Statistical methods for linguistic research: Foundational Ideas – Part I

Shravan Vasishth

University of Potsdam, Potsdam, Germany and

CEREMADE (Centre de Recherche en Mathématiques de la Décision), Université Paris Dauphine, Paris France

Bruno Nicenboim

University of Potsdam, Potsdam, Germany

May 31, 2016

Abstract

We present the fundamental ideas underlying statistical hypothesis testing using the frequentist framework. We start with a simple example that builds up the one-sample t-test from the beginning, explaining important concepts such as the sampling distribution of the sample mean, and the iid assumption. Then we examine the meaning of the p-value in detail, and discuss several important misconceptions about what a p-value does and does not tell us. This leads to a discussion of Type I, II error and power, and Type S and M error. An important conclusion from this discussion is that one should aim to carry out appropriately powered studies. Next, we discuss two common issues we have encountered in psycholinguistics and linguistics: running experiments until significance is reached, and the "garden-of-forking-paths" problem discussed by Gelman and others. The best way to use frequentist methods is to run appropriately powered studies, check model assumptions, clearly separate exploratory data analysis from planned comparisons decided upon before the study was run, and always attempt to replicate results.

Keywords: Frequentist methods, data analysis, linguistics, psycholinguistics

Please send correspondence to vasishth@uni-potsdam.de.

Introduction

Psycholinguistics has a long tradition of using experimental methods, but in recent years, linguists working in areas such as syntax, semantics, and pragmatics have also started to embrace empirical methods (see Arunachalam, 2013 for a review of the more commonly used methods). As a consequence, basic familiarity with experiment design is becoming a core requirement for doing linguistics. However, just knowing how to carry out an experiment is not enough; a good understanding of statistical theory and inference is also necessary. In this article, we present the most important issues that researchers need to be aware of when carrying out statistical inference using frequentist methods (as opposed to Bayesian approaches, see Part II of this review). We focus on frequentist methods in this article because these are the statistical tools of choice in psycholinguistics and linguistics; examples are the t-test, analysis of variance (ANOVA), and linear mixed models. Given software such as R (R Core Team, 2014), it is extremely easy to obtain statistics such as t- or F-values, and the corresponding p-values. However, it is equally easy to misunderstand what these mean; in particular, a misinterpretation of the p-value often leads researchers to draw conclusions from their data that are not supported by the underlying statistical theory. We start the paper illustrating the meaning of t- and p-value and discussing some common misconceptions by means of the one-sample t-test or (equivalently) the paired t-test. We use this test as an example because of its relative simplicity and because it happens to be a very frequently used one in linguistics and psycholinguistics. For ANOVAs and (generalized) linear mixed models, the situation is more complex, but the same logic and issues described below also apply. We then show the importance of power, and Type I, II, S, and M errors using simulations based on linear mixed models. Two further important topics discussed are: the problems involved in running participants till significance is reached, and the issues involved in experiments with multiple measures, multiple regions of interest, and too many degrees of freedom in analysis. Although our focus is on factorial repeated measures designs used in psycholinguistics and linguistics, our discussion is also relevant for research in other areas of linguistics that depends on null hypothesis significance testing.

Examples of such areas are corpus linguistics (Gries, 2009) and sociolinguistics (Wieling, Montemagni, Nerbonne, & Baayen, 2014).

A simple example: a two-condition repeated measures design

Consider the case of a two-condition repeated measures self-paced reading (Just, Carpenter, & Woolley, 1982) experiment, e.g., subject versus object relative clauses, where the dependent measure is reading time in milliseconds; assume that reading time is measured at a particular region of interest in the relative clause sentences. Suppose the experiment has n randomly sampled participants, each of whom read multiple instances of subject and object relative clauses in a counterbalanced Latin square design (Arunachalam, 2013). A typical approach taken is to calculate the mean of each participant's reading time for each relative clause type by aggregating over all the items that the participant saw. To make this example concrete, consider the simplified situation in Table 1 where a participant, labeled 1, sees three items for each condition; normally, of course, each participant will be shown many more items. The condition labels in Table 1 refer to subject relatives (condition a) and object relatives (condition b). If we average the three data points from the participant for condition a and for condition b, we obtain the aggregated data shown in Table 2.

participant id	item id	condition	reading time
1	1	a	500
1	2	a	600
1	3	a	700
1	4	b	450
1	5	b	550
1	6	b	650

Table 1
Hypothetical unaggregated reading time data from a two-condition experiment for one participant who saw six items, three from each condition.

This procedure, applied to each of the participants in an experiment, results in two data points from each participant, one for each condition. This is called a by-participants (or by-subjects) analysis. One can analogously do a by-items analysis by aggregating over participants.

participant id	condition	reading time
1	a	600
1	b	550

Table 2
The result of aggregating the data over items for participant 1 in Table 1.

After this aggregation procedure, we have n data points for subject relatives and n for object relatives. The data are paired in the sense that for each participant we have an estimate of their reading time for subject relatives and for object relatives. In the hypothetical example above, for participant 1, we have a mean subject relative reading time of 600 ms and an object relative reading time of 550 ms. If, for each participant, we take the difference in object vs subject relative reading time (for participant 1 this would be -50 ms), we have a vector of n values, $x_1, \dots, x_n$ that are assumed to be *mutually independent*, and represent the difference in OR vs SR reading times for each participant. Another assumption here is that these observed differences between RC types $x_1, \dots, x_n$ are generated from a normal distribution with some unknown mean μ and standard deviation σ. Since each of the data points is assumed to come from the same distribution, we say that they are *identically distributed*. The independence assumption mentioned above and the identical-distribution assumption are often abbreviated as iid—independent and identically distributed. The statistical test depends on the iid assumption, and the assumption that a simple random sample of participants has been taken. For example, if we were to do a t-test on the unaggregated data, we would violate the independence assumption and the result of the t-test would be invalid. When distributional assumptions (such as the normality assumption of residuals in linear mixed models, see Sorensen, Hohenstein, & Vasishth, 2015 for more discussion) are not met, the parametric bootstrap (Efron & Tibshirani, 1994) is an option worth considering. The bootstrap can also be used for linear mixed models; for an example, see Appendix D of Boston, Hale, Vasishth, & Kliegl, 2011.

Returning to the t-test under consideration, we begin by generating some fake data; all the R code used in this paper is available from the first author's home page. Let us simulate n data points representing differences in reading times, as an illustration. Since reading time data typically have a log-normal distribution (for reasons discussed later in this paper), we will simulate the difference from two log-normal distributions with mean 6 ($\exp(6) = 403.43$ ms) and 5.99 ($\exp(5.99) = 399.41$ ms), and the same standard deviation of 0.8 with $n = 1000$ draws each (Figure 1).

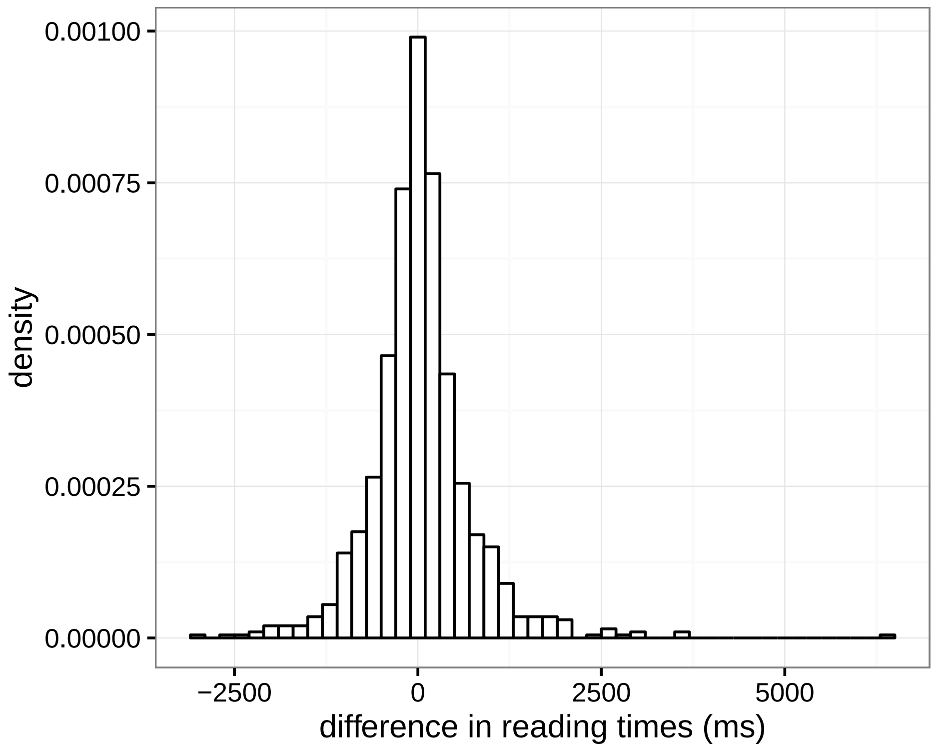

Figure 1. Histogram of simulated reading time data (differences in reading time between two conditions).

Given such a sample, the one-sample t-test (or, equivalently the paired t-test) works as follows. We first compute the sample mean and the sample standard deviation $\bar{x}$ and $\hat{\sigma}$; these are estimates of the unknown parameters, μ and σ, of the underlying distribution that is assumed to have generated the data. It is important to note the distinction between the sample mean $\bar{x}$ and the unknown true point value μ. The true parameter μ or β is unknown; we are reporting an estimate of this unknown value.

Statistical inference relies on an important property of the *sampling distribution of the sample means* under repeated sampling: For a large enough sample size n, the distribution of the sample means under repeated sampling will be normal with mean μ and standard deviation $\sigma/\sqrt{n}$; this is assuming that the underlying distribution that generated the data has a mean and standard deviation.[1] The preceding sentence is an informal statement of the *central limit theorem*. The standard deviation of the sampling distribution of the sample means is also called the *standard error* (SE), and can be estimated from the data by computing $\hat{\sigma}/\sqrt{n}$.

Statistical inference in the frequentist paradigm begins by positing a null hypothesis distribution, which is a statement about what the true sampling distribution of the sample means might look like. In our example, our null hypothesis is that the difference in means between the two RC types is 0. We will follow standard practice in writing this null hypothesis as $H_0 : \mu = 0$; μ represents the true, unknown difference in means between the two RC types. Next, we use the fact that the transformation $T = (\bar{x} - \mu)/SE$ has a t-distribution with $n - 1$ degrees of freedom, where n is the sample size (for large n, the t-distribution approximates the normal distribution with mean 0 and variance 1). Since we have hypothesized μ to be 0, and since we have estimated $\bar{x}$ and $\hat{\sigma}$ from the data, we can compute the observed t-value $t = (\bar{x} - 0)/SE$. This observed t-value is the distance between the sample mean and the hypothesized mean, in SE units; this is easy to see if we rearrange the equation as follows: $t \times SE = \bar{x} - 0$: the sample means $\bar{x}$ is t standard errors away from the hypothesized mean 0. Intuitively, if the t-value—the distance in SE units between the sample means and the hypothesized mean—is large, we feel entitled to reject the null hypothesis. It is traditional to compute the p-value associated with the observed t-value; this is the probability of observing a t-value at least as extreme as the one we observed, conditional on the assumption that the null hypothesis is true. It is also traditional to reject the null hypothesis if this conditional probability falls below 0.05.

In the above example with simulated data, our sample mean is 42.15, and the standard deviation is 706.98, leading to an observed t-value of 1.89. Can we reject the null here

[1] We mention this assumption because distributions like the Cauchy have no mean or variance. For such distributions, the above statement does not apply. This detail essentially has no practical implications in linguistics.

following our rule above, which is to reject if the p-value is below 0.05? We can start by visualizing the null hypothesis distribution and the observed t-value; see Figure 2. The p-value is the area under the curve to the right of the black dot (our observed t-value), plus the same area under the curve to the left of the gray dot. The black and gray dots mark the observed t-value 1.89 and -1.89 (if the observed t-value had been negative, i.e., if the sample means had been less than 0, we would have similarly considered both signs). We consider both sides of the distribution because we want to know the probability of seeing the absolute t-value or a value more extreme, regardless of sign; this is called the two-sided one-sample t-test. The p-value in our example is 0.06, because the observed t-value is not very far out in the tails of the null hypothesis distribution.[2]

So, in this simulated example, we would fail to reject the null hypothesis. That is the t-test in a nutshell, and it will serve as the basis for further discussion about statistical inference.

Four misconceptions about what a p-value tells us

It is very easy to overlook the fact that the p-value is a conditional probability. Neglecting to attend to this detail has led to quite a few misconceptions about what a p-value tells us. Here are some claims about the p-value that are incorrectly believed to be true. Note that the same misconceptions hold if one considers, instead of the p-value, the absolute value of the observed t-value, as is commonly done in connection with linear mixed models.

[2]For our current example with sample size $n = 1000$, we can easily compute this p-value "by hand" by using the function `pt` in R. This function computes the area under the curve—the probability—in the upper and lower tails of the relevant t-distribution (the t-distribution with parameter $n - 1$). Given an observed t-value `obs_t`, the p-value can be computed by running the R command: `pt(obs_t,df=999,lower.tail=FALSE)+pt(-obs_t,df=999,lower.tail=TRUE)`.

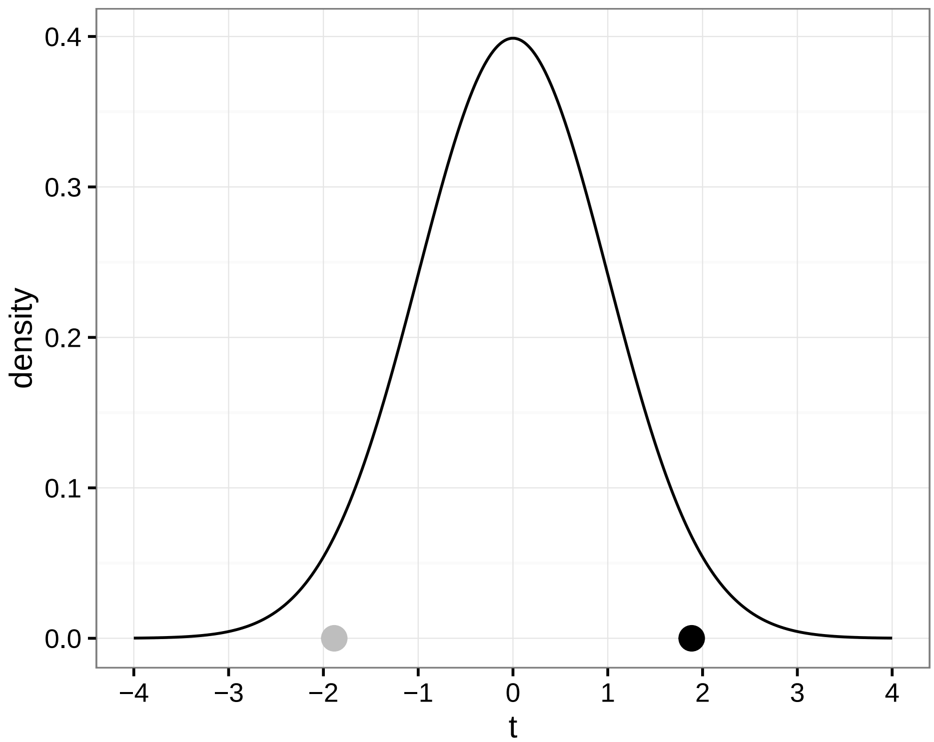

Figure 2. The null hypothesis distribution and the observed t-value.

Misconception 1: The smaller the p-value, the greater the confidence in the *specific* alternative hypothesis we are interested in verifying

The p-value provides a measure of the evidence against the null hypothesis. The smaller the p-value, the greater the confidence that the null hypothesis is false. The p-value is not a measure of the evidence in favor of our specific hypothesis of interest; it doesn't tell us *which* of the infinity of possible alternative μ is true. Rejecting the null doesn't give us any statistical evidence for the *specific* effect our theory predicts, it just gives us evidence against a very specific hypothesis that $\mu = 0$, and allows all other values of μ to be plausible, even ones we would not be happy to see as the outcome of an experiment. We want to know $\text{P}(H_1 \mid \text{data})$, i.e., how probable is it that our theory is correct (the specific alternative H_1) given the data that we have, but frequentist statistics tells us $\text{P}(\text{data} \mid H_0)$, i.e., the probability of the data (more accurately, the probability of seeing a test statistic as extreme or more extreme than the one observed) given the null hypothesis. Importantly, we cannot infer one conditional probability just by knowing its inverse. Dienes (2011) illustrates this with a very graphic example: The probability of dying given that a shark has bitten one's head clean off, P(dead | head bitten clean off by shark), is one. But most people die of other causes; given that one is dead, the probability that a shark has bitten one's head clean off, P(head bitten off by shark | dead), is close to zero.

In summary, the p-value answers *a* question, but it doesn't answer the question we are interested in. To answer the question we are actually interested in, namely whether the effect is positive or negative with a certain magnitude, we make an indirect inference by looking at the sign of the observed mean, and then drawing conclusions about whether our theory is supported or not. The p-value alone does not provide any evidence for our theory. The most informative piece of information we have about our specific hypothesis is actually the sample mean and the uncertainty associated with our estimate of this sample mean: the standard error.

Misconception 2: A p-value greater than 0.05 tells me that the null hypothesis is true

This is perhaps the commonest mistake seen in linguistics and psycholinguistics. Researchers in linguistics and psycholinguistics (and also psychology) routinely make the strong claim that *There is no effect of factor X on dependent variable Y*, based on their getting a p-value larger than 0.05. This claim can only be made when power is high, as discussed below.

This misconception arises because researchers do not consider the fact that the p-value is a conditional probability: the probability of getting a statistic as extreme or more extreme as the one we got, conditional on the null hypothesis being true. To conclude that the null is true when $p > 0.05$ is like arguing that, if the probability of the streets being wet given that it has just rained is higher than 0.05, then we can conclude that it has just rained.

Unless one has sufficient power (see below), the best one can say when we get a p-value larger than 0.05 is that “we failed to find an effect”. Instead, in linguistics and psycholinguistics it is routine to make the much stronger, and statistically invalid, claim that “there is no effect”.

Most probably, what researchers really mean when they decide to accept the null as true is not that $\mu = 0$, i.e., not that the true mean μ is literally 0, but rather that the true mean is in the vicinity of 0. For such hypotheses, one can reverse the null and alternative hypothesis using the equivalence testing approach (for an example, see Stegner, Bostrom, & Greenfield, 1996). This is a much better way to argue for the null hypothesis, but requires the researcher to make a commitment as to the range of values that count as representing “no effect”.

Misconception 3: Two nested comparisons allow us to draw conclusions about interactions

To illustrate the issue here, consider the results reported in Experiment 2 of Husain, Vasishth, & Srinivasan, 2014; we use their data because it is publicly available. In the published paper, the authors had used a Bayesian linear mixed model to report the results, but here we will do the same analysis using a series of one-sample t-tests discussed above. This experiment was a 2×2 factorial design, with one factor predicate type (complex vs simple) and the other factor distance between the verb and an argument noun (long vs short). We can look at the effect of distance within complex predicates and simple predicates separately. These comparisons can be done easily following the procedure of the one-sample t-test we discussed in the beginning of this paper. We convert reading times to the log scale to do the test. When we do these tests, we find a statistically significant effect of distance in complex predicates (t(59)=-2.51, p-value=0.02), but we don’t get a significant effect of distance in simple predicates (t(59)=0.52, p-value=0.61). This is shown graphically in Figure 3. Can we now conclude that the interaction between the two factors exists? No. One can check with a t-test whether the interaction is statistically significant. In the first test, the null hypothesis is that the true difference, δ_1, between the (unknown) means of the long vs short conditions in the complex predicate case is 0; in the second test, the null hypothesis is that the true difference, δ_2, between the (unknown) means of the long vs short conditions in the simple predicate case is 0. In the interaction, the null hypothesis is that the difference between these two differences, $\delta_1 - \delta_2 = \delta$ is 0. When we do the t-test with this null hypothesis (see accompanying code for details), we find that we cannot reject the null hypothesis that $\delta = 0$: (t(59)=-1.68, p-value=0.1).

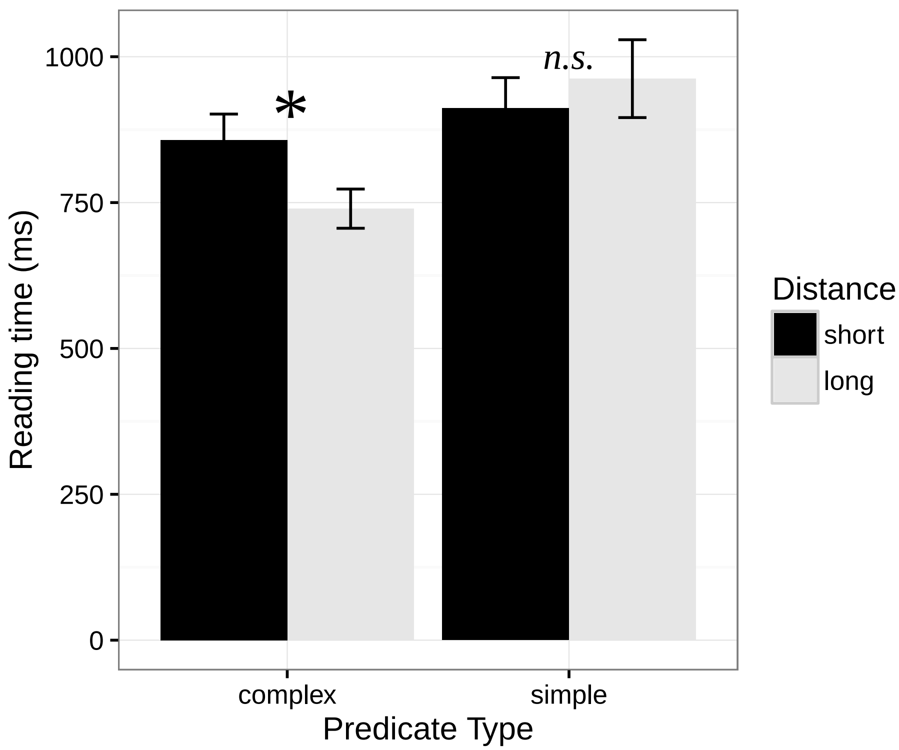

Figure 3. Example data, taken from Husain et al. 2014, showing a typical situation where one comparison is statistically significant, but the other is not. From this, we cannot conclude that the interaction is statistically significant.

The first comparison (the effect of distance in complex predicates) yields a difference of sample means -0.13 log ms, and the standard error of this difference is 0.05. The second comparison (the effect of distance in simple predicates) yields a difference of sample means 0.03, and the standard error of this difference is 0.06. From the first two comparisons, we cannot conclude that the interaction will necessarily be significant; indeed, in our example, the difference between these differences in means is -0.16, with standard error 0.09. Thus, one must always check whether the interaction is significant. This is a real issue in psychology and linguistics and has serious consequences for theory development; many papers have misleading conclusions that follow from this error. As evidence, Nieuwenhuis, Forstmann, & Wagenmakers, 2011 present a survey of published articles showing that approximately 50% of them (79 articles) draw this incorrect inference. A specific example of such an error comes from the first author of the present article: Vasishth and Lewis (2006, 787) argued that in one experiment (their Expt 2) a particular effect occurred, but in a subsequent experiment (their Expt 3) the effect disappeared, leading to the incorrect conclusion that the new factor in their Experiment 3 led to the disappearance of the effect. This would have been only valid if the relevant interaction had been found (this was never checked in Vasishth & Lewis, 2006).

Misconception 4: If $p < 0.05$, we have found out that the alternative is in fact true

This misconception is perhaps encouraged by the language we use to describe our success in getting a p-value; the effect is “reliable” or “real” or “generalizable”. The word “significant” also contributes to giving the feeling that something of importance has been found. It doesn’t help that textbooks and articles explaining the p-value often state that the p-value tells us whether “the effect is due to chance”. If this were literally what the p-value meant, it would be reasonable to conclude that if the p-value is low, then the effect is *not* due to chance. But this characterization of the p-value is not correct. The phrase “due to chance” is more accurately expanded to “due to chance under the null hypothesis”. Stated correctly in this way, if we get a very low p-value, we can only say that, *assuming that the null is true*, the probability of observing the t-value (or some value more extreme) is very low; it is on the basis of this low probability that we reject the null. No absolute certainty is afforded by the p-value, no matter how low it is.

In fact, no matter how low our p-value, we will have a 0.05 probability of having mistakenly rejected the null when the null is in fact true. Thus, a p-value (regardless of whether it is low or on the border of 0.05) alone should not convince us that an effect is "real" or "generalizable"; successful replications of the effect are much more convincing. We discuss this point further in the next section.

Type I, II error, power, and Type S, M error

The logic of frequentist methods is inextricably linked with hypothetical repeated sampling. If we were to repeatedly run an experiment, we would essentially get a different sample mean in every repeated sample; in some of these samples, we will reject the null, and in some other samples we will fail to reject the null. Under such repeated sampling—which we almost never have the luxury of doing in real life, incidentally—we can define the probability of incorrectly rejecting the null (when it's actually true); this is called *Type I* error (also called the α value) and is typically set at 0.05 by the researcher. Type I error is conventionally fixed at 0.05 before we run an experiment. Note that it is not the p-value: the p-value is computed based on the data you have at hand, and will vary depending on your data, whereas Type I error is a fixed rate of incorrect rejections of the null under hypothetical repeated sampling.

Under repeated sampling, the probability of incorrectly failing to reject the null hypothesis when it is false with some specific value is called *Type II* error. The quantity (1-Type II error) is called *power*, and is the probability of correctly rejecting the null.[3] Table 3 shows the four possible states when we consider the two possible states of the world (null true or false) and the binary decision we can take based on the statistical test (reject the null or not).

It may also help to see a visualization of Type I and II errors. Consider two different situations: in the first one, the true $\mu = 0$; i.e., the null hypothesis is true; in the second one, the true $\mu = 2$; i.e., the null hypothesis is false with a specific value for the alternative. Type I error will be relevant for the first situation, and it is illustrated as the black-colored

[3] Note that all our definitions here are with respect to the null hypothesis—it is a mistake to think that Type II error is the probability of failing to accept the alternative hypothesis when it is true. We can only ever reject or not reject the null; our hypothesis test is always with reference to the null.

Reality:	H_0 TRUE	H_0 FALSE
Decision: 'reject':	α **Type I error**	$1 - \beta$ **Power**
Decision: 'fail to reject':	$1 - \alpha$	β **Type II error**

Table 3
The two possible states of the world (the null be either true or false) and the two possible decisions we can take given our data.

area under the distribution representing the null hypothesis in Figure 4. The area under the curve in these regions gives us the total probability of landing in these regions under the null. The figure also shows Type II error; this is the gray-colored region under the *specific* alternative hypothesis $\mu = 2$. We determine this area by first drawing the two vertical lines representing the points beyond which we would reject the null; then we compute the probability of landing within these points *under the specific alternative*; this is exactly the gray-colored area. Power is not shown, but since power is 1-Type II error, it is all the area under the curve for the distribution centered around 2 excluding the gray-colored area.

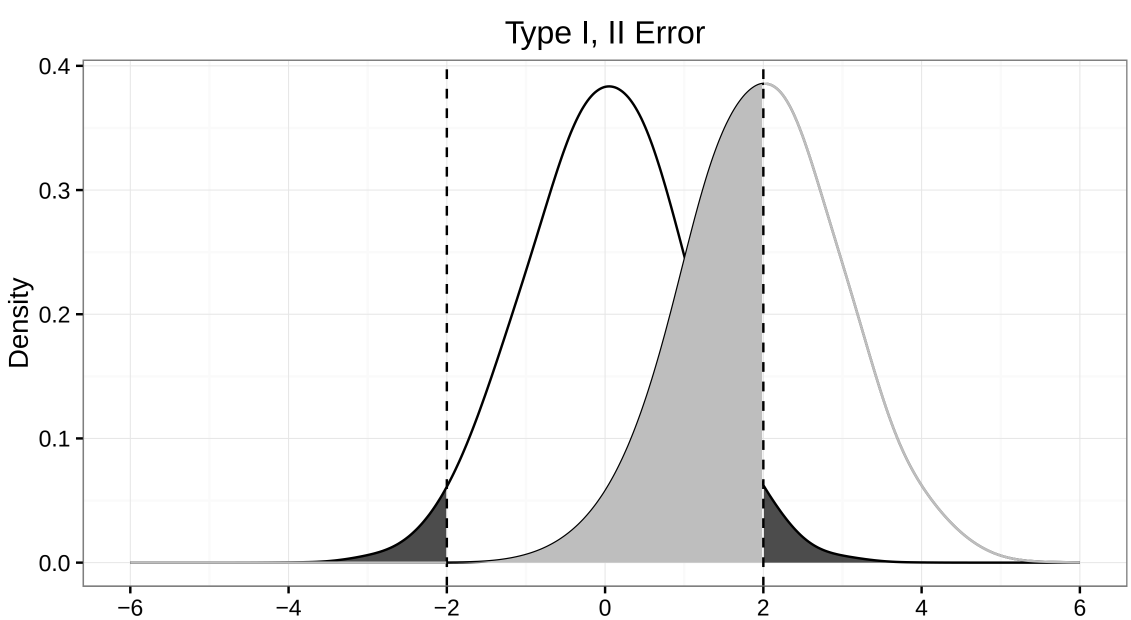

Figure 4. An illustration of Type I error (the area shaded black) and Type II error (the area shaded gray) for a specific case where the true value of μ is 2.

Power is best thought of as a function because we can only talk of power with reference to a specific alternative value for the magnitude (different from zero) of an effect (μ), and the sample size and standard deviation of the sample. For different magnitudes μ of the effect we are interested in, power will differ: in the scenario shown in Figure 4, the further away μ is from 0, the larger the power. The reader can verify this by considering what would happen to the Type II error region in Figure 4 if μ were 4 instead of 2. Clearly, Type II error would decrease, and therefore power would increase. Considering different values of μ in this manner, the power function is a curve that shows how power changes as μ changes (Figure 5). A practical implication is that one way to increase power is to design an experiment with a stronger manipulation, one which will lead to a larger effect. Another way to increase power and decrease Type II error in a study is by increasing the sample size. Figure 6 shows how power changes as sample size changes. Yet another way is to measure the dependent measure more precisely, thereby reducing the standard deviation. For example, eyetracking data is extremely noisy, which may lead to an overestimate of the standard deviation. More frequent recalibration, using better equipment and well-trained experimenters could yield better estimates.

It is especially important to do the best one can to achieve high power if we are interested in arguing for the null. The reason for this should be obvious at this point: low power implies high Type II error, which means that any failure (even repeated failures) to reject the null may just be due to the fact that the probability of accepting the null when the null is in fact false is very high. There are many situations when you might want to argue for the null (see, e.g., Phillips, Wagers, & Lau, 2009); here, it is imperative to put in extra effort into achieving as much power as possible (see Jäger, Benz, Roeser, Dillon, & Vasishth, 2015 for an example).

In order to be able to compute a reasonable estimate of power for a future study involving a comparison of two conditions, it is helpful to have an estimate of the difference between the conditions in ms or log ms. Now, this of course depends on having some way to determine a realistic estimate of the true effect size for a particular phenomenon. This

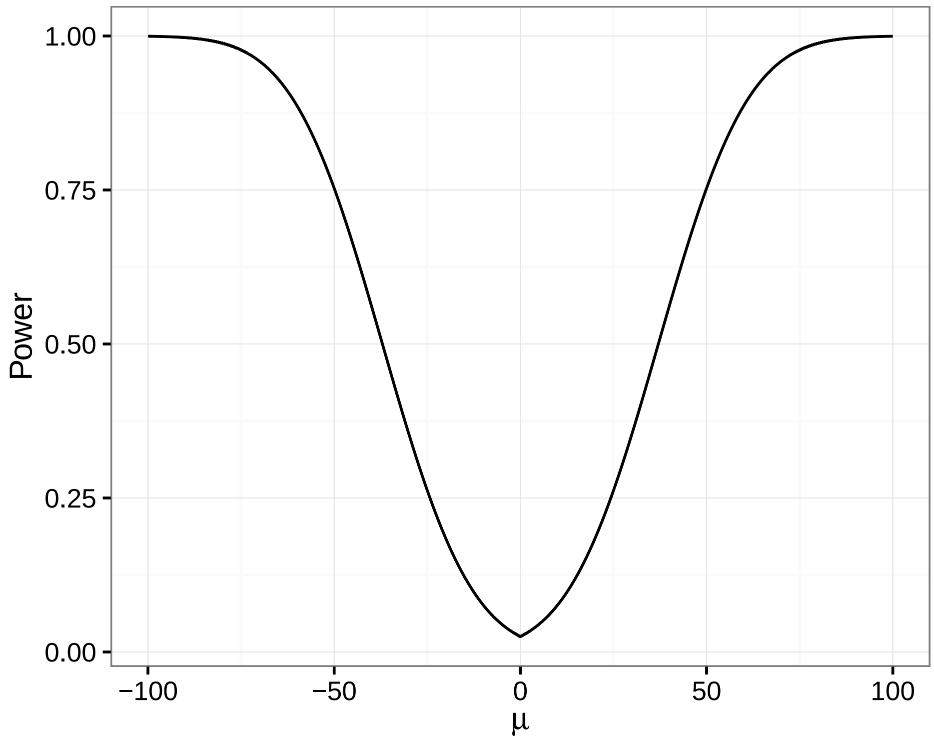

Figure 5. An example of a power function for different effect sizes, assuming (for illustration) a standard deviation of 40 and a sample size of 10.

could be done through a meta-analysis or literature review (Engelmann, Jäger, & Vasishth, 2016; Mahowald, James, Futrell, & Gibson, 2016; Vasishth, Chen, Li, & Guo, 2013), or knowledge elicited from experts on the topic you are studying (Vasishth, 2015). If such an estimate of the effect can be computed, then one can and should also compute Type S and M errors (Gelman & Carlin, 2014). These are defined as follows:

1. Type S error: the probability that the sign of the effect is incorrect, given that (a) the result is statistically significant, or (b) the result is statistically non-significant.

2. Type M error: the expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size (conditional on whether the result is significant or not). Gelman and Carlin also call this the exaggeration ratio, which is perhaps more descriptive than "Type M error".

To illustrate Type S and M errors, we simulate data with similar characteristics as the data from Gibson and Wu (2013); their experiment had a simple two-condition design

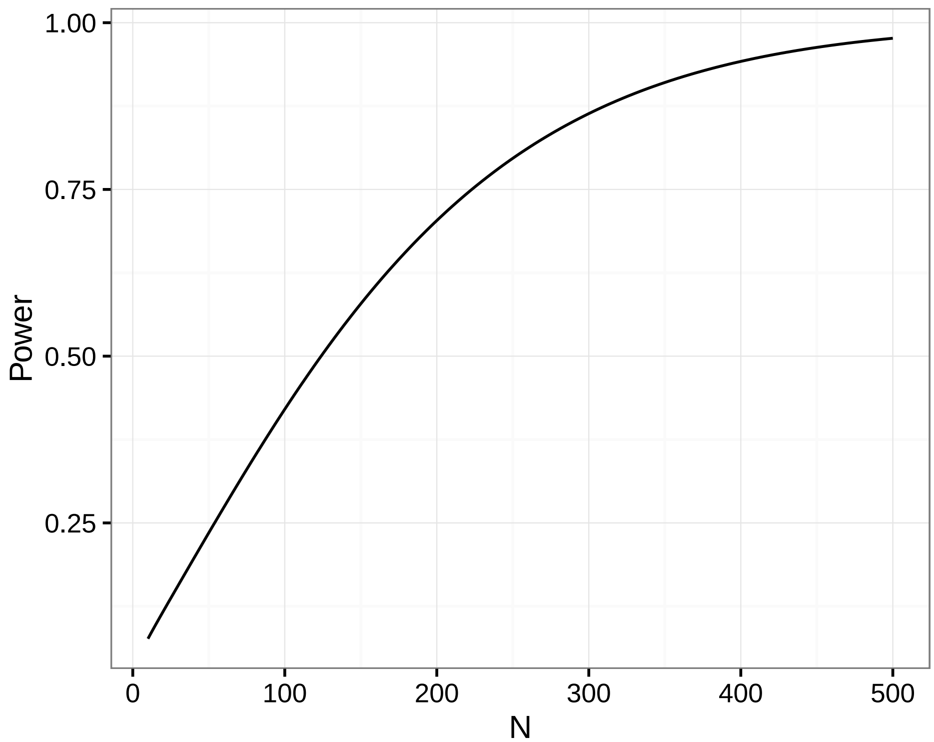

Figure 6. An example of a power function for sample sizes (N), assuming (for illustration) a standard deviation of 40 and a true effect of 10.

and was originally run with 40 participants and 16 items. As an example, we assume a *true* effect size of 0.01 for the log-transformed dependent variable, that is a difference of ≈4 ms from a grand mean (the mean reading time of all the data) of 550 ms. The magnitude of the effect may strike the reader as smaller than they'd expect in a psycholinguistic study; however, note that we do not know the true effect, and previously published studies may be giving us overestimates (due to Type S and M errors) if they have low power. The choice of a small magnitude of effect here is just to illustrate what happens when power is low.

For the simulation, we generated 1000 data sets, which we fit using linear mixed models with a log-transformed dependent variable (using the package *lme4*; Bates, Maechler, Bolker, & Walker, 2015). We used models with varying intercepts only for subjects and for items, in order to have the highest power, at the cost of anticonservativity (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2015); the discussion here does not depend on whether maximal models in the sense of Barr, Levy, Scheepers, and Tily (2013) are fit or not. The simulation shows the best-case scenario, since we generated the dependent variable (reading times) with a perfectly lognormal distribution and no outliers; Ratcliff (1993) shows that outliers can reduce power. The proportion of models under repeated sampling that show a significant effect (an absolute observed t-value greater than two) gives us the power, which is only 0.09. Type S error rate is 0.11 (calculated as the proportion of models with a significant effect, but with an estimated effect in the opposite direction of the true effect), and a Type M error of 5.08 (calculated as the mean ratio of the estimated effect to the true effect). This means that, in the present scenario, with a power of 0.09, we are likely to get an inflated estimate, and we have a 11 percent probability of having the wrong sign if we were to run the experiment repeatedly.

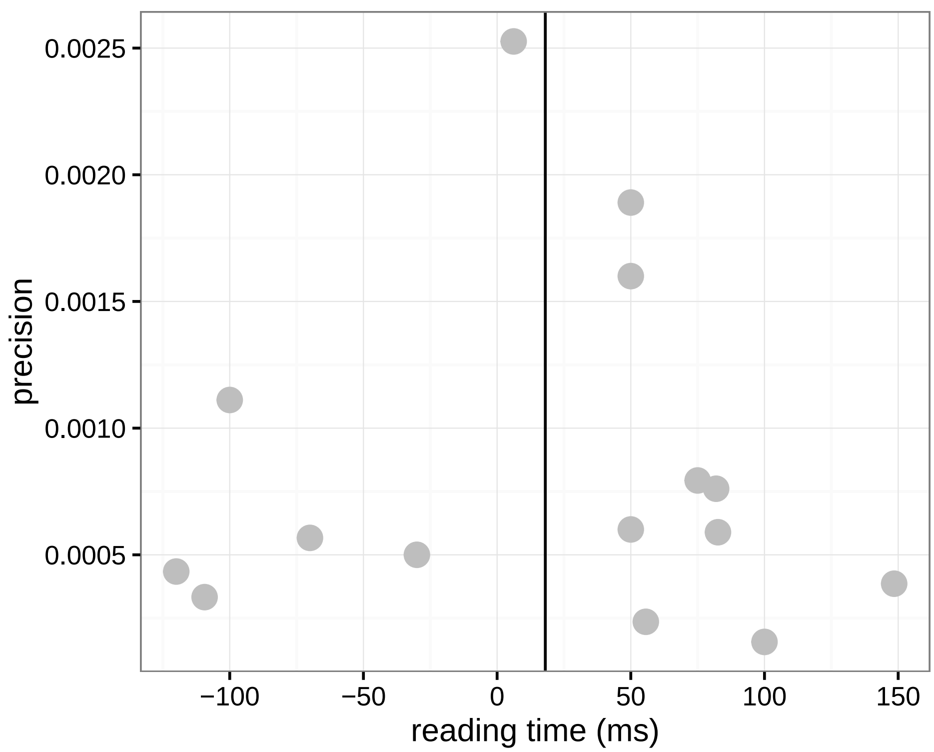

Figure 7. A funnel plot for 15 Chinese relative clause studies data, showing precision plotted against mean effects. The vertical line marks the mean effect of +18 msec (which represents a subject relative advantage).

Thus, in low power situations, Type S and M errors will be high; the practical consequence is that, if power is low, the magnitude and sign of the effect we find in our experiments may not be useful for calculating power in future experiments, or for modeling. This can be illustrated by considering the case of the Chinese relative clause controversy. There is a great deal of empirical disagreement regarding the question: in Chinese relative clauses, are subject relatives easier to process than object relatives (subject relative advantage)? Some researchers have claimed that object relatives are easier to process (at the head noun), and others have found the opposite, even in studies with the same design. As discussed in Vasishth 2015, we can see the Type S and M errors in action in this real-life example. We use the funnel plot (Duval & Tweedie, 2000) for this purpose. When we display the means of 15 published studies, we see that studies with lower precision (the inverse of the variance, $1/SE^2$) have a bigger spread about the grand mean of these means (Figure 7). Since extreme values will be further away from the mean and will influence the grand mean, an obvious consequence is that if we want accurate estimates of the true effect, we should run as high-powered experiments as we can.[4]

Of course, in many situations (e.g., when studying special populations such as individuals with aphasia), there are practical limits to the sample size. Indeed, one could also argue that it is unethical to run unnecessarily large-sample studies because this would be a waste of resources and maybe even a misuse of participants. In such restrictive contexts, the researcher may be better off using a Bayesian framework with (mildly) informative priors (see Part II of this review). Our point here is that there is no substitute for attempting to calculate power before running an experiment, using the best estimates one can obtain. Note also that it has been repeatedly pointed out by statisticians that it is a mistake to use 'observed' power, computed after the experiment has been run, to justify that one had adequate power for a study (Hoenig & Heisey, 2001). At the very least, observed power could lead one astray if the "true" power happens to be actually quite low; in such a case, Type S and M errors will lead to overestimates of power.

The next simulation illustrates the problem of low power by showing potential differences between estimates and various true effect sizes. We simulate the data of experiments

[4] As an aside, note that such funnel plots can also be used to identify publication bias: gaps in the funnel are suggestive of unpublished data that should have been observed under repeated sampling.

similar to the one presented before but with different values of true effect sizes (0.01, 0.02, 0.03, 0.05, 0.1) and in two flavors: a small sample experiment (but still publishable) with 30 subjects and 16 items, and a medium-sized experiment with 80 subjects and 40 items.

Figure 8 shows the results of this simulation. We simulated a total of ten (one small and one medium size experiment for each of the five effect sizes) sets of 200 experiments. Every point in the graph represents an experiment with a significant effect. The x-axis shows the true effect size on the log scale, while the y-axis shows the estimate of the effect from linear mixed models with significant results. The power of each simulation also appears in the figure. The simulation shows that exaggerated estimates (Type M errors) are more common for low-powered experiments. In addition, when the underlying effect is very small, some experiments will show results with the incorrect sign (Type S error). The dashed line shows the ideal situation where the estimate and the effects are the same.

This means that if power is very low, and if journals are publishing larger-than-true effect sizes—and we know that they have an incentive to do so because researchers mistakenly believe that lower p-values, i.e., bigger absolute t-values, give stronger evidence for the specific alternative hypothesis of interest—then power calculations based on published data are probably overestimating power, and thus also overestimating the replicability of our results. One can visualize the extent of overestimation of power for different scenarios by considering different true power levels. As shown in Figure 9, if true power happens to be low (e.g., if the true effect is much smaller than one expects), then the power estimate based on a pilot experiment's data could end up being an overestimate, due to Type M error.

To make matters worse, if we assume a normal distribution but the true distribution has a skew and possibly also occasional extreme values, this can also reduce power (Ratcliff, 1993), thereby increasing Type S and M errors. Latencies such as reading or response times are limited on the left by some amount of time and they are right-skewed; as a consequence, assuming, as is standardly done in linguistics and psychology, that the underlying distribution generating the data is normal, can lead to loss of power. Reading time or reac-

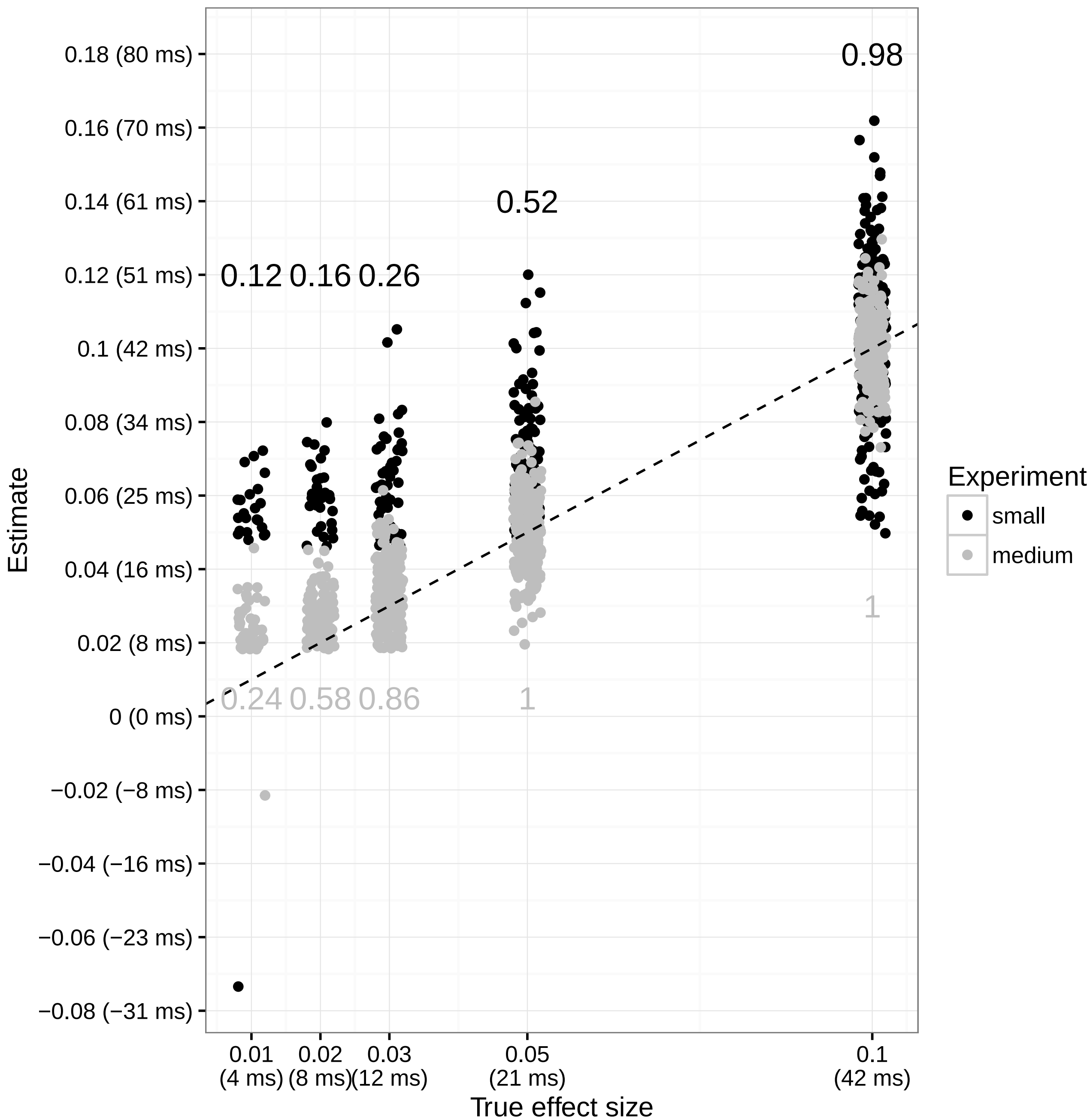

Figure 8. Plot illustrating the distribution of estimates under different true effect sizes and two experiment sizes. Each point represents an experiment with a significant effect, and for each effect size and experiment size, we simulated 200 experiments. The power is shown within the figure for the two sample sizes and for the different effect sizes.

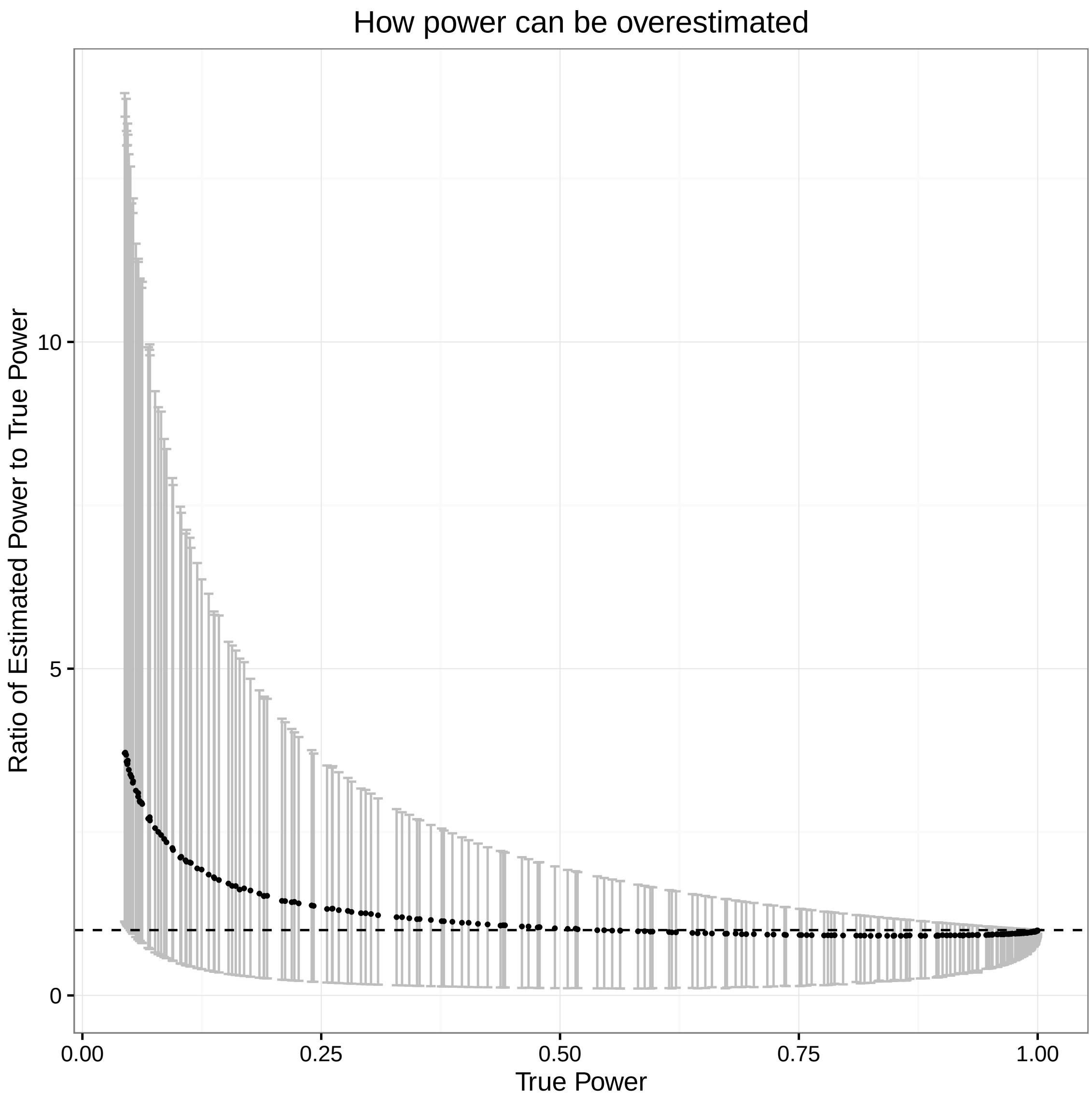

Figure 9. A simulation showing the effect of overestimation of power for different true power levels. Shown are the mean ratio of estimated power to true power; the error bars are 95% confidence intervals. For low true power, the estimates of power can be dramatic over-estimates: the ratio of estimated to true power can be quite large.

tion time distributions are best fit with three parameter distributions such as ex-Gaussian (the convolution of a Gaussian and an exponential distribution), shifted lognormal (the log-transformed normal distribution shifted to the right) or shifted Wald (the inverse of the normal distribution shifted to the right); see, for example, Luce 1986, Ratcliff 1993, and Rouder 2005. These distributions, however, are difficult to fit using standard frequentist software such the `lme4` function. A middle-ground solution is to apply a transformation on the dependent variable. This can reduce the impact of the skew (and of outliers) by compressing larger values to a greater extent than smaller values. The Box-Cox procedure (Box & Cox, 1964; Osborne, 2010) can be used to find out the best transformation; but, for reading times, we find that the reciprocal- (especially for self-paced reading) or the log-transformation (especially for eye-tracking measures) are adequate, and have the advantage of being easy to interpret. In sum, model assumptions matter. Even in the simple case of the one-sample t-test, violating the normality assumption leads to a fall in power.

A major practical implication of the Type M and S error problem is that if a study has low power, then it doesn't matter much whether you got a significant result or not. Theory development based on low power studies will have a very weak empirical basis, regardless of the p-values obtained. The main take-away point here is that we should run high powered studies, and attempt to replicate our results. There's really nothing as convincing as a replication. In our experience, reviewers and editors don't appreciate the importance of replications, but hopefully with increasing awareness of the issues, the culture will change.

Finally, Type S and M errors also relate to the point we have made earlier about interpreting p-values. Researchers sometimes think that when they get a low or very low p-value such as 0.00001 from a single experiment, the specific alternative is almost certain to be true. This would be the correct conclusion *if* power were high. But if power is low, we are at the mercy of Type S and M errors. This point is illustrated in Figure 10 shows that, indeed, a true effect and high power will almost guarantee a very low p-value. However, we don't know whether the null is true or not; if we already knew this, we wouldn't be running the experiment in the first place (and recall that it is non-trivial to determine that we are in a high-power situation). Notice that for point null hypotheses (and when the dependent measure is continuous), the distribution of p-values is uniform under the null hypothesis regardless of sample size (Hung, O'Neill, Bauer, & Kohne, 1997); this means that if the null is true, we are as likely to find a very low p-value as a very high one. So how do we know whether we have a very low p-value because (a) the null is false, or because (b) the null is in fact true, and therefore a very low p-value is as likely as any other value? From a single p-value we cannot know this. Therefore, a *single* p-value shouldn't give us much confidence on our theory.[5]

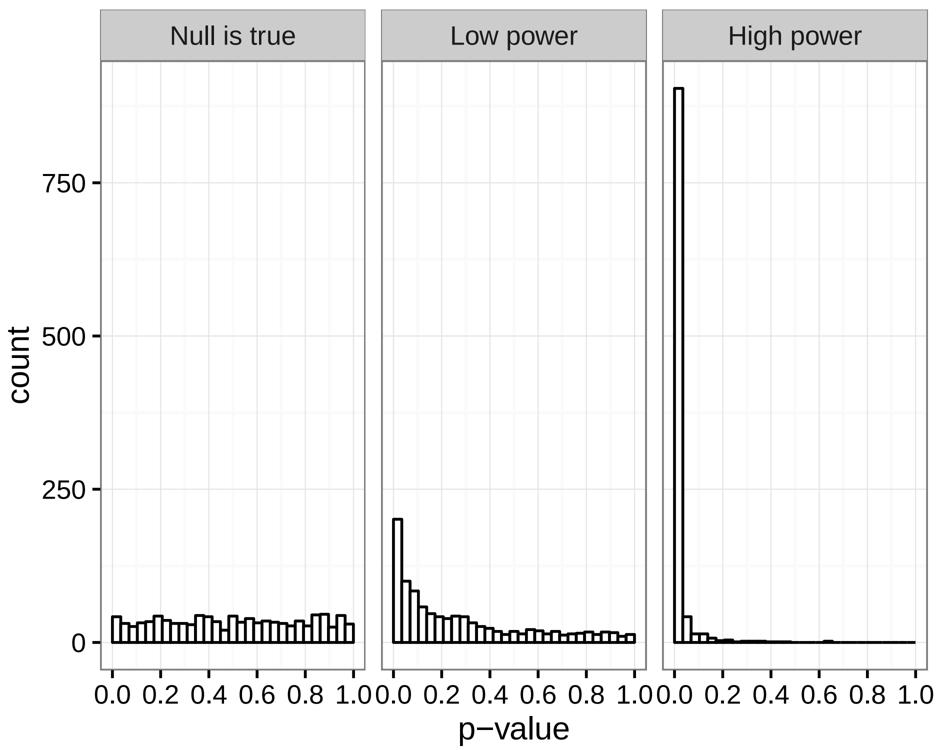

Figure 10. Simulated p-value distributions under a true $\mu = 0$ and under a true $\mu \neq 0$ with low and high power (0.25 and 0.93 respectively).

Since we don't know whether the null is true or not, a low p-value from a single experiment leaves us with no way to make a decision about the effect. One straightforward way to convince oneself whether an effect is present is to attempt a replication; repeatedly finding the same effect is much more convincing than any single hypothesis test.

[5] For more details on p-value distributions, see Hung et al. 1997.

Running till significance is reached

Given the importance that is attributed to significant findings regardless of sample size, replicability, multiple comparisons, and so forth, an interesting and troubling practice that we have encountered in linguistics and psycholinguistics is running an experiment until significance is reached. The experimenter gathers n data points, then checks for significance (whether $p < 0.05$ or not). If the result is not significant, he/she gets more data (n more data points) and checks for significance again. We can simulate data to get a feeling for the consequences of such a procedure. A typical initial n might be 15. This approach would give us a range of p-values under hypothetical repeated sampling. Theoretically, under the standard assumptions of frequentist methods, we expect a Type I error to be 0.05.

This is true if we fix the sample size in advance and repeatedly sample. But the statistical theory no longer holds if we adopt the strategy we outlined above: run some participants, test for significance, and if non-significant, run more participants. If we track the distribution of the t-statistic for this approach, we will find that Type I error is much higher (in our simulation, approximately 15%).

Let's compare the distribution, under repeated sampling, of the t-statistic in the standard case vs with the above stopping rule (gray line) (Figure 11). We get bumps in the tails with the above stopping rule because, under repeated sampling, some proportion of trials which have $p > 0.05$ will be replaced by trials in which $p < 0.05$, leading to a redistribution of the probability mass in the t-distribution. This redistribution happens because we give ourselves more opportunities to get the desired $p < 0.05$ under repeated sampling. In other words, we have a higher Type I error than 0.05. It would of course be reasonable to take this approach if we appropriately adjust the Type I error (Pocock, 1997); but this is unfortunately not standard practice. Thus, when using the standard frequentist theory, one should either adjust the Type I error when deploying stopping rules, or fix one's sample size in advance based on a power analysis.

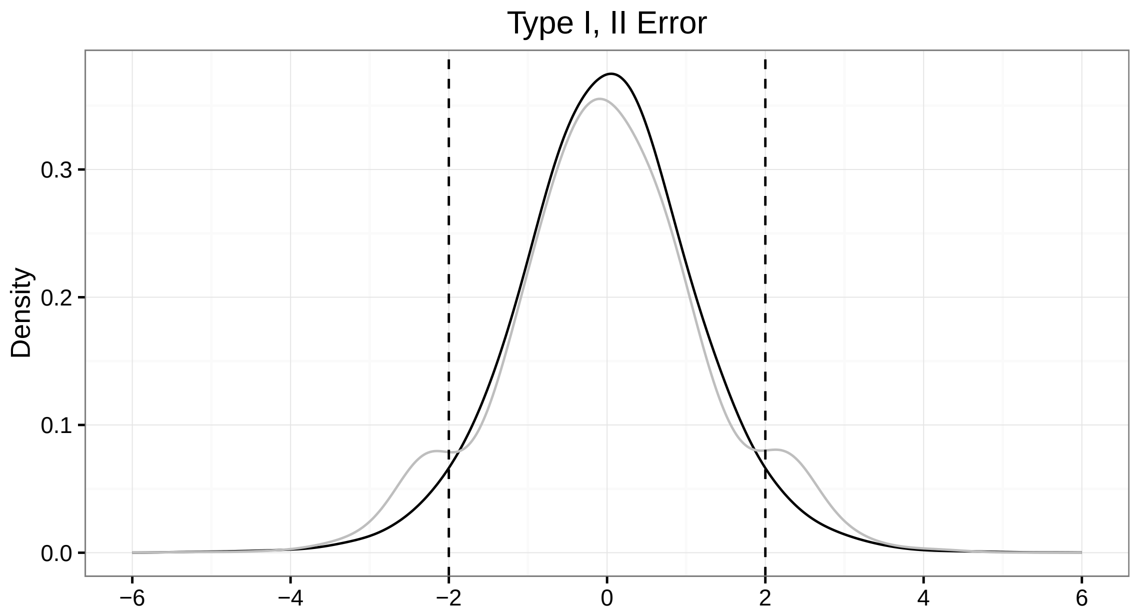

Figure 11. The distribution of observed t-values under repeated sampling using the stopping rule of run-till-significance. The dashed vertical lines mark the boundaries beyond which the p-value would be below 0.05.

Multiple measures, multiple regions, degrees of freedom in analysis

Even without a flexible stopping rule, we might be rejecting the null more than 5% of the time. This can happen due to researchers degrees of freedom (Simmons, Nelson, & Simonsohn, 2011) or the "garden of forking paths" (Gelman & Loken, 2013).

After the data have been collected from a planned experiment or an observational study, researchers usually have to make decisions such as: the choice of the statistical test, decisions on what constitutes an outlier, which covariates to include, whether the dependent variable should be transformed and which transformation to use, what the critical region in the sentences we are investigating, and so on. In many cases, it is impractical for researchers to make all these decisions beforehand: after gathering the data, for example, some models may not converge, and may need to be simplified (Barr et al., 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek et al., 2015).

In principle, none of this is problematic: fitting many models to a data-set as part of a model selection process is a central component of the data analysis process and is part of standard statistical practice. As Gelman and Hill (2007) put it (p. 513): "There are generally many options available when modeling a data structure, and once we have successfully fit a model, it is important to check its fit to data. It is also often necessary to compare the fits of different models."

The problem arises when researchers present models with statistically significant results (or ones without, depending on what the theoretical claim is). It is common for researchers to explore various alternatives and then reason that the one that produces a significant result is more likely to be the most reliable. The main problem arises when the researcher bases inferences only on the significant result, ignoring the alternative analyses (Humphreys, de la Sierra, & Van der Windt, 2013). Gelman and Loken (2013) point out that since there are in general so many ways to analyze the data, once we start looking hard enough, it is possible to find a significant difference and then tell a post-hoc theory that fits very neatly together. For that to happen, it is not necessary that a researcher would go on a "fishing expedition" (Gelman, 2013), that is, it is not necessary that he/she would be actively trying to report any comparisons that happen to yield a *p*-value lower than 0.05; sometimes it is just not clear what the right way is to analyze the data.

We discuss here four common types of situations where the focus on statistical significance can lead to difficulties:

1. Variance decreasing artificially due to aggregation.
2. Presenting post-hoc discoveries as predicted effects.
3. Multiple analyses for several regions of interest and several dependent measures.
4. Different potential analyses.

Variance decreases artificially due to aggregation

Given a choice between statistical tests, researchers have an incentive to choose the test that yields a significant result over one that doesn't. For example, as discussed by Barr et al. (2013), when a linear mixed model with crossed subject and item random effects is suggested by the design, using an ANOVA or t-test can artificially reduce the sources of variance due to aggregation, with the result that effects that are not statistically significant under a linear mixed model end up being significant once one aggregates the data. The LMM is useful in linguistic and psychology experiments precisely because it can take all sources of variance into account simultaneously (for further discussion, see Baayen, Davidson, & Bates, 2008; Barr et al., 2013; Bates, Kliegl, et al., 2015; Matuschek et al., 2015).

Presenting post-hoc discoveries as predicted effects

Other situations increase the degrees of freedom of the researcher in such a way that the results are clearly exploratory, but are presented as if they confirm the hypothesis that was proposed before the experiment was run. This is not uncommon especially in reading studies, where it is possible to flexibly analyze different regions of interest (often aggregating different regions post-hoc) until a $p < 0.05$ result is found. For example, in Badecker & Straub, 2002, the region of interest was the four words following the originally targeted region (the reciprocal), and this was the only combination of positions that yielded a statistically significant result. In their results section for experiment 5, Badecker and Straub write: "No significant differences emerged for any individual words or two-word regions. However when reading times are collapsed across the four positions following the reciprocal, reading times for this region were 48 ms longer in the multiple-match than in the single-match condition…". This finding might well reflect a result of theoretical interest and may be reproducible; we know that effects often spill over to subsequent regions. It is, however, a fairly typical example of failing to find an effect in the region where it was expected a priori, and then expanding the regions of interest post-hoc until an effect is found.

Multiple analyses for several regions of interest and several dependent measures

As mentioned above, one problem that the researcher faces is that in methods like self-paced reading and eye-tracking, it is often not clear if the effect should appear at only one specific word or several, or even what the critical region is. Effects can indeed be delayed and appear in spillover regions. However, even though this is rarely acknowledged, fitting models (t-test, ANOVA, linear mixed model, etc.) for each possible word where the effect may appear raises the problem of multiple comparisons, increasing the chances of finding a false positive.

In addition to the multiple regions problem, eye-tracking-while-reading raises the problem that multiple measures which are highly correlated (first fixation duration, single fixation duration, gaze time, etc.) are routinely analyzed as if they were separate sources of information (von der Malsburg & Angele, 2015). The motivation for analyzing multiples measures is that the theories being tested usually do not make explicit predictions about which measure is relevant. For EEG, even though it is not always the case, without a clear prediction about the ERP component, this issue is even more serious: it is possible to look for effects in too many different groups of electrodes (which are correlated) and in different time windows (Frank, Otten, Galli, & Vigliocco, 2015).

Different potential analyses

The problem is not only the analyses that have been explored, but the ones that could have been done if the data would have been different (Gelman & Loken, 2013). This may sound strange, but we can illustrate it with a very simple example. Imagine a researcher, R, that collects data from a self-paced reading experiment. R may decide that observations below 150ms are very likely to be outliers even before looking at the data. R removes them and fits an ANOVA. The p-value is lower than 0.05, and thus R writes a paper reporting the results. The problem, according to Gelman and Loken (2013), is that R may have behaved differently had he not found a p-value lower than 0.05. R would have looked at the data more closely and he would have removed perhaps also slower outliers, or he would have removed also observations that are below 200ms.

Some proposals in the literature for resolving the garden-of-forking-paths problem

It is necessary to recognize here that model fitting is as much science as art, and a one-size-fits-all approach can be misleading. It is clear that in most cases researchers are not doing tests until they find a comparison showing a significant difference. But in many novel experiments, it is just not possible to know ahead of time where an effect will appear, what statistical model will converge, what will be the right transformation of the dependent variable, and so forth. However, the problem of researcher degrees of freedom in the analyses may hinder replicability and give a false feeling of certainty regarding the results.

There are ways to mitigate these problems. One interesting idea is that the researcher may analyze the data "blindly" (Nuzzo, 2015). This requires that the researcher would write a program that creates an alternative dataset by, for example, adding random noise or moving observations to different experimental conditions. The researcher should handle the fake data set as usual, and when he or she is satisfied with the analysis, the analysis should be applied to the real data.

Second, Simmons et al. (2011) provide some guidelines for both authors and reviewers. Their guidelines for authors emphasize transparency: authors should report failed manipulations, and whether the results depend on the removal of outliers; releasing the data and the analysis code would also contribute greatly to transparency. Simmons et al.'s guidelines for reviewers emphasize that reviewers should be more tolerant of imperfections in results, but should require authors to show that their results do not depend on arbitrary decisions.

Third, linear mixed models can solve some of the multiple comparisons problems if all relevant research questions can be represented as parameters in one coherent hierarchical model (Gelman, Hill, & Yajima, 2012), since the point estimates and their corresponding intervals are shifted toward each other via "shrinkage" or "partial pooling" (Gelman et al., 2012). However, building a single hierarchical model that addresses all the research questions is not always trivial. For several regions of interest, it may be possible to fit a single model using Helmert contrasts (as in Nicenboim, Logačev, Gattei, & Vasishth, 2015). This type of contrast compares each region with the average of the previous ones, such that it is possible to discover a change in the pattern of the effects. However, it is unclear if the effects should appear for all the trials in the same region, since some participants in some trials could start a certain process sooner predicting the structure of the item or could delay it due to fatigue, lapse of attention, or because they have not finished a previous cognitive process. Another solution proposed recently by von der Malsburg and Angele (2015) is to fit independent models but to apply some type of correction such as the Bonferroni correction.

Finally, when new data can be easily gathered, an attractive solution is to take results as exploratory until being confirmed with new data (De Groot, 1956,2014; Tukey, 1977). The exploratory data is used to identify the relevant regions, measures and/or ERP components, and to make decisions about the model and outliers. Only these potential effects are then tested on the confirmatory analysis. Researchers could pair each new experiment with a preregistered replication (Nosek, Spies, & Motyl, 2012), or gather more data than usual so that the full data set could be divided into two subsets (for an example with EEG data, see Frank et al., 2015).

The broader point here, stressed in Gelman and Loken (2013), is that researchers should be more aware of the choices involved in the data analysis. If the exploratory analysis is clearly separated from the confirmatory one, researchers can still be creative about exploring patterns in the data. It is relatively easy to separate the primary analyses from the exploratory ones; see Safavi, Husain, and Vasishth (2016) for a specific example.

Conclusion

In this article, we attempted to summarize some well-known pitfalls associated with using the frequentist data-analytic approach. The frequentist approach has a very important role to play in statistical inference but it should be used as intended. Given the growing realization that many claims in the scientific literature may be false (Ioannidis, 2005), and the recent failures to replicate many of the published results (Open Science Collaboration, 2015), now is as good a time as any to take stock and carefully consider how we use these tools. Checking model assumptions, running well-powered studies to the best of one's ability, avoiding exploratory analyses under the guise of hypothesis testing, and always attempting to replicate one's own and others' results; these are steps that can be realized in linguistics. It is true that replication of one's own and others' data may be difficult in some cases, and running high power studies can also be a challenge when, for example, one is working with special populations. Regardless of these and other barriers, the overall goal should be transparency. One way to achieve complete transparency is to release all data and code associated with a published paper, so that the reader can retrace the authors' steps.

Finally, in situations where data is sparse and/or the researcher needs to go beyond standard models, Bayesian data-analysis methods should be considered (Gelman et al., 2014); we review this approach in Part II.

Acknowledgements

We would like to thank the anonymous reviewers, Daniela Mertzen, Dario Paape, and Titus von der Malsburg for helpful comments on previous versions of this paper.

References

Arunachalam, S. (2013). Experimental methods for linguists. *Language and Linguistics Compass*, *7*(4), 221–232.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 748-769.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models.* Retrieved from `http://arxiv.org/abs/1506.04967` (ArXiv e-print)

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software.* (In Press)

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

De Groot, A. (1956,2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta psychologica*, *148*, 188–194.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC Press.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2016). *The determinants of retrieval interference in dependency resolution: Review and computational modeling.* (Manuscript submitted)

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015, Jan). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. Retrieved from `http://dx.doi.org/10.1016/j.bandl.2014.10.006` doi: 10.1016/j.bandl.2014.10.006

Gelman, A. (2013). *Too good to be true. Slate, 24 Jul.*

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third ed.). Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012, Apr). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. Retrieved from `http://dx.doi.org/10.1080/19345747.2011.618213` doi: 10.1080/19345747.2011.618213

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time.

Gibson, E., & Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155.

Gries, S. (2009). *Quantitative corpus linguistics with R: A practical introduction.* New York–London: Routledge.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55(1)*, 19–24.

Humphreys, M., de la Sierra, R. S., & Van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, *21*(1), 1–20.

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, *53*, 11–22.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, *9*(7), 1–14.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), 40–47. doi: 10.1371/journal.pmed.0020124

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., & Vasishth, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. *Frontiers in Psychology*, *6*(506). Retrieved from `http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00506/abstract` doi: 10.3389/fpsyg.2015.00506

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111(2)*, 228–238.

Luce, R. D. (1986). *Response times* (No. 8). Oxford University Press.

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). *A meta-analysis of syntactic priming.* (Unpublished draft)

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2015). Balancing Type I error and power in linear mixed models. *arXiv preprint arXiv:1511.01864*.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2015). *When high-capacity readers slow down and low-capacity readers speed up: Working memory differences in unbounded dependencies.* (Resubmitted)

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature neuroscience*, *14*(9), 1105–1107.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature News*, *526*, 182–185.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). Retrieved from `http://www.sciencemag.org/content/349/6251/aac4716.abstract` doi: 10.1126/science.aac4716

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, *15*(12), 1–9.

Phillips, C., Wagers, M. W., & Lau, E. F. (2009). Grammatical illusions and selective fallibility in real time language comprehension. *Language and Linguistics Compass*.

Pocock, S. J. (1997). *Clinical trials: A practical approach.* John Wiley & Sons.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510.

Rouder, J. N. (2005, Jun). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377-381. Retrieved from http://dx.doi.org/10.1007/s11336-005-1297-7 doi: 10.1007/s11336-005-1297-7

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in persian separable complex predicates: Implications for expectation and memory-based accounts. *Frontiers in Psychology*, *7*. (Special Issue on Encoding and Navigating Linguistic Representations in Memory) doi: 10.3389/fpsyg.2016.00403

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2015). *Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists.* Retrieved from http://arxiv.org/abs/1506.06201 (ArXiv e-print)

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*, *19*(3), 193–198.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Vasishth, S. (2015). *A meta-analysis of relative clause processing in Mandarin Chinese using bias modelling.* Sheffield, UK. Retrieved from http://www.ling.uni-potsdam.de/~vasishth/pdfs/VasishthMScStatistics.pdf

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013, 10). Processing Chinese relative clauses:

Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), 1–14.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767-794.

von der Malsburg, T., & Angele, B. (2015). False positive rates in standard analyses of eye movements in reading. *arXiv preprint arXiv:1504.06896*.

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, *90*(3), 669–692.