Fair Statistical Communication in HCI

Pierre Dragicevic

Fair Statistical Communication in HCI

Pierre Dragicevic

Author version v.1.6.4.
The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-26633-6.
This version differs by its page layout.

Abstract Statistics are tools to help end users accomplish their task. In research, to be qualified as usable, statistical tools should help researchers advance scientific knowledge by supporting and promoting the effective communication of research findings. Yet areas such as human-computer interaction (HCI) have adopted tools — i.e., *p*-values and dichotomous testing procedures — that have proven to be poor at supporting these tasks. The abusive use of these procedures has been severely criticized in a range of disciplines for several decades, suggesting that tools should be blamed, not end users. This chapter explains in a non-technical manner why it would be beneficial for HCI to switch to an *estimation* approach, i.e., reporting informative charts with effect sizes and interval estimates, 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 tests or *p*-values.

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 not able to guarantee either clarity, accuracy, or verifiability in scientific communication.

Pierre Dragicevic
Inria, France, e-mail: pierre.dragicevic@inria.fr

A distinguishing feature of these statistical procedures is their mechanical nature: data is fed to a machine called "statistics", and a binary answer is produced: either we can trust the data or not. The idea is that for the work to qualify as scientific, inference from data should be as objective as possible and human judgment should be put aside. Few HCI researchers see the contradiction between this idea and the values they have been promoting — in particular, the notion that "humans in the loop" are often 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 a substitute 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 (see Chapter 3), 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 chapter challenges the view that human judgment can be left out when doing confirmatory analysis.

By mechanical decision procedures I refer to a family of statistical procedures termed *null hypothesis significance testing (NHST)*. This chapter compares NHST with *interval estimation of effect sizes* (or *estimation* for short), an alternative approach that consists of reporting effect sizes with interval estimates and offering nuanced interpretations (Cumming, 2013). The chapter skips many technical details, widely available elsewhere. The key difference between the two approaches lies in their usability, and it can be summarized by the following illustration:

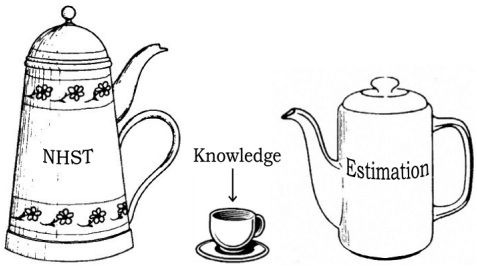

Fig. 1 If empirical knowledge was coffee and articles were coffee cups, experiments would be coffee machines and statistical tools would be coffee pots. Drawing inspired from Norman (2002).

NHST as it is typically carried out involves *i)* computing quantities called *p*-values and then *ii)* applying a cut-off to these *p*-values to determine "statistical significance". Section 2 focuses on the notion of *p*-value divorced from the notion of a cut-off. Confidence intervals, a particular type of interval estimate closely related to *p*-values, will be used as a baseline of comparison. Section 3 discusses the use of cut-offs to determine statistical significance and contrasts this approach with estimation. Section 4 offers practical advice on how to achieve fair (i.e., clear and truthful) statistical communication through estimation. Readers seeking practical advice can jump to Section 4, while those seeking for justifications can keep on reading.

2 p-values, Effect Sizes and Confidence Intervals

Though the aim of this chapter is not to offer an introduction to statistics, it is useful to start by briefly reviewing a few basic concepts. This will make sure we understand the examples offered throughout this chapter, and will also clarify our assumptions.

2.1 A Minimalistic Example and Quick Reminders

Imagine you need to help your best friend decide whether or not she should buy an expensive pill for losing weight, and you find a scientific paper assessing the pill's efficacy. From the statistical report you gather the following information:

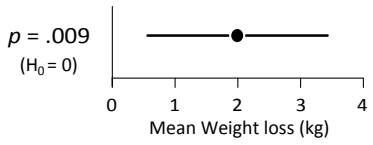

Fig. 2 Results of an imaginary study on the effectiveness of a weight-loss pill.

What can you conclude from this figure? How confident can you be? Where does the uncertainty come from, exactly?

Sample and population. There is uncertainty as to the true efficacy of the pill, partly because the pill has only been tested on a few volunteers. Ideally, these volunteers constitute a random *sample* from a *population* of interest (e.g., all overweight US citizens), to which we assume your friend belongs. The mean weight loss only informs us about the *sample*, but a much better measure would be the weight loss averaged across the entire *population*. Neither measure will tell for sure what will happen to your friend, but the population average would be a much better indication.

Statistical inference. Since the population average is a better measure of efficacy, we decide it is really our measure of interest, even though it can only be guessed. This guessing process about a hypothetical population is essentially what is meant by statistical "inference". It is only part of what we can do with statistics but it is central in HCI and other domains, and this is what this chapter focuses on. Note that other interpretations of statistical inference exist that are perhaps more accurate and realistic for HCI experiments (Frick, 1998). However, random sampling is by far the most widespread and we will stick to it for simplicity.

Replication. In statistics, a replication typically refers to a hypothetical sample that could have been obtained from the same (also hypothetical) population on a different experiment. For example, had the researcher above chosen a different set of randomly selected volunteers, the results would have been different. A concrete illustration of multiple replications will be provided in Section 2.4.4.

Point estimate. The black dot on the chart is the *point estimate* of the population-wise weight loss: it is our *best bet* on how much weight your friend will lose, in the

absence of any other information. A simple method for computing the point estimate is to take the sample mean (for other methods see Section 4.3). Different replications would yield a different point estimate, hence the uncertainty.

Interval estimate. The bar on the figure is an *interval estimate*. It indicates the uncertainty around the point estimate. The most common type of interval estimate is the confidence interval. Let us assume that a 95% confidence interval has been provided, as it is commonly the case. Strictly speaking, a 95% confidence interval is an interval that is obtained from a procedure that satisfies a certain property, this property being that the intervals it generates capture the population mean 95% of the time across many replications. In practice, it is simpler to think of a confidence interval as a *range of plausible values* for the population mean (Cumming, 2012; Fidler and Loftus, 2009; Schmidt and Hunter, 1997). The point estimate is the *most plausible*, and plausibility *smoothly decreases* as we move away from it — in typical cases, the point estimate is about *seven times more plausible* than the confidence limits, i.e., the interval's upper and lower ends (Cumming, 2013, p.17). Plausibility does not suddenly drop when crossing the limits, as values outside are implausible but *not impossible*. This interpretation is an approximation and there is debate over whether it is a good one (more on this in Section 3.1), but for now let us trust it.

***p*-value**. The number to the left is the *p*-value for a *null hypothesis of no effect*. This null hypothesis is the devil's advocate claim that the pill yields exactly zero weight loss on average across the entire population. If this was true, any result would be caused by sampling error alone. However, not all results would be equally likely: a consistent and massive weight loss, for example, would be quite unlikely under the null hypothesis. This is what the *p*-value captures: it is the probability of observing results as extreme as (or more extreme than) what was actually observed if the null hypothesis was true. In practice, it is easier to think of a *p*-value as *a measure of strength of evidence against the null hypothesis*; The closer p is to 0, the more evidence that the pill has a some effect overall, or more specifically has a strictly positive effect, since here the point estimate is positive. This is how R.A. Fisher, who introduced *p*-values, thought we should understand them (Goodman, 1999, p.997; Gigerenzer, 2004, p.593). A different view will be given in Section 3.1.

There is no way of interpreting Figure 2 that would satisfy all statisticians and methodologists, but a reasonable interpretation is that we can trust the pill to be effective ($p = .009$ is low), and that on average the weight loss is most likely between 0.5kg to 3.5kg, maybe not too far from 2kg. Now let us see how useful p really is.

2.2 Choosing a Pill

Your friend has now decided to buy a pill to lose weight, but there are many options and she cannot make up her mind (this problem is inspired from Ziliak and McCloskey (2008, p.2303)). 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, shown on the next page.

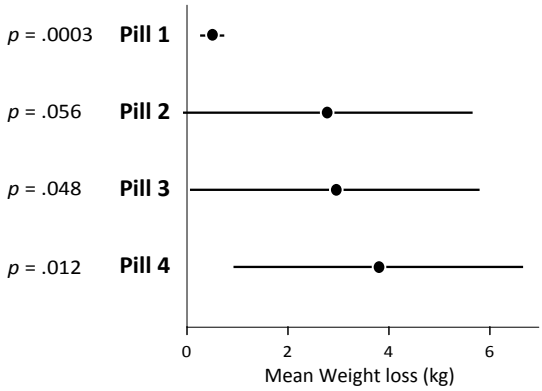

Fig. 3 Chart showing the results from four (imaginary) studies on the effectiveness of different weight-loss pills. Error bars are 95% confidence intervals and *p*-values assume a null hypothesis of no effect.

Note that you *only have access to the four study reports*. So even if you can do statistics and would like to compute the *p*-values for all pairwise differences between pills, you cannot. This scenario is meant to illustrate to what extent *already published studies* can be used to inform decisions, depending on how results are *reported*. A researcher who writes a literature survey does not usually download all datasets to re-run analyses. Also, it does not matter whether the *p*-values are used to assess conditions individually (as it is the case here) or differences between conditions (as is more often the case in HCI). It may help to think of the reported weight losses as differences between pills and a common baseline, e.g., a placebo.

Any trained scientist will have immediately noticed the enormous amount of uncertainty in the data — except apparently for the first pill[1] — and should not feel compelled to draw any conclusion. 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. This is indeed a sensible answer: it is reasonable to favor a pill that yields the maximum expected utility — here, weight loss. Recall that each point estimate shows your friend's most likely weight loss with that pill. For your friend, pill 4 is the best bet, and it is certainly a much better bet than pill 1.

Now suppose that pill 4 does not exist. Which pill would you pick among the remaining ones? Look at Figure 3 carefully. With some hesitation, most people reply pill 3. Now also remove pill 3. More hesitation ensues: some people choose pill 2 while others choose pill 1. 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.

[1] The width of confidence intervals generally increases with the variability of observations and decreases (somehow slowly) with sample size (Cumming, 2012). So either pill 1 has a much more consistent effect or the number of subjects was remarkably larger. It is not very important here.

2.3 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. 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 always 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:

Table 1 The p-value for each pill

Pill 1	$p = .0003$
Pill 2	$p = .056$
Pill 3	$p = .048$
Pill 4	$p = .012$

Ranking the pills based on this information only 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 ($p = .012$), 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, at the very least, account for the point estimates in Figure 3, i.e., our best bets.

2.3.1 The Importance of Conveying Effect Sizes

Broadly speaking, an effect size is *anything that might be of interest*[2] (Cumming, 2012, p.34). An effect size can be, e.g., the average completion time difference between two techniques. In our case, effect sizes are simply average weight losses.

p-values capture what is traditionally termed *statistical significance*, while effect sizes capture *practical significance* (Kirk, 2001). For example, the effect of pill 1 can be said to exhibit a high statistical significance, but only a moderate practical significance compared to others.

Practical significance is our primary focus, both in research and in real-world decision making. Thus it is widely recognized that effect sizes should be reported (APA, 2010). What methodologists generally mean by this is that we should report all *point estimates* of interest, or equivalently — assuming we are interested in simple effect sizes — all *sample means* of interest. But since a point estimate only conveys our best guess about the population, it is crucial to also convey information on uncertainty. The next page shows two ways of doing this.

[2] The term *effect size* is often used in a narrower sense to refer to *standardized effect sizes* (Coe, 2002, see also Chapter 5). Although sometimes useful, reporting standardized effect sizes is not always necessary nor is it always recommended (Baguley, 2009; Wilkinson, 1999, p.599).

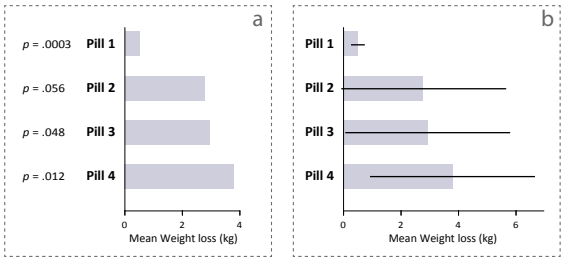

Fig. 4 Showing the most plausible effect sizes and their associated uncertainty using a) *p*-values with point estimates of effect sizes (here shown as bar charts); b) 95% CIs around point estimates.

Above, each black dot has been replaced by a bar, but this is only a matter of presentation (see Figure 9 in Section 4). The option *a* (left) follows the orthodoxy and the common recommendation to report *p*-values together with effect sizes (Thompson, 1998). The option *b* (right) follows an estimation approach that consists of reporting point and interval estimates for effect sizes, without *p*-values (Cumming, 2012). In the simplest cases, *a* and *b* are theoretically equivalent and convey the same information — readers can even learn to mentally convert from *b* to *a* (Cumming, 2012, pp.106–108). However, it seems harder to mentally convert from *a* to *b*, especially when confidence intervals are asymmetrical (e.g., confidence intervals on proportions, correlations, transformed data, or bootstrap confidence intervals). Regardless, the option *b* is clearly easier to read and more informative.

Methodologists who remain attached to *p*-values (APA, 2010; Abelson, 1995; Levine et al, 2008b) suggest reporting everything: *p*-values, point estimates of effect sizes, and their 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 more rigorous to complement a confidence interval with a *p*-value that captures accurately how far it is from zero. Later I offer arguments against this idea, which can be summarized using another illustration:

Fig. 5 Merging a bad design with a good design does not necessarily yield a good design. In statistical communication, reporting everything just in case can produce unnecessary clutter and prompt misinterpretations.

2.3.2 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.

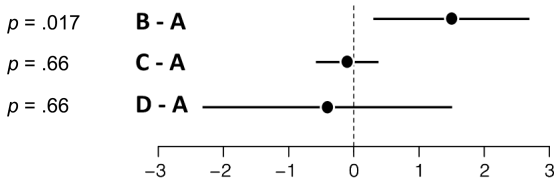

Fig. 6 95% confidence intervals showing differences between conditions.

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 such a high p-value: it tells us that zero is plausible, but 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 would not 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, 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 reveal the uncertainty in the experimental data. This is crucial information.

2.4 Usability Problems with p-values

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

2.4.1 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 an average, and they are used by the media, for example when reporting opinion polls (Cumming and Williams, 2011). Another important difference is that confidence intervals have natural visual representations while *p*-values do not.

One issue specific to confidence intervals is their lack of standardization. They are visually represented by error bars, which are also used to show several other types of information, including standard errors (typically about half the size of 95% CIs) and standard deviations. Researchers simply need to become more consistent and get used to clearly indicating what error bars refer to (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 incorrect 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. Because confidence intervals are 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 damaging.

2.4.2 Misinterpretations Regarding Probabilities

Again, 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 think of 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).

As will be discussed in Section 3.1.2, stating that a particular 95% confidence interval has a 0.95 probability of capturing the population mean is also generally incorrect. However, 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 (Fidler and Loftus, 2009). Shown visually, confidence intervals look less rigorous, and do not prompt overconfidence 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 (see Chapters 8 and 9). In particular, tools exist for computing confidence intervals that convey probabilities, as will be further discussed in Section 3.1.

2.4.3 Misinterpretation of High p-values

Although strictly speaking, *p*-values do not capture any practically meaningful probability, we can use them, like Fisher, as 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 6), 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 Figure 6 eliminates the problem.

2.4.4 Misinterpretations Regarding Reliability

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

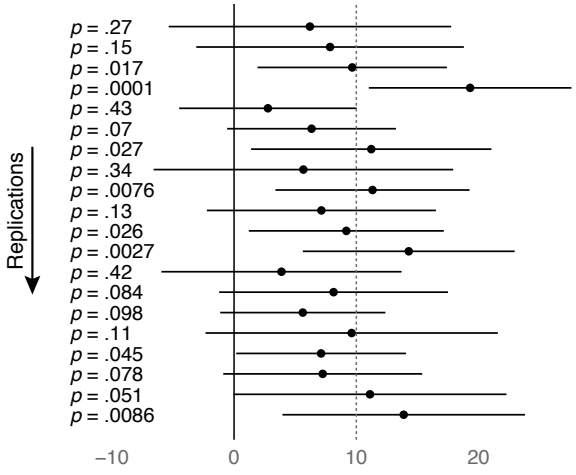

Fig. 7 *p*-values and 95% confidence intervals obtained by simulating replications of an experiment (normally distributed population with $\mu = 10$ and $\sigma = 20$; $n = 20$; statistical power 0.56). After Cumming (2009a).

Running an experiment amounts to closing your eyes and picking one of the *p*-values (and confidence interval) in this figure. With a statistical power[3] of about 0.5 (typical in both 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). Suppose an experiment yields $p = .05$ for a *t*-test. If the experiment is repeated with different participants, there is a 20% chance that the new *p*-value will fall *outside* the interval (.00008, .44). Even if the initial experiment yielded an impressive $p = .001$, there is still a 20% chance that the new *p*-value falls outside the interval (.000006, .22). *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 obtain a *p*-value from 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 a function of the mean and the standard deviation (and sample size, if not held constant). Thus the *p*-value obtained would be 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 7). By definition (see Section 2.1), only 95% of these will capture the population mean in the long run. Being fully aware of the dance of confidence intervals is certainly an important prerequisite for their correct use and interpretation. Watching replication simulations (e.g., from Cumming (2009a)) is enough to get a good intuition, and one can hardly claim to understand 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 likely perceptual and cognitive differences: confidence intervals, especially shown graphically, may not give the same illusion of certainty and truth as *p*-values reported with high numerical precision.

Here is the R code for Figure 7 to help you play with your own simulations:

```
require(ggplot2)
require(plyr)

replications <- 20
sampleSize <- 20
populationMean <- 10
populationSd <- 20
plotRange <- c(-15, 35)
```

[3] Briefly, statistical power is the probability of correctly detecting an effect whose magnitude has been postulated in advance. The more participants, the larger the effect size and the lower the variability, the higher the statistical power (see also Chapter 5).

```
createReplication <- function(replication) {
  #set.seed(replication) # uncomment this to get the same results each time
  obs <- rnorm(sampleSize, populationMean, populationSd)
  ttest <- t.test(obs)
  data.frame(mean = mean(obs), ci.lower = ttest[4]$conf.int[1],
             ci.upper = ttest[4]$conf.int[2], pvalue = ttest[3]$p.value)
}

dance <- ldply(1:replications, createReplication)

format_p <- function(p) {
  paste("p =", substring(prettyNum(p, digits=2, scientific=FALSE), 2))
}

ggplot(data = dance, aes(x = 1:replications, y = mean, label=format_p(pvalue))) +
  geom_pointrange(aes(ymin=ci.lower, ymax=ci.upper), size=0.7) +
  geom_text(y=plotRange[1], hjust=0) +
  geom_abline(intercept = 0, slope = 0) +
  geom_abline(intercept = populationMean, slope = 0, lty = 2) +
  ylim(plotRange) + coord_flip() +
  theme_bw() + theme(
    axis.title = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank(),
    panel.border = element_blank(),
    text = element_text(size=17))
```

2.5 Conclusion

There rarely seems to be a good reason to report *p*-values in an HCI research paper, since confidence intervals can present the same information and much more, and in a much clearer, more usable manner. Perhaps the only remaining argument in favor of *p*-values is that they are useful for formally establishing statistical significance. But as we will now see, the notion of binary significance testing is a terrible idea for those who want to achieve fair statistical communication.

3 Null Hypothesis Significance Testing vs. Estimation

We previously mentioned that *statistical significance* can be quantified in a continuous manner 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 a major component of null hypothesis significance testing (NHST).

3.1 A Few More Reminders

To put things in context and further clarify our underlying assumptions, let us recall a few under-discussed but important statistical ideas before proceeding.

3.1.1 Frequentist Statistics, Fisher and Neyman–Pearson

For the sake of simplicity let us equate the null hypothesis to the hypothesis of no effect. Suppose that *1)* many replications of an experiment are carried out; *2)* each time, the researcher concludes that there is an effect *iif* $p < \alpha$; *3)* there is in truth no effect. In the long run, the researcher will be wrong $(100 \times \alpha)\%$ of the time. A known proportion of the time, she will be committing what is called a Type I error.

This way of interpreting *p*-values is termed *frequentist* because it involves long-run frequencies. Originally put forward by Fisher, it was formalized into a rigorous procedure by Neyman and Pearson (Goodman, 1999, p.998; Gigerenzer, 2004, p.590–591). According to this procedure, the researcher sets α before carrying out the experiment, and if $p < \alpha$, the researcher behaves as if there was an effect[4]. If all researchers were to apply this procedure and agree on, say, $\alpha = .05$, then only 5% of all significance tests where there is in truth no effect would yield a Type I error.

Neyman and Pearson insisted that if $p < \alpha$, the researcher should *behave as if* there was an effect, and *nothing else*. The researcher should not only ignore *p*, but also refrain from reasoning or holding any belief (Gigerenzer et al, 1990, pp.98–105). This view of statistics can be characterized as strictly frequentist and *behavioristic*. It seems well suited for automating repeated decisions (e.g., in quality control), but not so much for actual research practice. Fisher, who was an applied researcher, advocated an *epistemic* view of statistics, where the *p*-value brings knowledge about the data. Although he suggested in his earlier writings that conventional α cut-offs can be useful (see Chapter 5), he viewed *p* as a continuous measure of strength of evidence. He rejected the Neyman–Pearson procedure as *"childish"*, *"remote from scientific research"*, and intellectually *"horrifying"* (Gigerenzer, 2004, p.593). In turn, Neyman and Pearson criticized Fisher for lacking rigor and consistency. Although this may be true, fair statistical communication seems deeply incompatible with the Neyman–Pearson view of scientists as brainless decision machines.

Today no researcher uses a strict Neyman–Pearson procedure, since virtually any researcher carries out statistical analyses for *epistemic* reasons: for learning things, drawing conclusions and making arguments. Yet some aspects of the procedure crept into research practice and textbooks. Researchers report and often interpret *p*-values, but they also apply an α cut-off and use it to make dichotomous "decisions" about what we should believe. NHST as it is carried out today consists of this incoherent mix of Fisher and Neyman–Pearson methods (Gigerenzer, 2004).

[4] Strictly speaking, Neyman–Pearson's procedure involved choosing between the null hypothesis and an *alternative hypothesis* generally stating that the effect exists and takes some precise value. Accepting the null if the alternative hypothesis is true is a Type II error. Its frequentist probability is noted β, and power is defined as $1 - \beta$. These notions are not important to the present discussion.

3.1.2 On Interpreting Confidence Intervals

Having covered frequentist statistics, it is now possible to discuss interpretation issues with confidence intervals. Perhaps surprisingly, confidence intervals were first introduced by Neyman. They were designed to be used within his strict frequentist and behavioristic framework: the researcher states that the confidence interval contains the population mean, and nothing else (Morey et al, 2015, p.3). She does not reason or holds beliefs, only *behaves as if* this was true. If the confidence level is 95%, in the long run she will be wrong about 5% of the time.

There is another link between confidence intervals and Neyman–Pearson testing. Confidence intervals can be used to carry out statistical significance tests, since examining whether a $100 \times (1-\alpha)\%$ CI contains the value v is the same as examining whether the p-value for $H_0 = v$ is lower than α. This can be verified in all previous figures for $\alpha = .05$ and $v = 0$. This use of confidence intervals is common practice but since it is essentially the same as NHST, it inherits all of its drawbacks.

Advocates of estimation reject both interpretations of confidence intervals and recommend instead the more nuanced epistemic interpretation offered in Section 2.1 (Schmidt and Hunter, 1997, p.13; Cumming, 2012). This approach focuses on extracting as much useful information as possible from confidence intervals while recognizing that they cannot be fully trusted.

There is a caveat, though. Confidence intervals are defined in a frequentist way (see Section 2.1), and this definition is permissive enough to allow for many different types of confidence interval procedures, including absurd ones. For example, a random procedure that returns the real line ($\mathbb{R}$) 95% of the time and the empty set ($\varnothing$) 5% of the time is a valid 95% confidence interval procedure. This challenges the notion that any given confidence interval will necessary capture the range of plausible values. Other pathologic cases are illustrated by Morey et al (2015).

Bayesian interval estimates, or credible intervals (see Chapters 8 and 9), are the only interval estimates for which the "range of plausible values" interpretation is formally correct (Morey et al, 2015). In addition, they produce more reasonable and more informative interval estimates if there is reliable *a priori* knowledge about the possible range of effect sizes (Gelman, 2013a).

Nevertheless, there are practical reasons to use confidence intervals. In many common situations, confidence intervals agree with so-called *objective* credible intervals (Greenland and Poole, 2013). This is true for exact confidence intervals (Bayarri and Berger, 2004, p.63) and bootstrap confidence intervals (Bååth, 2015). In addition, confidence intervals are easier to compute than credible intervals, they are more widely used, and they are currently better supported by statistical tools. In the context of this chapter, their mathematical equivalence with statistical significance testing also allows us to clearly contrast estimation thinking with dichotomous testing. Confidence intervals can be seen as the "poor man's" credible intervals, and as a good bridge between mindless NHST and sophisticated Bayesian reasoning.

With these issues in mind, let us now compare significance testing as it is carried out today (i.e., using α as an epistemic tool) with estimation as it is done today (i.e., using confidence intervals as approximations to objective credible intervals).

3.2 How Useful is the α cut-off?

The insights yielded by the use of an α cut-off can be assessed by returning to our first scenario and considering again the respective merits of our four pills:

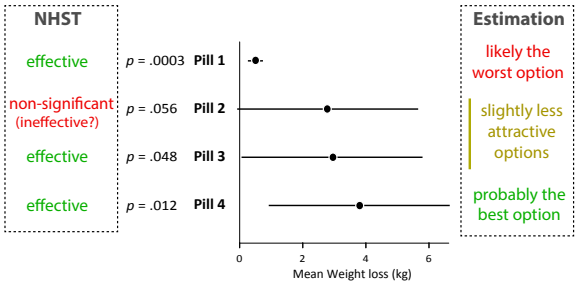

Fig. 8 The same four pills, ranked based on the outcome of statistical tests (left), and based on an examination of effect sizes and interval estimates (right).

As we saw previously, a sensible ranking scheme (shown to the right) would give a preference to pill 4, then pills 2–3 (whose results are almost identical), then pill 1. Nothing is certain and we may well be wrong, especially about pills 2, 3, and 4 for which the data is very unreliable. But since we need to decide we are forced to rank. In a scientific paper one would typically be much 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 methods: 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.

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 interval estimates). It is only logical to assume that statistical significance in its binary form cannot carry more.

3.3 More Usability Problems Brought by the α Cut-Off

Binary significance testing is based on p-values and therefore inherits their usability problems. The use of a binary decision rule based on a cut-off also introduces a range of additional usability problems that are discussed next.

[5] The sharp distinction between pills 2 and 3 is not a caricature. Due to Neyman–Pearson's heritage, even pointing out that a non-significant p-value is close to .05 is often considered a serious fault.

3.3.1 Misjudgments of Uncertainty

p-values give a seductive 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; Hoekstra et al, 2006). For example, this amounts to saying that in Figure 4a, 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 4b.

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 2.4.4). Thus many HCI experiments effectively amount to tossing a coin (Dragicevic et al, 2014).

3.3.2 Misinterpretations Regarding Comparisons

Few researchers are fully aware of the disturbing paradoxes yielded by the use of a cut-off when comparing findings. The results for pills 2 and 3, for example, appear very different despite being virtually identical (Figure 8). In fact, pill 2 has close to a 50% chance of ending up better than pill 3 on the next replication (remember the dance in Figure 7). 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 for 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' minds, but in the design of NHST tools.

With estimation, results are labeled with confidence intervals, whose comparison is not always trivial (Cumming, 2012, Chap. 6) but is certainly much less problematic. For example, instead of simply writing *"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 write *"our mean improvement of 1.9 s, 95% CI [-0.7, 4.4] is consistent with the improvement of 3.1 s, 95% CI [1.7, 4.7] reported by Schmidt (2010) but seemingly lower than the improvement of 5.2 s, 95% CI [4.1, 6.6] reported by John (2012)"*.

3.3.3 Misinterpretations Regarding Type I Error Rates

Due to sampling error, any statistical analysis is error-prone. The idea that a researcher can take control over the likelihood of making false discoveries is very appealing, and so is the idea that among all published results a known proportion will be wrong. But Neyman–Pearson's Type I error rate captures neither of these, even remotely (Pollard and Richardson, 1987; Colquhoun, 2014). Like p (see Section 2.4.2), the Type I error rate is computed with the assumption that the null hypothesis is true. In many disciplines a Type I error is impossible, and one can only fail to detect the effect[6], or commit sign errors and magnitude errors (Gelman, 2004). In addition, the fact that not all results are published renders theoretical error rates mostly irrelevant for assessing research reliability (also see Section 3.3.5). The Type I error rate is only a theoretical convenience that captures an idealized situation. It is a useful and powerful thinking tool, but the current obsession with Type I error rates and insistence on maintaining them at a precise 5% level sound more like a magical ritual than something that will necessarily guarantee reliable research.

3.3.4 Multiple Levels of Significance

A practice that has become less popular in HCI (although it is still sometimes 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 Neyman–Pearson's strict frequentist approach and Fisher's approach of using exact p-values as a measure of strength of evidence (Gigerenzer, 2004). Few, if any, statistical methodologists recommend the use of multiple levels of significance.

3.3.5 Issues Regarding Publication Bias

Since statistical significance is a major criterion for publishing study papers, conference proceedings and journals 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 are never published and those that are published because of statistical luck or flawed analyses are never 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).

[6] Since computing β (or the probability of a Type II error) requires assigning a precise value to the population mean, β is also very unlikely to correspond to an actual probability or error rate.

3.3.6 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 are significant (Abelson, 1995, p.55). Such practices 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; Gelman and Loken, 2013). They 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. Information obfuscation can also occur after *p*-values have been computed, e.g., by selectively reporting results (*cherry picking*), using post-hoc α cut-offs (Gigerenzer, 2004), or elaborating evasive narratives (Abelson, 1995, p.55). NHST makes it easy to dissimulate unscientific practices under the appearance of objectivity and rigor. Since humans excel at unconsciously taking advantage of fuzzy lines between honest and dishonest behavior (Mazar et al, 2008), merely promoting scientific integrity is likely futile. To be usable, statistical tools should be designed so that they do not leave too much space for questionable practices and self-deception. Estimation approaches do not draw a sharp line between interesting and uninteresting results, and thus make "torturing" data much less useful. As I will discuss later, planned analyses are another very effective safeguard.

3.3.7 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). However, many dichotomies are 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[7].

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).

[7] For elements of discussion concerning this particular dichotomy, see Stewart-Oaten (1995); Norman (2010); Velleman and Wilkinson (1993); Wierdsma (2013); Abelson (1995, Chap. 1) and Gigerenzer (2004, pp.587–588).

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 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 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 seems much more likely to promote clear statistical thinking.

3.3.8 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, and statistical hypotheses are experiment-specific statements derived from research hypotheses in order to assess the plausibility of the 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; Gelman and Loken, 2013).

Many research hypotheses are dichotomous: the Higgs boson either exists or not; the acceleration of a falling object is either a function of its mass or it is not; a pointing method either obeys Fitts' Law or some other (say, Schmidt's) law. Such dichotomies are justified: although there is the possibility that a pointing method 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 (see Chapters 8 and 9). That said, in many cases choosing a hypothesis is a decision that is informed both by data and by extraneous considerations, so estimation methods (e.g., for goodness of fit) can still be beneficial in this context.

Regardless, the vast majority of HCI studies are *not* conducted to test 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 presents a mere hunch (or hope) as something it is not, a scientific theory that needs to be tested. It is sufficient to simply ask a question. Since the respective merits of two techniques cannot be meaningfully classified into sharp categories, it is preferable to ask questions in a quantitative manner, and use estimation to answer them.

3.3.9 End User Dissatisfaction

NHST has been severely criticized for more than 50 years by end users to whom fair 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" in the 1920s — came to reject 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; Open Science Collaboration, 2015).

3.4 Conclusion

Null hypothesis significance testing rests on important theoretical ideas that can help reflect on difficult notions in statistics, such as statistical power and multiple comparisons (briefly covered in the next Section). However, it is now widely understood that it is not a good tool for scientific investigation. I — as many others before — have pointed out a range of usability problems with NHST procedures. 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., Loftus, 1993; Schmidt and Hunter, 1997; Lambdin, 2012; Cumming, 2013) with weak reformists who think it should be *i)* de-emphasized and *ii)* properly taught and used (e.g., Abelson, 1995; Abelson, 1997; Levine et al, 2008a; Levine et al, 2008b). I have already given arguments against *i)* by explaining that *p*-values are redundant with confidence intervals (Section 2). Concerning *ii)*, I suggested that the problem lies in the tools' 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 Fair Statistical Communication Through Estimation

What do we do now? There are many ways to analyze data without using NHST or *p*-values. Two frequently advocated alternatives are estimation and Bayesian methods, although the two address different issues and can be combined. As we mentioned in Section 3.1.2, there is a Bayesian version of estimation, and much of the justification and discussion of interpretation of CIs can be transferred to these methods. Again, we focus here on estimation with confidence intervals because it is simple and accessible to a wide audience of investigators and readers (thus it emphasizes simplicity as discussed next). Keep in mind, however, that some Bayesians strongly reject any kind of frequentist tool, including confidence intervals for the reasons outlined in Section 3.1.2 (Trafimow and Marks, 2015; Morey et al, 2015).

Confidence intervals have been studied extensively, and statistical packages like *R* offer extensive libraries for computing them. However, there is a lack of pedagogical material that brings all of these 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. Since this is a vast topic, in this section we only discuss a few general 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. Most research tasks require expertise, judgment, and creativity. The analysis and communication of empirical findings is no exception. This task is necessarily subjective, but it is our job as scientists to carry it out (Thompson, 1999; Lambdin, 2012).

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 fair statistical communication. From this general principle one can derive 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 free minds who can judge it. Thus a study report should be an exercise of pedagogy as much as an exercise of rhetoric.

Transparency. All decisions made when carrying out an analysis should be communicated as explicitly as possible, because the results of an analysis cannot be fairly assessed if many decisions remain concealed (see *p*-hacking in Section 3.3.6).

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[8]. This is a corollary of the principle of clarity, as any analysis that departs from this principle is misleading about the data.

Noncontingency. Ideally, no decision subtending an analysis should be contigent 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 the principles of clarity, transparency and simplicity, because data-contingent procedures are hard to explain and easy to leave unexplained (Gelman and Loken, 2013). 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 high statistical *precision* (Cumming, 2012, Chap. 13), should also be a goal to pursue.

4.2 Before Analyzing Data

Experiment design and statistical analysis are tightly coupled (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 fair 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 in purely exploratory studies and when building multivariate models, complex experiment designs — i.e., many factors or many conditions per factor — are best 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 one addresses the questions and issues raised by the previous one.

[8] 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.

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 precision (Rawls, 1998). For measurements such as task errors or age, prefer continuous metrics to binary or categorical scales[9].

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 uncontingency 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 3.3.6). 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 "Unplanned Analyses" subsection.

Tip 6: There is no magic number of participants. The idea that there is a "right" number of participants in HCI is part of the folklore and has no theoretical basis. One issue is statistical precision, and it will be discussed next. A separate issue is meeting statistical assumptions. Concerning statistical assumptions, about twenty participants put the researcher in a safe zone for analyzing about any numerical data (see Tips 13 and 14). Analyses do not suddenly become invalid below that — just possibly less accurate. If all scales are believed 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; Norman, 2010, p.628).

Tip 7: Anticipate precision. It is important to achieve high statistical precision, i.e., narrow confidence intervals (Cumming, 2012, Chap. 13). Therefore, 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 at a safe distance from zero is a good rule of thumb. The $p < .05$ criterion has so much psychological influence on reviewers that it is not unreasonable to try to meet it. However, it is better to forget about it in the analysis stage.

---

[9] 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. Likert scales are easy to analyze if they are constructed adequately, i.e., by averaging responses from multiple question items (see Carifio and Perla (2007)). If responses to individual items are of interest, it can be 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 8: Hypotheses are optional. Hypotheses have their place, especially 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 can benefit transparency (Rosenthal and Fode, 1963; Rosenthal, 2009), but expectations do not need to be called hypotheses. Expectations can also 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; Gelman and Loken, 2013).

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 existing material. Some of them are at odds with orthodox practices as popularized by textbooks, but they are in better accordance with fair statistical communication and are supported by compelling arguments from the methodology literature. I have also tried to include common pitfalls that I have committed or observed while working with students.

Tip 9: As many observations as participants. Perhaps the only serious mistake that can be made when computing confidence intervals is by not aggregating data. Suppose you recruit 20 subjects, show them various conditions (e.g., technique $\times$ task type), and for each condition you ask them to perform 10 similar tasks. Multiple measurements can greatly help reduce statistical noise, but reporting confidence intervals based on all measurements (n=200) would be wrong (Lazic, 2010). This is because the purpose of statistical inference in HCI is typically to generalize data to a population of people (see Section 2.1), not of trials[10]. Measurements need to be aggregated (e.g., averaged) so that each participant ends up with a *single observation* before any confidence interval is computed. NHST has developed notations that make it possible for 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 observations were obtained. Raw measurements can be converted into different units 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 (Velle-

[10] Both types of inferences can be combined using hierarchical or multi-level models, and tools exist for computing hierarchical confidence intervals (see Chapter 11).

man 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 value (*truncation*, see Ulrich and Miller (1994)) or based on spread (*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 the mean outperforms all other measures (Wilcox, 1998). In general there is 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 skip such procedures.

Tip 12: Consider the log transform. The log transform corrects for positive skewness in time measurements and gives less weight to extreme observations, 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 prevent the embarrassing situation where a confidence interval extends to negative values. The procedure consists in log-transforming all raw time measurements, performing all analyses as usual, then converting back (*antilogging*) the means and confidence 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 will become ratios (Gardner and Altman, 1986, p.750). As it turns out, ratios between completion times are easier to interpret than differences because they are unitless (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. Bootstrapping is a very useful method that has not received enough attention (Kirby and Gerlanc, 2013; Wood, 2004, 2005). Briefly, it consists of generating many alternative datasets from the experimental data by randomly drawing observations with replacement. The variability across these datasets is assumed to approximate sampling error and is used to compute so-called *bootstrap confidence intervals*. This way of calculating confidence intervals is recent in the history of statistics because it requires computers, but it is very versatile and works for many kinds of distributions (Kirby and Gerlanc, 2013). Also, since bootstrapping relies on a simple algorithm, the computer scientists in HCI may find it easier to intuitively grasp than the traditional analytical approaches (Ricketts and Berry, 1994; Duckworth and Stephenson, 2003). 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 NHST procedures for testing normality, which are logically and practically unsound (Wierdsma, 2013; Stewart-Oaten, 1995, p.2002). When computing exact confidence intervals, departures from normality are not such a big deal: as with the *t*-test, the normality assumption does not concern the population distribution but the sampling distribution of the sample mean[11]. As per the central limit theorem, this distribution turns out to be approximately normal for almost any population distribution shape, provided that the sample size is large enough (Norman, 2010, p.628). One difficulty is that it is often unclear how large is large enough, as it also depends on how much the original population departs from a normal distribution. Another issue with exact confidence intervals is that they are necessarily symmetric, so they do not reflect skewed distributions very well and may cover impossible values. Thus there are merits to using alternative methods (see Tips 12 and 13) if there are reasons to think that the population distribution is not normal. Measurement scales that are strictly positive (e.g., time) or bounded (e.g., percents) cannot be normally distributed. Strictly positive scales are typically positively skewed and approximate a normal distribution once logged (Tip 12). When in doubt, use bootstrapping (Tip 13).

Tip 15: Report interval estimates for 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 interval estimates that convey the numerical uncertainty around those estimates. Many sources are available in textbooks and online on how to compute such intervals. Be aware, however, that not all confidence interval procedures are reliable, in the sense that in some special cases they may produce incorrect intervals (Morey et al, 2015).

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 fair 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 plots do not lie — they just conceal less. 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]? While this is not immediately clear with numerical data, the graphical representation in Figure 3 makes comparison much easier. In addition, plots generally appear less precise than numbers, which likely reduces dichotomous thinking and overconfidence in results.

[11] For more on the important concepts of sampling distribution and the central limit theorem, see, e.g., Cumming (2013, Chap. 3) and the applet at `http://tinyurl.com/sdsim`.

Tip 17: Use numbers wisely. If plots with confidence intervals are already provided, numerical values are not very useful and only produce clutter. However, because the numerical format is more compact, it can be used for reporting secondary results. A complete list of numerical confidence intervals can also be included as a table or in the accompanying material to facilitate comparison with future studies and meta-analysis. However, 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: Do not ignore conventions. When plotting confidence intervals, aim for simplicity and try to stick to the few existing conventions. Ideally, figures should be interpretable with as little contextual information as possible. Changing the level of confidence from the standard 95% to 50% or 99% does not help. Similarly, do not use procedures that "adjust" or "correct" the length of confidence intervals unless there are good reasons to do so. 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. Finally, do not show standard errors (SEs) in your plots. As Cumming and Finch (2005, p.177) have 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."*

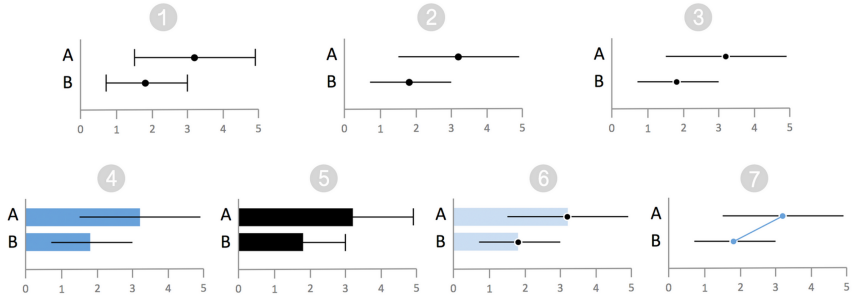

Fig. 9 Seven ways of plotting effect sizes with confidence intervals.

Tip 19: Be creative. The scarcity of graphical standards should be taken as an opportunity to explore custom visual designs, within the limits suggested by Tip 18. For example, there are many options for displaying error bars (Figure 9): while the design (1) is widely used, (2) is a common alternative that has the advantage of de-emphasizing confidence limits; The variant (3) improves the legibility of point estimates; Error bars can also be combined with bar charts (4); Bars help compare magnitudes on a ratio scale (Zacks and Tversky, 1999), but they introduce visual asymmetry (Newman and Scholl, 2012; Correll and Gleicher, 2014) and tend to de-emphasize error bars; This is evident in the so-called "dynamite plots" (5); The design (6) supports ratio comparison while maintaining emphasis on error bars; Finally, error bars can be combined with line charts (7) to convey temporal ordering (Zacks and Tversky, 1999) or within-subject factors (Cumming, 2012, p.172).

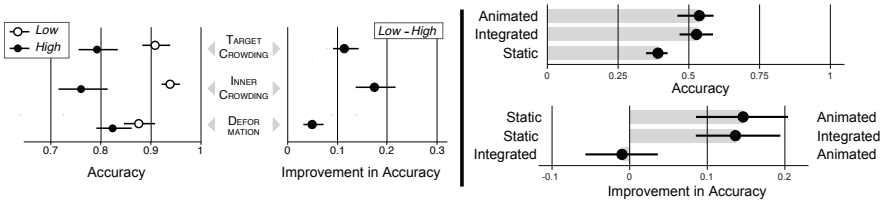

Fig. 10 *Left:* the effects of three animation complexity metrics (one per row) on visual tracking accuracy. The first plot shows mean subject accuracy depending on whether animations are low or high on that metric, while the second plot shows mean within-subject improvement when switching from high to low (after Chevalier et al (2014)). *Right*: the upper plot shows map reading accuracy for three terrain visualization techniques. On the lower plot, 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 between conditions. 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). When comparing multiple conditions (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 the 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 10 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. See Dragicevic (2015) for illustrations. 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 way consists of sorting conditions or pairwise comparisons by effect size. If effects are similar, every replication will lead to a different ordering — thus the plot misleads. Instead, 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 plot, always try to imagine how it could "dance" across replications.

Tip 22: Think beyond averages. Inferences about population means are an important but limited part of statistical analysis and communication. Distributions and individual differences can also be insightful (Vicente and Torenvliet, 2000, pp.250–

[12] *Visual robustness* is related to the concept of *visual-data correspondence* recently introduced in infovis (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).

253). Some empirical data — especially categorical or ordinal data like questionnaire responses — is also hard to faithfully capture with a single aggregated measure. As an alternative, such data can be conveyed without loss of information using compact visualization methods such as matrix displays (Perin et al, 2014). Showing individual observations next to error bars can also be informative and pedagogical (Drummond and Vowler, 2011; Ecklund, 2012). Finally, while confidence and credible intervals are useful for conveying uncertainty about population averages, alternatives such as *tolerance intervals* and *prediction intervals* may be better suited in some cases (Nelson, 2011). Unfortunately, using interval estimates with very different meanings may exacerbate confusions surrounding the meaning of error bars.

4.5 Interpreting Confidence Intervals

Interpreting confidence intervals is a key aspect of estimation. It is hard to master, and it could almost be called an "art". Despite this, not much has been written on the topic. Here is a list of recommendations 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 those created by Cumming (2009a) are a powerful tool for building this intuition, perhaps more so than mathematical formulas. Watch simulations over and over again, and 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 $\alpha = .05$ level. This is only a convenient reference point, *not a hard rule* to be applied mindlessly (see Tip 25). Cumming also explains how to visually compare confidence intervals in between-subjects designs: if two error bars overlap by less than $1/4$ of their average length, then the difference is statistically significant at the $\alpha = .05$ level. This rule is also just a convenient rule of thumb, and should not be used in a binary way. In within-subject designs, the $1/4$ overlap rule is often (but not necessarily) conservative. Pairwise differences of interest thus need to be plotted and interpreted separately (Cumming and Finch, 2005, p.176) (see Figure 10).

Tip 25: Ban dichotomous interpretations. Using confidence intervals to provide yes/no answers defeats the whole purpose of estimation: *"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). Plausibility does not suddenly drop when crossing the confidence limits (See Section 2.1). A confidence interval can be thought of as abstracting a continuous "plausibility" function (Cumming, 2012, pp.98–102; Cumming, 2013, p.17). While a recent study has explored alternatives to error bars that visually convey this continuity (Correll and Gleicher, 2014), the classical error bar has

the benefit of being visually cleaner and more economical in terms of space and data-ink ratio (Wainer, 1984, p.139). The edges of error bars offer visual reference points, whereas other representations such as color gradients may not.

Tip 26: Use vague language and hedges. We could say that Figure 6 *"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 necessary for acknowledging and honestly conveying the uncertainty present in effect size estimates. Vague language — which is not the same as ambiguous language — plays a key role in reasoning (van Deemter, 2010). In the face of uncertainty and complexity, the only alternatives to vagueness are false clarity (van Deemter, 2010, Chap. 1) and pseudo-objectivity (Thompson, 1999). The term "seem" in the text above is a *hedge*, and hedges are also important in science communication (van Deemter, 2010, Chap. 6). There are many ways Figure 6 can be described using text, and different investigators will use different wordings. The subjective nature of this task should not make the researcher feel uncomfortable. It is important to be objective when performing planned analyses and turning them into numbers and plots, but after that one can afford to be subjective, knowing that no reader is forced to accept one's conclusions. That said, wordings that misrepresent or exaggerate findings naturally tend not to give a 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"*.

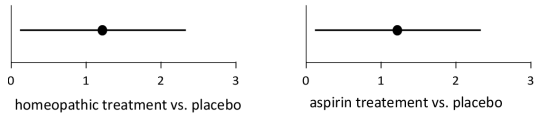

Fig. 11 Identical confidence intervals calling for different interpretations.

Tip 28: Use external information. A key part of empirical research consists of interpreting results in relation to externally available information. For example, the imaginary plot on Figure 11 (left) shows the results of a study assessing the clinical efficacy of a homeopathic treatment. While the data speaks in favor of this treatment, the scientific consensus is that such treatments are ineffective, and thus the result 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 only 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 result. Confidence intervals have both the drawback and the advantage of moving the burden

of Bayesian reasoning to the investigator and the readers. The *cost of error* is also important. 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).

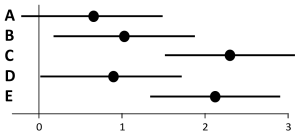

Fig. 12 In this plot, each confidence interval needs to be interpreted in relation to other confidence intervals.

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 12, there is nothing wrong with 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 3.3.2 for the danger of not doing so). A less misleading way of describing Figure 12 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 for information about 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 reported and interpreted separately. Providing non-misleading interpretations of figures with confidence intervals requires judgment, and no mechanical decision procedure can carry out this job better than a thoughtful investigator.

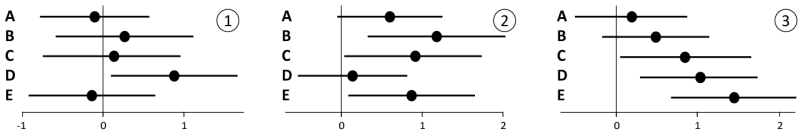

Fig. 13 Three possible sets of confidence intervals that can be tricky to summarize.

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 who purports to have extrasensory perception (ESP) abilities is subjected to testing sessions labeled *A*-*E* (Figure 13-1). 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 particularly low. The chances of observing a similar deviation in *any* of the five tests is even higher: about 20%. If 40 tests were conducted, the possibility of a false positive would be about 90% — almost certain. 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 13-2 may conclude that the technique is imperfect because one task yields inconclusive results. However, such an outcome is likely

to occur even if the technique is better in all respects (e.g., a consistent population difference of 1 for all tasks — see again the dance in Figure 7).

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, Chap. 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"*. Complex results should not be oversimplified in the paper's abstract or conclusion.

Finally, you may find on occasion a striking pattern, as in Figure 13-3. 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, it is best to express a pattern with a single number (e.g., a grand mean, difference, interaction or correlation) and report a confidence interval. However, such analyses are largely uninformative if they are unplanned (see Tip 5).

Tip 31: Do not mindlessly correct for multiplicity. Correction procedures (e.g., Bonferroni correction) that account for the ESP scenario in Tip 30 (Figure 13-1) are part of the NHST ritual, and have been adapted to confidence intervals (Bender and Lange, 2001; Tryon, 2001). Such procedures are a powerful safeguard against fallacious reasoning, but they are too conservative in the many situations where results are not combined in a disjunctive manner. They also have the drawback of changing the meaning of error bars (Cumming, 2009b, p.209) (see Tip 18). Multiplicity correction procedures are far from having reached a consensus (Bender and Lange, 2001) and there are strong arguments against their systematic use (Wilson, 1962; Stewart-Oaten, 1995, p.2003). Between the fervent defenders and the strong opponents, estimation advocates tend to take a laissez-faire position (Cumming, 2012, pp.421–423; Cumming and Finch, 2005, p.177; Baguley, 2012, pp.173–174). The principle of simplicity provides an argument for skipping such procedures, while the principles of transparency and clarity requires that issues related to joint probabilities are kept in mind and pointed out if necessary. If a conclusion follows from a disjunctive combination of many confidence intervals, it can be useful to also report multiplicity-corrected confidence intervals. The best solution 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 fair 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 chapter focuses on uncertainty due to sampling 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, pp.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 to make *"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 essential in order for other researchers to understand the details of a study and greatly facilitates replication. Similarly, experimental data (all data tables and possibly analysis scripts) is necessary for conducting re-analyses and meta-analyses. To be useful this material should be freely shared online upon paper acceptance.

For errata and updates, go to `www.aviz.fr/badstats`.

5 Conclusion

When assessing the quality of a statistical analysis in an HCI paper, reviewers often tend to exclusively focus on the rigorous application of statistical procedures. This reveals several misconceptions about statistics. One is the belief that there exists a set of "correct", universally-accepted 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 alternative methods in order to produce *"the illusion that there is only one tool"*. Another belief is that statistical procedures can produce rigorous knowledge, just because they can output precise numbers. However, there is no such information in the data. Data is uncertain and messy, and so are statistics.

This chapter has introduced some basic principles of *fair statistical communication*, i.e., principles for conveying uncertainty in empirical data in a way that does not prompt misinterpretations, so that as-wide-as-possible an audience can judge and decide whether or not to trust the authors' conclusions. There are many questions this chapter does not address (e.g., what is a good research question? How to design experiments?) and it does not stand by itself as a guide to statistics — interested readers will need to go through some of the literature. Also, none of the tips offered in Section 4 should be taken as a dogma. Closely following guidelines will never be a necessary nor a sufficient condition for an article to merit publication. However, I do think that the general principles of fair statistical communication outlined in Section 4.1 should be given more consideration in peer reviewing.

Among the worst enemies of good statistical communication is dichotomous thinking. False dichotomies exist at many levels in current statistical practice, and have been greatly 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 and trying to achieve clarity and transparency while fully embracing the messiness of our data (Giner-Sorolla, 2012) can not only benefit science, but can also make statistical analysis a much more exciting and rewarding experience.

Many issues outlined in this chapter are old and have been raised in many disciplines. But things seem to be changing — in the most recent 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, 2010). Meanwhile, high-impact psychology journals are starting to enforce the use of estimation (Eich, 2014). Large collective initiatives, whose values overlap a lot with the idea of fair statistical communication, are also being started under the umbrella of the "open science" movement[13]. In parallel, more and more articles on bad statistics are being published in newspapers and blogs, raising the public's awareness of those issues. It may well be that statistical practice will be very different only a few years from now.

Since statistics are nothing but user interfaces meant to help researchers in their task of producing and disseminating knowledge, the fields of HCI and infovis can 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 for communicating 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

Many thanks to Elie Cattan, Fanny Chevalier, Geoff Cumming, Steven Franconeri, Steve Haroz, Petra Isenberg, Yvonne Jansen, Maurits Kaptein, Heidi Lam, Judy Robertson, Michael Sedlmair, Dan Simons, Chat Wacharamanotham and Wesley Willett for their helpful feedback and comments.

References

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

Abelson RP (1997) A retrospective on the significance test ban of 1999. What if there were no significance tests pp 117–141

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

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

Bååth R (2015) The non-parametric bootstrap as a Bayesian model. Online, URL `http://tinyurl.com/bayes-bootstrap`

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

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

Bayarri MJ, Berger JO (2004) The interplay of Bayesian and frequentist analysis. Statistical Science pp 58–80

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

Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science 1(3):140,216

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]. URL `http://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

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, URL http://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

Dragicevic P (2012) My technique is 20% faster: Problems with reports of speed improvements in HCI. Research report

Dragicevic P (2015) The dance of plots. Online, URL http://www.aviz.fr/danceplots

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

Eich E (2014) Business not as usual (editorial). Psychological Science 25(1):3–6, URL http://tinyurl.com/psedito

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

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

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

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

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 (2004) Type 1, type 2, type S, and type M errors. Online, URL http://tinyurl.com/typesm

Gelman A (2013a) Commentary: p-values and statistical practice. Epidemiology 24(1):69–72
Gelman A (2013b) Interrogating p-values. J Math Psychol 57(5):188–189
Gelman A, Loken E (2013) The garden of forking paths. Online article
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. J Socio Econ 33(5):587–606
Gigerenzer G, Kruger L, Beatty J, Porter T, Daston L, Swijtink Z (1990) The empire of chance: How probability changed science and everyday life, vol 12. Cambridge University Press
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? J Exp Educ 71(1):83–92
Goldacre B (2012) What doctors don't know about the drugs they prescribe [TED talk]. Online, URL http://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
Greenland S, Poole C (2013) Living with p values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology 24(1):62–68
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
Hoekstra R, Finch S, Kiers HA, Johnson A (2006) Probability as certainty: Dichotomous thinking and the misuse of p values. Psychonomic Bulletin & Review 13(6):1033–1037
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: Proc. 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

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

Lambdin C (2012) Significance tests as sorcery: Science is empirical, significance tests are not. Theory & Psychology 22(1):67–90

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

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

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

Loftus GR (1993) A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. Behavior Research Methods, Instruments, & Computers 25(2):250–256

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. Q J Exp Psychol 43(4):907–912

Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ (2015) The fallacy of placing confidence in confidence intervals (version 2). Online draft, URL http://tinyurl.com/cifallacy

Nelson MJ (2011) You might want a tolerance interval. Online, URL http://tinyurl.com/tol-interval

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. Psychon Bull Rev 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
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716+
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
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. Chem Eng 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. Hum Commun Res 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

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, URL http://tinyurl.com/trafimow

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/

Wainer H (1984) How to display data badly. Amer Statist 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

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: Proc. 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