do not hesitate to run your own.

Tip 24: Know inference by eye. Cumming offers useful rules of thumb for doing statistical inference "by eye" (Cumming and Finch, 2005; Cumming, 2009b). The most basic rule follows from the equivalence between CIs and NHST, that is, if a certain value is outside a 95% CI, it would be rejected as a null hypothesis at the α=.05 level. This is only a convenient reference point, *not a hard rule* to be applied mindlessly (see Tip 25). Similarly, Cumming offers the "overlap rule" for visually comparing two confidence intervals in between-subjects designs. This rule also is nothing more than a convenient rule of thumb, and should not be used to draw dichotomous conclusions. In within-subject designs, the overlap rule is generally (but not always) conservative. It is thus important to plot and interpret confidence intervals on pairwise differences (Cumming and Finch, 2005, p.176) (See Figure 9).

Tip 25: Ban dichotomous interpretations. Using confidence intervals to provide yes/no answers defeats the whole purpose of estimation. There is enough literature to resist pressure from misinformed reviewers: *"It seems clear that no confidence interval should be interpreted as a significance test"* (Schmidt and Hunter, 1997, p.3-15); *"CIs can prompt better interpretation if NHST is avoided"* (Cumming, 2013, p.17). The point estimate is the best bet for the population mean, and plausibility smoothly decreases with distance to the point estimate – the edges of a 95% CI being about seven times less plausible (Cum-

---

[12]*Visual robustness* is related to the concept of *visual-data correspondence* recently introduced in information visualization (Kindlmann and Scheidegger, 2014). The counterpart of robustness, i.e., a visualization's ability to reveal differences in data, has been variously termed *distinctness* (Rensink, 2014), *power* (Hofmann et al, 2012), and *unambiguity* (Kindlmann and Scheidegger, 2014).

ming, 2012, p.100). There is no sharp decrease when crossing the edges. Their location depend on the confidence level chosen, and there is clearly nothing magic about them. The "cat's eyes" representation can help fully realize this (Cumming, 2012, pp.98–102; Cumming, 2013, p.17). While some have recently suggested to adopt such representations in figures (Correll and Gleicher, 2014), the classical error bar with the conventional 95% level has the benefit of being visually cleaner and more economical in terms of space and data-ink ratio (Wainer, 1984, p.139). Furthermore, the edges of error bars offer convenient visual reference points, whereas other representations may not. In particular, gradients as suggested in (Correll and Gleicher, 2014) are perceived differently depending on contrast settings.

Tip 26: Use vague language and hedges. We could say that Figure 5 *"provides good evidence that B outperforms A, whereas C and A seem very similar, and results are largely inconclusive concerning the difference between D and A."*. The terms "good evidence", "very similar" and "largely inconclusive" are *vague*. The use of vague language is appropriate for expressing effect sizes and degrees of evidence conveyed by confidence intervals. Vague language – which is different from ambiguous language – plays an important role in reasoning (van Deemter, 2010). It is not unscientific, only false clarity (van Deemter, 2010, Ch. 1) and false objectivity (Thompson, 1999) are. The term "seem" in the text above is a *hedge*, and hedges are important in scientific communication (van Deemter, 2010, Ch. 6). There are many ways Figure 5 can be turned into textual form, and different investigators would come up with different wordings. The subjective nature of this task should not make the researcher feel uncomfortable. It is important to be objective when doing planned analyses and turning them into numbers and plots, but after that the job is done. One can then afford being subjective, knowing that no reader is forced to accept one's conclusions. That said, interpretations that are overly biased or clearly misleading naturally tend not to give a very good image of their authors.

Tip 27: Never say "no effect". Avoid suggesting that there is no effect, even using hedges (e.g., *"results suggest that A has no effect on B"*). Almost any experimental manipulation has *some* effect on human behavior (Cohen, 1994), so an effet of exactly zero is highly implausible. Better wordings include *"the direction of the effect is uncertain, but it is likely small"*, or *"we were not able to measure an effect"*.
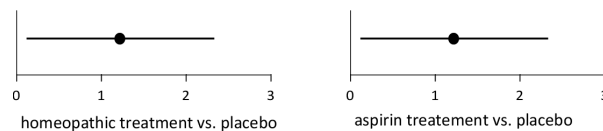

Figure 10: Identical confidence intervals calling for different interpretations.

Tip 28: Use external information. A key part of the exercise of interpreting empirical data consists in putting it into context with externally available information (Perin et al, 2014). For example, the imaginary plot on Figure 10 (left) shows the results of a study assessing the clinical efficacy of a homeopathic treatment. While the data speaks in favor of it, the scientific consensus is that such medications are ineffective, thus results should be interpreted with skepticism. Following the skeptical inquiry principle that *"extraordinary claims require extraordinary evidence"*, perhaps the investigator should require that the confidence interval be much further from zero. In contrast, the right plot concerns a drug whose efficacy has been already firmly established, so one just needs to see it as a successful replication. The two results are *identical*, yet their interpretation is *very different*. Although Bayesian statisticians would typically attempt to incorporate such knowledge into the statistical analysis itself, not every reader needs to agree on the *a-priori* plausibility of a scientific result. Confidence intervals have both the drawback and the advantage of moving the burden of Bayesian reasoning to the investigator and the readers. Another crucial piece of external information (rarely accounted for in Bayesian analysis) is the *cost of error*. While HCI studies on safety-critical systems require cautious interpretations, excessive caution can slow down exploration and be detrimental to progress in studies that simply investigate new user interface technologies (i.e., most of HCI).
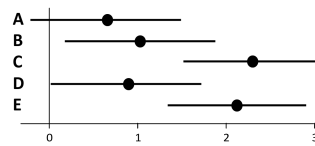
Figure 11: In this plot, each confidence interval needs to be interpreted in the context of other intervals.

Tip 29: Use internal information. Individual confidence intervals should also be interpreted according to internal information, i.e., other pieces of information obtained from the same study. For example, in Figure 11, there is nothing wrong in interpreting *A* as providing *"only moderate evidence for an effect"*. However, it would be misleading to then interpret *B* as exhibiting *"clear evidence for an effect"*. The principle of robustness requires that two similar results are interpreted in a similar way (also see Section 3.2.2 for the danger of not doing so). Perhaps the least misleading way of commenting Figure 11 would be to group results, e.g., by stating that the figure provides evidence *"for a small effect for A, B, D, and for a larger effect for C and E."*. Also see Tip 24 on how to read overlaps in confidence intervals. On the other hand, care must be taken *not* to suggest that the effects within each group are the same or similar. If pairwise differences are of interest, they should be interpreted separately. Elaborating non-misleading interpretations of figures with confidence intervals requires judgment, and no mechanical decision procedure exists that can carry out this job better than a thoughtful investigator.
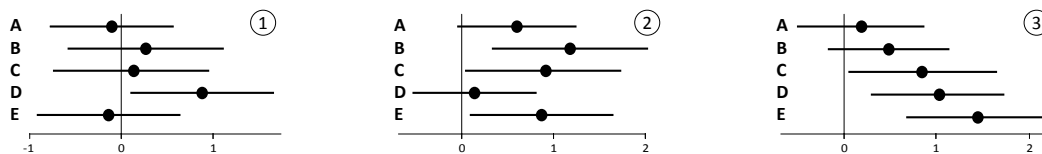

Figure 12: Three possible sets of confidence intervals that can be tricky to summarize.

Tip 30: Combine results very carefully. Drawing conclusions from a set of confidence intervals without any consideration for joint probabilities is particularly dangerous. Suppose that a person with purported extrasensory perception (ESP) abilities is subjected to testing sessions labeled *A-E* (Figure 12-1). The investigator, observing that test *D* appears positive, may conclude that ESP is real. But even if ESP does not exist, the chances of observing *D* is about 5%, already not that low. The chances that a similar deviation is observed in *any* of the five tests is even higher: about 20%. With 40 tests, it would have been about 90%. An HCI researcher who concludes that a technique is promising because it succeeded in one out of five tasks commits the same fallacy. The converse fallacy (unsurprisingly less common) is also possible: a researcher observing Figure 12-2 may conclude that the technique is imperfect because one task yields inconclusive results. But such an outcome is likely to occur even if the technique is better in all respects (e.g., a consistent population difference of one – see again the dance in Figure 6).

Disjunctive (*or* operator) and conjunctive (*and* operator) logical combinations are best avoided when interpreting multiple results. Results should be *averaged*, or expressed with *buts* (Abelson, 1995, Ch. 6): e.g., *"results are inconclusive except perhaps for one task out of five"*, and *"the technique seems consistently better, although possibly not for one out of the five tasks"*. Nuanced findings should *not* be oversimplified in the paper's abstract or conclusion. Finally, it happens that a striking pattern emerges from a figure, as in Figure 12-3. Such a linear trend is unlikely to emerge from a dance of confidence intervals (unless conditions have been sorted, see Tip 21), so the investigator should not be shy to point it out. Nevertheless, best is to express a pattern with a single number (e.g., a grand mean, difference, interaction, correlation, or regression slope) and compute a CI for it. Such analyses are however largely uninformative if they have not been planned (see Tip 5).

Tip 31: Do not correct for multiplicity. Correction procedures (e.g., Bonferroni correction) that account for the scenario in Figure 12-1 are part of the NHST ritual, and have been adapted to confidence intervals (Bender and Lange, 2001; Tryon, 2001). However, such procedures do not account for the many situations where results are not combined in a disjunctive manner (see Tip 30) and they have the drawback of changing the meaning of error bars (Cumming, 2009b, p.209) (see Tip 18). Besides, multiplicity correction procedures are far from having reached a consensus (Bender and Lange, 2001) and the mere idea of systematically correcting for multiple tests has been criticized as logically and practically unsound (Wilson, 1962; Stewart-Oaten, 1995, p.2003). Between the fervent defenders and the strong opponents, many methodologists take a laissez-faire position, especially among those advocating estimation (Cumming, 2012, pp.421–423; Cumming and Finch, 2005, p.177; Baguley, 2012, pp.173-174). The principle of simplicity provides an argument for avoiding such procedures, while the principles of transparency and clarity requires that issues related to disjunctive conclusions are kept in mind and pointed out. Best is to avoid reporting many confidence intervals by keeping experiments simple (Tip 2) and planning all analyses ahead of time (Tip 5) based on clear research questions (Tip 1).

Tip 32: Point out possibly spurious patterns. For a researcher who is committed to truthful statistical communication, it is not enough to write an irreproachable analysis. Such a researcher should also anticipate possible misinterpretations of figures by statistically less sophisticated readers and take the necessary precautions.

Tip 33: Defer judgment. NHST and the idea of statistical significance made us forget that no single study can provide conclusive evidence. Although this article focuses on uncertainty due to statistical error, there are many other sources of uncertainty (Brewer, 2000; Meehl, 1967; Rosenthal, 2009; Vicente and Torenvliet, 2000, pp.264–266). According to Schmidt and Hunter (1997, p.3-16) *"it is best for individual researchers to [...] refrain from attempting to draw final conclusions about research hypotheses."*. This echoes Fisher's belief that we should grant our readers the right of *"making their own decisions"*. It is fine for a study to *"provide strong evidence for X"*, but not to *"show that X"*. Authors tend to be especially oversimplifying and brash in conclusions and abstracts, the very parts on which hurried readers (and journalists of course) tend to focus their attention. Using hedges as suggested in Tip 26 and providing nuanced conclusions as suggested in Tip 30 can help.

Tip 34: Share study material. Finally, sharing as much experimental material as possible (stimuli and data) is important as it greatly facilitates peer scrutiny and replication. Being able to run the experimental software and examine what participants saw (the techniques, tasks, instructions, and questions asked) is necessary for researchers to understand the details of a study and greatly facilitates replication. Experimental data (all data tables and possibly analysis scripts) is necessary for conducting re-analyses and meta-analyses. To be useful this material should not be simply attached to the submission, but also published online upon paper acceptance.

For new and updated tips, go to `www.aviz.fr/badstats`.

5 Conclusion

When judging the quality of a statistical analysis in a paper, reviewers often tend to exclusively focus on the correct application of procedures and the precision of calculations. This reveals several misconceptions about the role of statistics in HCI and more generally in science. One is the belief that there exists a set of standard, well-accepted, "correct" statistical procedures, a myth largely cultivated by textbooks and introductory statistics courses (Gliner et al, 2002). Gigerenzer (2004, pp. 587–588) tells the story of a textbook author who was forced by his editor not to mention non-orthodox methods in order to produce *"the illusion that there is only one tool"*. Another belief is that statistical procedures can produce precise and rigorous knowledge, just because they can spit precise numbers. This ignores the fact that no statistical procedure can produce accurate knowledge, for the simple reason that this information is not in the data. Data is uncertain and messy, and so are stats.

In this article I tried to introduce some very basic principles of truthful statistical communication, i.e., principles for conveying uncertainty in empirical data in a way that does not prompt misinterpretations, so that an as-wide-as-possible audience is able to judge and decide whether or not to trust the authors' conclusions. Among the worst enemies of truthful statistical communication is dichotomous thinking. Dichotomization exists at many levels in current statistical practice, and has been even legitimized and encouraged by NHST procedures. Judging articles based on the outcomes of such procedures reveals a deep misunderstanding of the purpose of statistics and is damaging to science, because it encourages questionable practices, information obfuscation and publication bias. Dropping mindless statistical testing procedures, trying to achieve truthful statistical communication and rethinking how we should communicate scientific results (Giner-Sorolla, 2012) can not only benefit science, but can also make statistical analysis a much more exciting and rewarding experience.

Such issues are not specific to HCI; They are old and have been raised in many disciplines. But things seem to be changing: in the last edition of its highly influential "Publication Manual" (APA, 2010), the American Psychological Association deemphasizes *p*-values and stresses the importance of reporting and interpreting effect sizes and confidence intervals (Fidler et al, 2010). High-impact psychology journals are starting to enforce the use of estimation techniques (Eich, 2014). Large collective initiatives, whose values overlap a great deal with the idea of truthful statistical communication, are being started under the umbrella term of "open science" movement[13]. In parallel, more and more articles on bad science and bad stats get published online, raising the public's awareness of those issues. It may well be that stats will be done very differently in a few years from now.

Since stats are nothing but "user interfaces" meant to help researchers in their task of producing and disseminating knowledge, the fields of HCI and infovis should logically take a head start and show the way to other disciplines. HCI and infovis researchers also have the exciting opportunity to contribute new research, for example by studying new visual representations to communicate study results (Correll and Gleicher, 2014; Perin et al, 2014). Such representations do not have to be static, and there are many ways computers, animation and interactivity can be used to teach statistics and convey scientific evidence to a wide audience (Victor, 2011).

6   Acknowledgements

References

Abelson R (1995) Statistics as principled argument. Lawrence Erlbaum Associates

Anderson G (2012) No result is worthless: the value of negative results in science. Online, URL `http://tinyurl.com/anderson-negative`, last accessed Mar 2015

APA (2010) The Publication manual of the APA (6th ed.). Washington, DC

Baguley T (2009) Standardized or simple effect size: What should be reported? British Journal of Psychology 100(3):603–617

Baguley T (2012) Calculating and graphing within-subject confidence intervals for ANOVA. Behavior Research Methods 44(1):158–175

---

[13]see, e.g., `http://centerforopenscience.org/` and `https://osf.io/`.

Barry J (2011) Doing bayesian data analysis: A tutorial with R and BUGS. Europe's Journal of Psychology 7(4):778–779

Beaudouin-Lafon M (2008) Interaction is the future of computing. In: Erickson T, McDonald DW (eds) HCI Remixed: Reflections on Works That Have Influenced the HCI Community, The MIT Press, pp 263–266

Bender R, Lange S (2001) Adjusting for multiple testing—when and how? Journal of clinical epidemiology 54(4):343–349

Beyth-Marom R, Fidler F, Cumming G (2008) Statistical cognition: Towards evidence-based practice in statistics and statistics education. Statistics Education Research Journal 7(2):20–39

Brewer MB (2000) Research design and issues of validity. Handbook of research methods in social and personality psychology pp 3–16

Brodeur A, Lé M, Sangnier M, Zylberberg Y (2012) Star wars: The empirics strike back. Paris School of Economics Working Paper (2012-29)

Carifio J, Perla RJ (2007) Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. Journal of Social Sciences 3(3):106

Chevalier F, Dragicevic P, Franconeri S (2014) The not-so-staggering effect of staggered animated transitions on visual tracking. Visualization and Computer Graphics, IEEE Transactions on 20(12):2241–2250

Coe R (2002) It's the effect size, stupid. In: Paper presented at the British Educational Research Association annual conference, vol 12, p 14

Cohen J (1990) Things i have learned (so far). American psychologist 45(12):1304

Cohen J (1994) The earth is round (p<.05). American psychologist 49(12):997

Correll M, Gleicher M (2014) Error bars considered harmful: Exploring alternate encodings for mean and error. Visualization and Computer Graphics, IEEE Transactions on 20(12):2142–2151

Cumming G (2008) Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science 3(4):286–300

Cumming G (2009a) Dance of the p values [video]. tinyurl.com/danceptrial2

Cumming G (2009b) Inference by eye: reading the overlap of independent confidence intervals. Statistics in medicine 28(2):205–220

Cumming G (2012) Understanding the new statistics : effect sizes, confidence intervals, and meta-analysis. Multivariate applications series, Routledge Academic, London

Cumming G (2013) The new statistics: why and how. Psychological science

Cumming G, Finch S (2005) Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist 60(2):170

Cumming G, Williams R (2011) Significant does not equal important: why we need the new statistics [podcast]. ABC Ockham's Razor tinyurl.com/geoffstalk

Cumming G, Fidler F, Vaux DL (2007) Error bars in experimental biology. The Journal of Cell Biology 177(1):7–11

Dawkins R (2011) The tyranny of the discontinuous mind. New Statesman 19:54–57

van Deemter K (2010) Not Exactly: in Praise of Vagueness. Oxford University Press

Dienes Z (2014) Using bayes to get the most out of non-significant results. Frontiers in psychology 5

Downey AB (2013) Think Bayes. O'Reilly Media, Inc.

Dragicevic P (2012) My technique is 20% faster: Problems with reports of speed improvements in hci. Research report, URL `http://hal.inria.fr/hal-00739237`

Dragicevic P, Chevalier F, Huot S (2014) Running an HCI experiment in multiple parallel universes. In: CHI Extended Abstracts, ACM, New York, pp 607–618

Drummond GB, Vowler SL (2011) Show the data, don't conceal them. Advances in physiology education 35(2):130–132

Duckworth WM, Stephenson WR (2003) Resampling methods: Not just for statisticians anymore. In: 2003 Joint Statistical Meetings

Ecklund A (2012) Beeswarm: the bee swarm plot, an alternative to stripchart. R package version 01 5

Eich E (2014) Business not as usual. Psychological Science 25(1):3–6

Fekete JD, Van Wijk JJ, Stasko JT, North C (2008) The value of information visualization. In: Information visualization, Springer, pp 1–18

Fidler F, Cumming G (2005) Teaching confidence intervals: Problems and potential solutions. Proceedings of the 55th International Statistics Institute Session

Fidler F, Loftus GR (2009) Why figures with error bars should replace p values. Zeitschrift für Psychologie/Journal of Psychology 217(1):27–37

Fidler F, et al (2010) The american psychological association publication manual sixth edition: Implications for statistics education. Data and context in statistics education: Towards an evidence based society

Fisher R (1955) Statistical methods and scientific induction. Journal of the Royal Statistical Society Series B (Methodological) pp 69–78

Forum C (2015) Is there a minimum sample size required for the t-test to be valid? Online, URL `http://tinyurl.com/minsample`, last accessed Feb 2015

Franz VH, Loftus GR (2012) Standard errors and confidence intervals in within-subjects designs: Generalizing loftus and masson (1994) and avoiding the biases of alternative accounts. Psychonomic bulletin & review 19(3):395–404

Frick RW (1998) Interpreting statistical testing: Process and propensity, not population and random sampling. Behavior Research Methods, Instruments, & Computers 30(3):527–535

Gardner MJ, Altman DG (1986) Confidence intervals rather than p values: estimation rather than hypothesis testing. BMJ 292(6522):746–750

Gelman A (2013a) Commentary: p-values and statistical practice. Epidemiology 24(1):69–72

Gelman A (2013b) Interrogating p-values. Journal of Mathematical Psychology 57(5):188–189

Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press

Gelman A, Stern H (2006) The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician 60(4):328–331

Gigerenzer G (2004) Mindless statistics. The Journal of Socio-Economics 33(5):587–606

Giner-Sorolla R (2012) Science or art? how aesthetic standards grease the way through the publication bottleneck but undermine science. Perspectives on Psychological Science 7(6):562–571

Gliner JA, Leech NL, Morgan GA (2002) Problems with null hypothesis significance testing (NHST): what do the textbooks say? The Journal of Experimental Education 71(1):83–92

Goldacre B (2012) What doctors don't know about the drugs they prescribe [ted talk]. tinyurl.com/goldacre-ted

Goodman SN (1999) Toward evidence-based medical statistics. 1: The p value fallacy. Annals of internal medicine 130(12):995–1004

Hager W (2002) The examination of psychological hypotheses by planned contrasts referring to two-factor interactions in fixed-effects ANOVA. Method Psychol Res Online 7:49–77

Haller H, Krauss S (2002) Misinterpretations of significance: A problem students share with their teachers. Methods of Psychological Research 7(1):1–20

Hofmann H, Follett L, Majumder M, Cook D (2012) Graphical tests for power comparison of competing designs. Visualization and Computer Graphics, IEEE Transactions on 18(12):2441–2448

Hornbæk K, Sander SS, Bargas-Avila JA, Grue Simonsen J (2014) Is once enough?: on the extent and content of replications in human-computer interaction. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems, ACM, pp 3523–3532

Ioannidis JP (2005) Why most published research findings are false. PLoS medicine 2(8):e124

Jansen Y (2014) Physical and tangible information visualization. PhD thesis, Université Paris Sud-Paris XI

Kaptein M, Robertson J (2012) Rethinking statistical analysis methods for CHI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 1105–1114

Keene ON (1995) The log transformation is special. Statistics in medicine 14(8):811–819

Kerr NL (1998) HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review 2(3):196–217

Kindlmann G, Scheidegger C (2014) An algebraic process for visualization design. Visualization and Computer Graphics, IEEE Transactions on 20(12):2181–2190

Kirby KN, Gerlanc D (2013) BootES: An R package for bootstrap confidence intervals on effect sizes. Behavior research methods 45(4):905–927

Kirk RE (2001) Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement 61(2):213–218

Kline RB (2004) What's wrong with statistical tests–and where we go from here. American Psychological Association

Lai J, Kalinowski P, Fidler F, Cumming G (2010) Dichotomous thinking: A problem beyond NHST. Data and context in statistics education: Towards an evidence based society

Lakens D, Pigliucci M, Galef J (2014) Daniel lakens on p-hacking and other problems in psychology research [podcast]. Rationally Speaking Podcast tinyurl.com/lakens-podcast

Lazic SE (2010) The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC neuroscience 11(1):5

Lee MD (2014) The "new statistics" are built on fundamentally flawed foundations. (Online)

Levine TR, Weber R, Hullett C, Park HS, Lindsey LLM (2008a) A critical assessment of null hypothesis significance testing in quantitative communication research. Human Communication Research 34(2):171–187

Levine TR, Weber R, Park HS, Hullett CR (2008b) A communication researchers' guide to null hypothesis significance testing and alternatives. Human Communication Research 34(2):188–209

MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. Psychological methods 7(1):19

Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: A theory of self-concept maintenance. Journal of marketing research 45(6):633–644

Meehl PE (1967) Theory-testing in psychology and physics: A methodological paradox. Philosophy of science pp 103–115

Miller J (1991) Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. The quarterly journal of experimental psychology 43(4):907–912

Newcombe RG (1998a) Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in medicine 17(8):873–890

Newcombe RG (1998b) Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in medicine 17(8):857–872

Newman GE, Scholl BJ (2012) Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. Psychonomic bulletin & review 19(4):601–607

Norman DA (2002) The Design of Everyday Things. Basic Books, Inc., New York, NY, USA

Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. Advances in health sciences education 15(5):625–632

Nuzzo R (2014) Scientific method: statistical errors. Nature 506(7487):150–152

Osborne JW, Overbay A (2004) The power of outliers (and why researchers should always check for them). Practical assessment, research & evaluation 9(6):1–12

Perin C, Dragicevic P, Fekete JD (2014) Revisiting Bertin matrices: New interactions for crafting tabular visualizations. Visualization and Computer Graphics, IEEE Transactions on 20(12):2082–2091

Perneger TV (1998) What's wrong with Bonferroni adjustments. Bmj 316(7139):1236–1238

Pollard P, Richardson J (1987) On the probability of making Type I errors. Psychological bulletin 102(1):159

Rawls RL (1998) Breaking up is hard to do. Chemical & engineering news 76(25):29–34

Reips UD, Funke F (2008) Interval-level measurement with visual analogue scales in internet-based research: Vas generator. Behavior Research Methods 40(3):699–704

Rensink RA (2014) On the prospects for a science of visualization. In: Handbook of Human Centric Visualization, Springer, pp 147–175

Ricketts C, Berry J (1994) Teaching statistics through resampling. Teaching Statistics 16(2):41–44

Rosenthal R (2009) Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books. Oxford University Press Oxford

Rosenthal R, Fode KL (1963) The effect of experimenter bias on the performance of the albino rat. Behavioral Science 8(3):183–189

Rosnow RL, Rosenthal R (1989) Statistical procedures and the justification of knowledge in psychological science. American Psychologist 44(10):1276

Rossi JS (1990) Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical psychology 58(5):646

Sauro J, Lewis JR (2010) Average task times in usability tests: what to report? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 2347–2350

Schmidt FL, Hunter J (1997) Eight common but false objections to the discontinuation of significance testing in the analysis of research data. What if there were no significance tests pp 37–64

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science 22(11):1359–1366

Smith RA, Levine TR, Lachlan KA, Fediuk TA (2002) The high cost of complexity in experimental design and data analysis: type I and type II error rates in multiway ANOVA. Human Communication Research 28(4):515–530

Stewart-Oaten A (1995) Rules and judgments in statistics: three examples. Ecology pp 2001–2009

The Economist (2013) Unreliable research: Trouble at the lab. Online, URL http://tinyurl.com/trouble-lab, last accessed Mar 2015

Thompson B (1998) Statistical significance and effect size reporting: Portrait of a possible future. Research in the Schools 5(2):33–38

Thompson B (1999) Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. Theory & Psychology 9(2):191–196

Trafimow D, Marks M (2015) Editorial. Basic and Applied Social Psychology 37(1):1–2

Tryon WW (2001) Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. Psychological methods 6(4):371

Tukey JW (1980) We need both exploratory and confirmatory. The American Statistician 34(1):23–25

Ulrich R, Miller J (1994) Effects of truncation on reaction time analysis. Journal of Experimental Psychology: General 123(1):34

Velleman PF, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. The American Statistician 47(1):65–72

Vicente KJ, Torenvliet GL (2000) The Earth is spherical (p < 0.05): alternative methods of statistical inference. Theoretical Issues in Ergonomics Science 1(3):248–271

Victor B (2011) Explorable explanations. Online, URL `http://worrydream.com/ExplorableExplanations/`, last accessed Mar 2015

Wainer H (1984) How to display data badly. The American Statistician 38(2):137–147

Wickham H, Stryjewski L (2011) 40 years of boxplots. Am Statistician

Wierdsma A (2013) What is wrong with tests of normality? Online, URL `http://tinyurl.com/normality-wrong`, last accessed Nov 2014

Wilcox RR (1998) How many discoveries have been lost by ignoring modern statistical methods? American Psychologist 53(3):300

Wilkinson L (1999) Statistical methods in psychology journals: Guidelines and explanations. American psychologist 54(8):594

Willett W, Jenny B, Isenberg T, Dragicevic P (2015) Lightweight relief shearing for enhanced terrain perception on interactive maps. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, CHI '15, pp 3563–3572

Wilson W (1962) A note on the inconsistency inherent in the necessity to perform multiple comparisons. Psychological Bulletin 59(4):296

Wood M (2004) Statistical inference using bootstrap confidence intervals. Significance 1(4):180–182

Wood M (2005) Bootstrapped confidence intervals as an approach to statistical inference. Organizational Research Methods 8(4):454–470

Zacks J, Tversky B (1999) Bars and lines: A study of graphic communication. Memory & Cognition 27(6):1073–1079

Ziliak ST, McCloskey DN (2008) The cult of statistical significance. Ann Arbor: University of Michigan Press 27

Contents