Science Practice Hub Workshop | Frequentist statistical analysis: foundational ideas and common mistakes

Please mute your microphone to prevent feedback.

Turn on live captions to follow along as the presenter speaks if you like.

To ask a question please use the chat box or the raise your hand function.

This talk will be recorded. Please comment in the chat box if you do not consent to being recorded.

Frequentist statistical analysis: foundational ideas and common mistakes

DR MALGORZATA WOJTYS

18/11/2020

WORKSHOP IN QUANTITATIVE ANALYSIS

Who I am...

- Lecturer in Statistics at the University of Plymouth, since 2013
- PhD in Mathematical Statistics, Warsaw University of Technology (2011)
- Research interests:
  - Model selection criteria
  - Copula modelling
  - Flexible estimation using splines
  - Sample selection models
- Teaching:
  - Statistical inference for undergraduate maths students
  - Machine learning for MSc Data Science students

Contents of the talk

- Concepts around hypothesis testing
  - P-value
  - Type I and Type II errors
  - Power
- Common mistakes:
  - Not checking statistical assumptions
  - Small sample sizes
  - Misinterpreting p-values

Introduction

- Easy to obtain results from statistical packages such as R, SPSS, etc.
- Equally easy to misinterpret them…
- … leading to conclusions not supported by statistical theory.
- Growing realization that many claims in the scientific literature may be false (Ioannidis, 2005).
- Failures to replicate many of the published results (Open Science Collaboration, 2015).

Example 1: reading experiment

- Two-condition repeated measures self-paced reading experiment
- subject versus object relative clauses
- the dependent measure is reading time in milliseconds
- Question: Do the population means of reading time between two RC types differ?
- reading time is measured at a particular region of interest in the relative clause sentences
- *n* randomly sampled participants, each of whom read multiple instances of subject and object relative clauses in a counterbalanced Latin square design.

Example 1: reading experiment

Participant id	Item id	Condition	Reading time (ms)
1	1	SR	505
1	2	SR	601
1	3	SR	710
1	4	OR	452
1	5	OR	550
1	6	OR	640

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

Example 1: reading experiment

- For each participant: mean reading time for SR and for OR
- Data:

$$x_1, x_2, \dots, x_n$$

where x_i = the difference between mean reading time for SR and for OR for ith participant, $i = 1,2, \dots, n$.

- Assumptions:
  - each x_i comes from the same distribution with population mean μ and std deviation σ,
  - all x_i's are mutually independent
- μ represents the true, unknown difference in means between the two RC types.

Example 1: reading experiment

- The hypotheses:

$$H_0: \mu = 0$$
$$H_1: \mu \neq 0$$

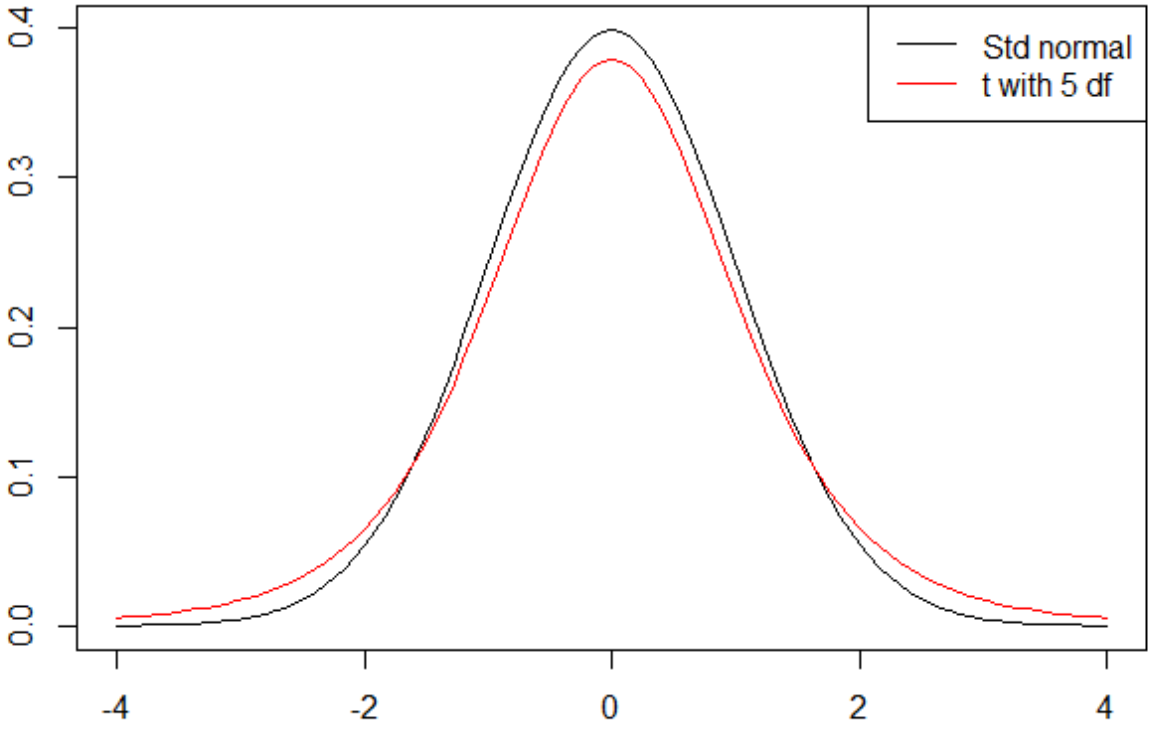
- The test statistic:

$$T = \frac{\bar{x}}{SE(\bar{x})}$$

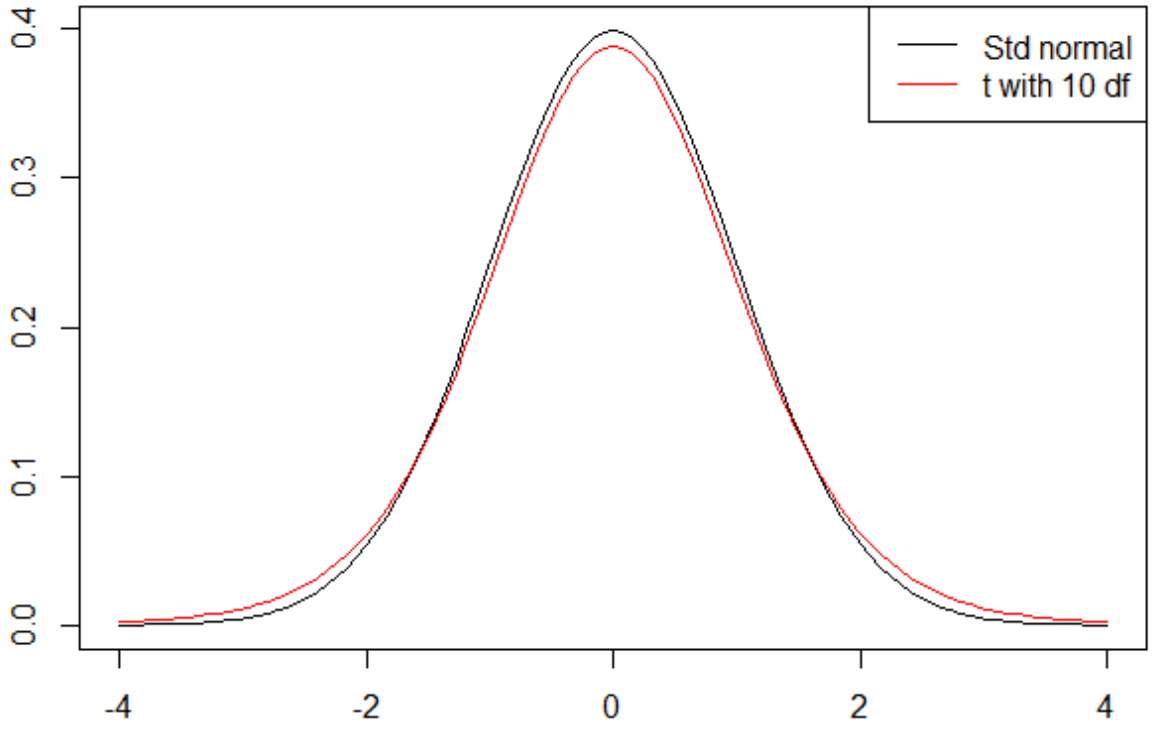
where $SE(\bar{x}) = {}^{S}/_{\sqrt{n}}$ and S is the sample standard deviation of $x_1, x_2, \ldots, x_n$.

- The distribution of T under H_0:
  - If $x_1, x_2, \ldots, x_n$ are normally distributed then $T \sim t_{n-1}$ for any n.
  - If n is large then $T \sim N(0,1)$ approximately.

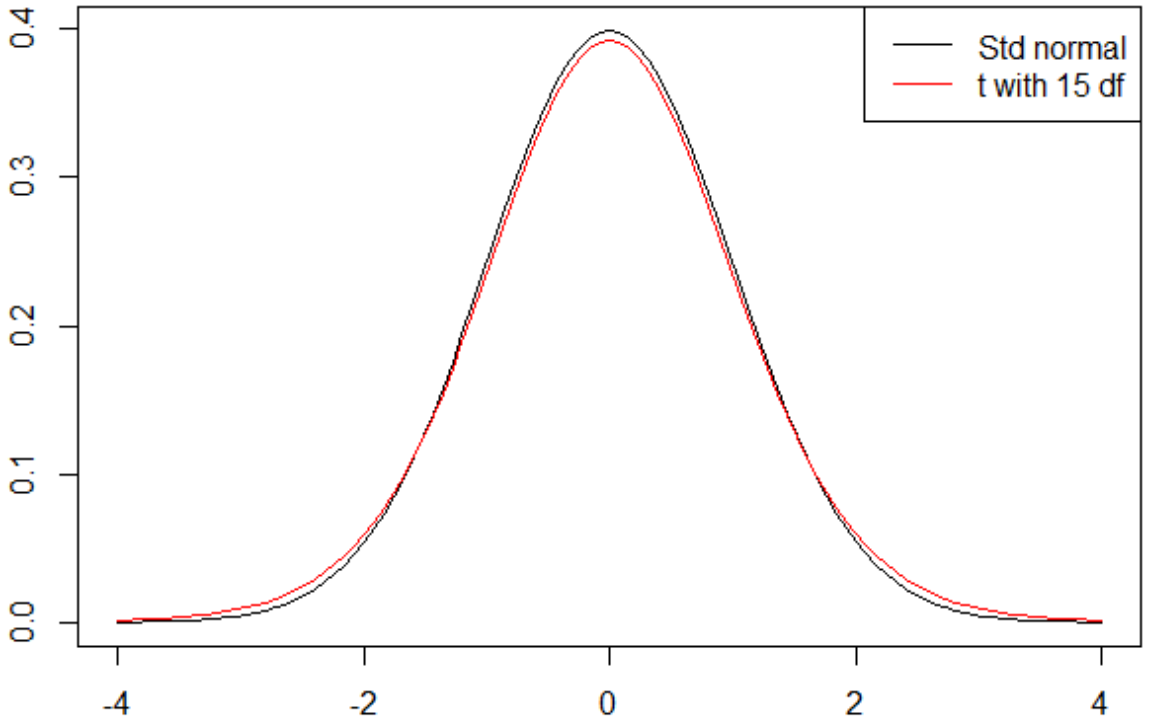
T versus Normal distribution



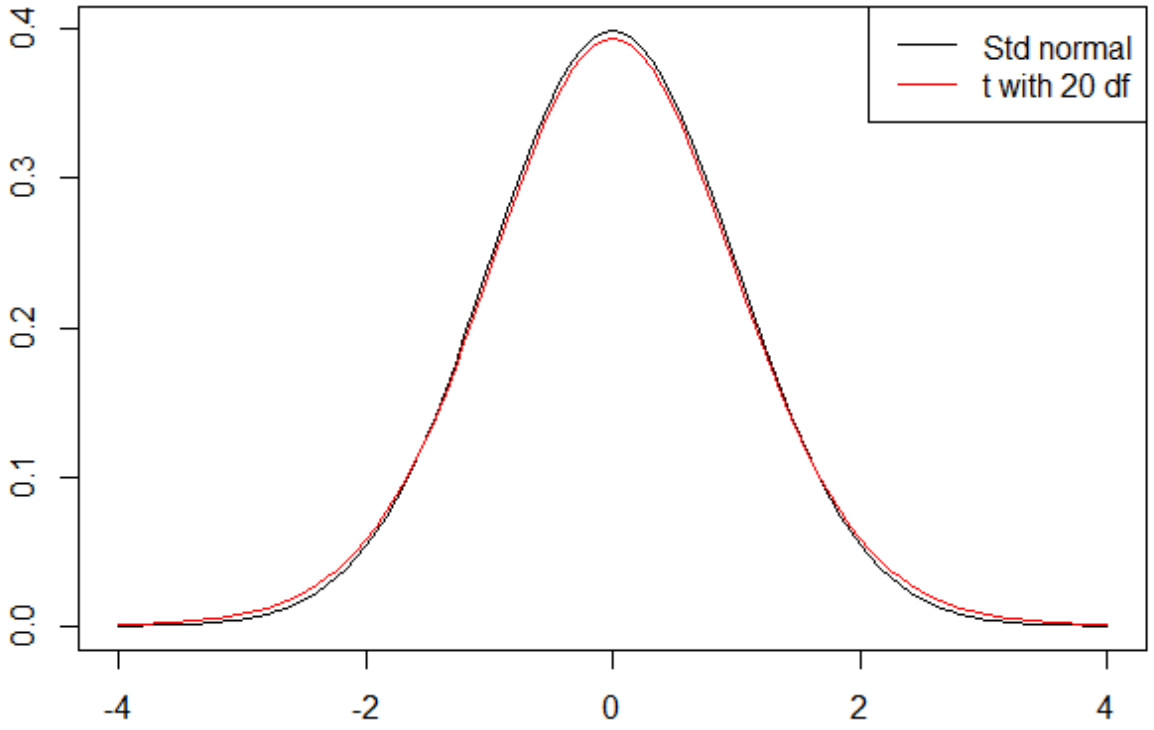
T versus Normal distribution



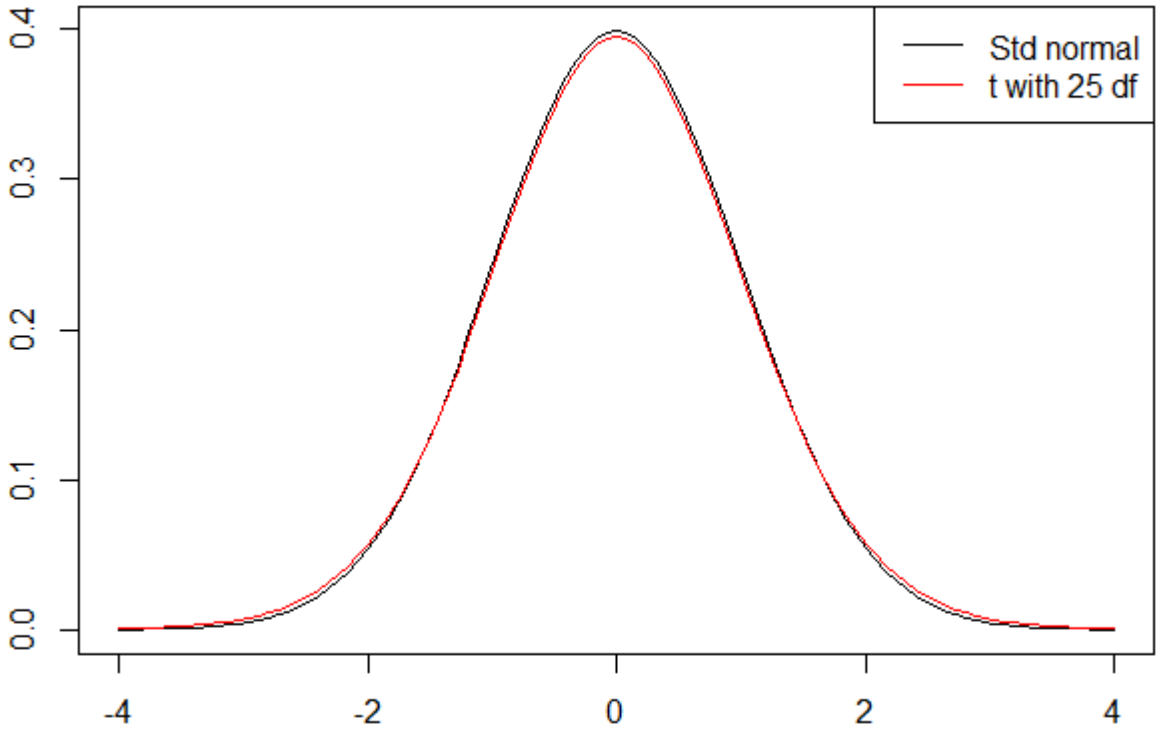
T versus Normal distribution



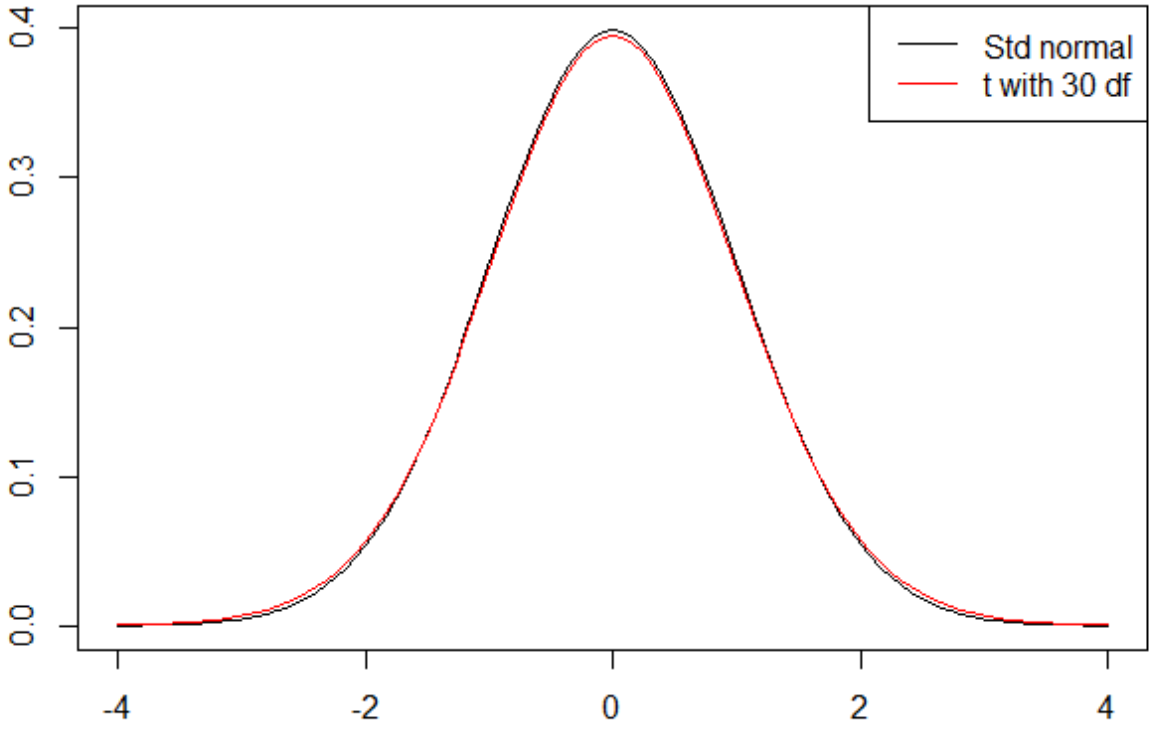
T versus Normal distribution



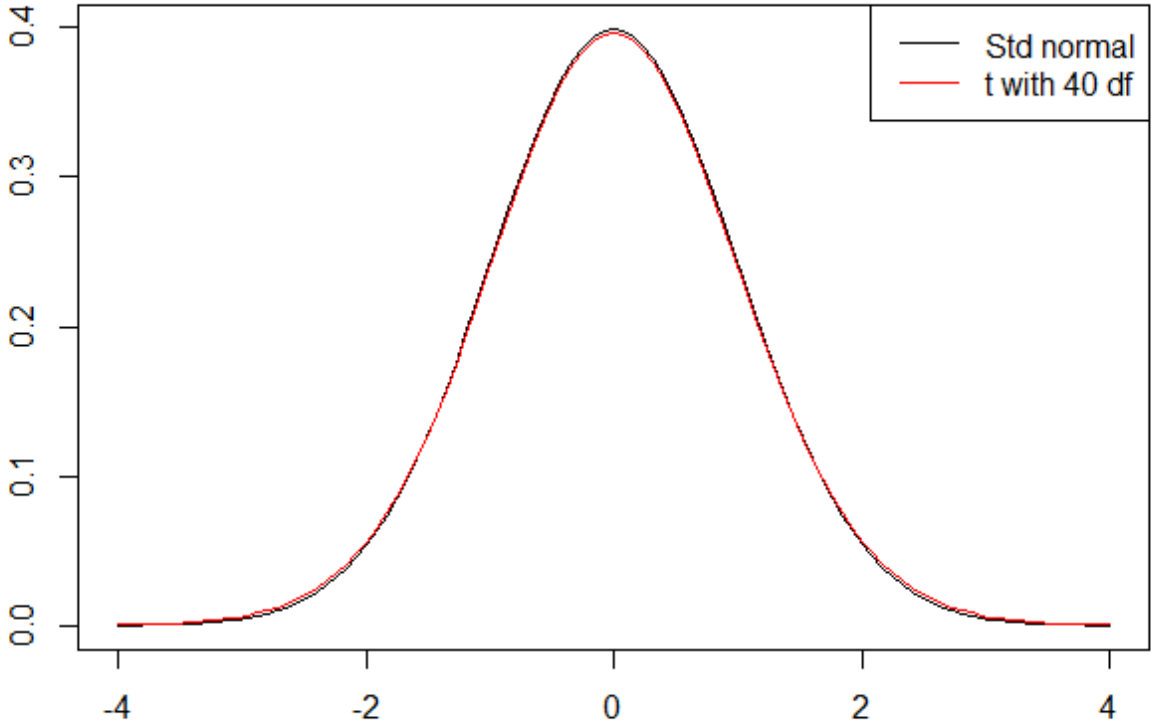
T versus Normal distribution



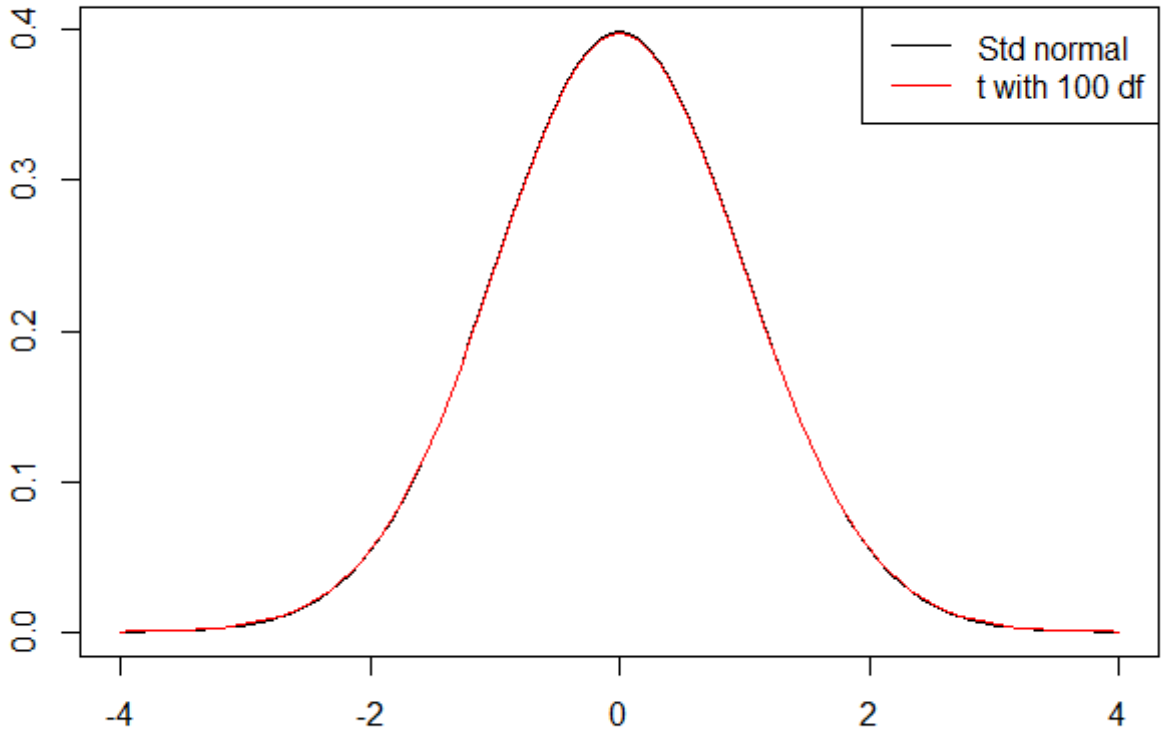
T versus Normal distribution



T versus Normal distribution



T versus Normal distribution

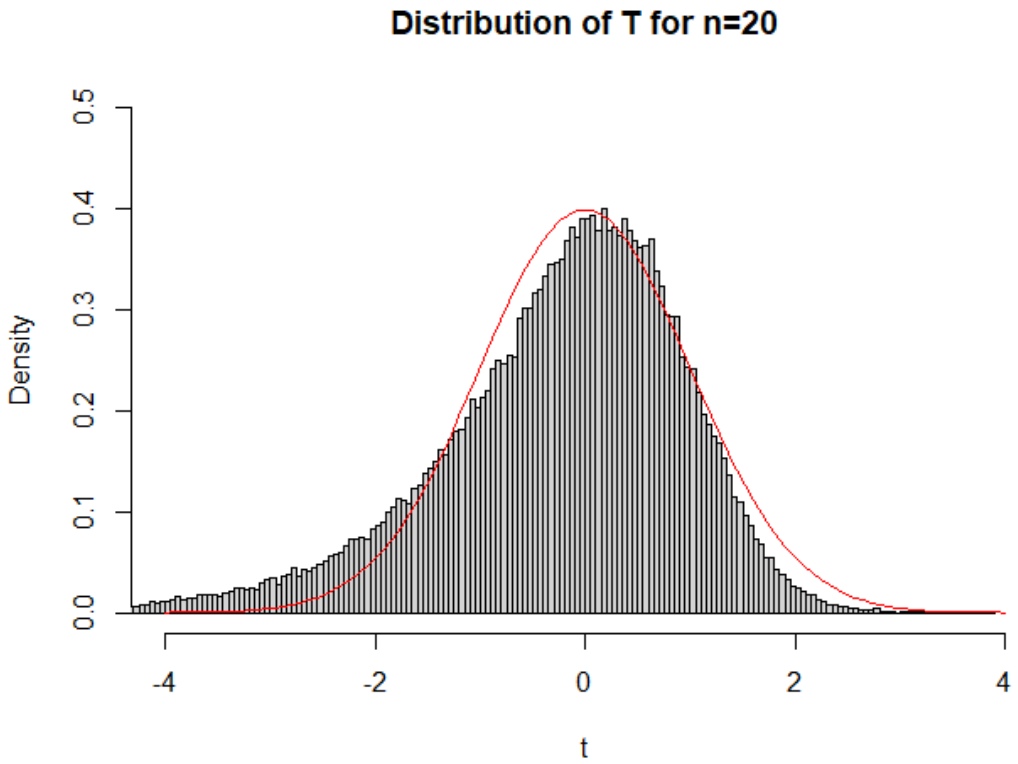


T versus Normal distribution

- Roughly speaking, for n>50 the two distributions are practically the same…
- For large sample size we may use N(0,1) distribution in place of t-distribution.
- For large sample size the Central Limit Theorem holds: the sample mean follows a normal distribution regardless of the distribution of data.
- So, we do not necessarily need normally distributed data when we deal with sample means and sample size is large.

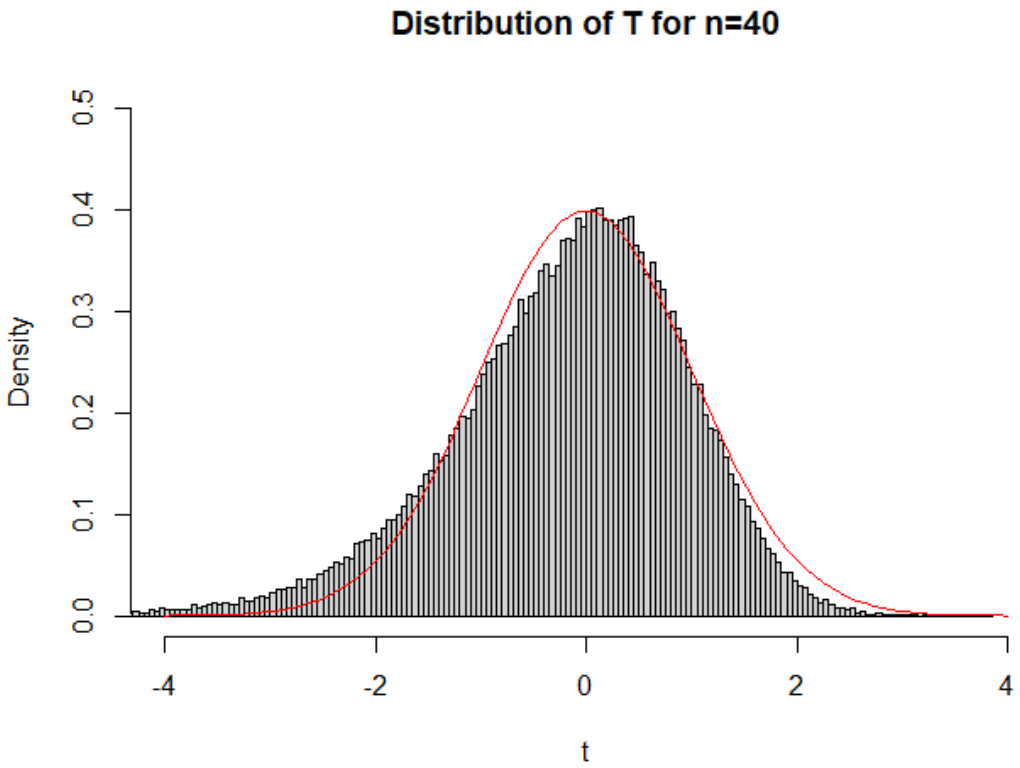
Distribution of T for skewed data

- Data: $x_1, x_2, \ldots, x_{20} \sim Exp(1)$
- $T = {}^{\bar{x}-1}/_{S/\sqrt{n}}$



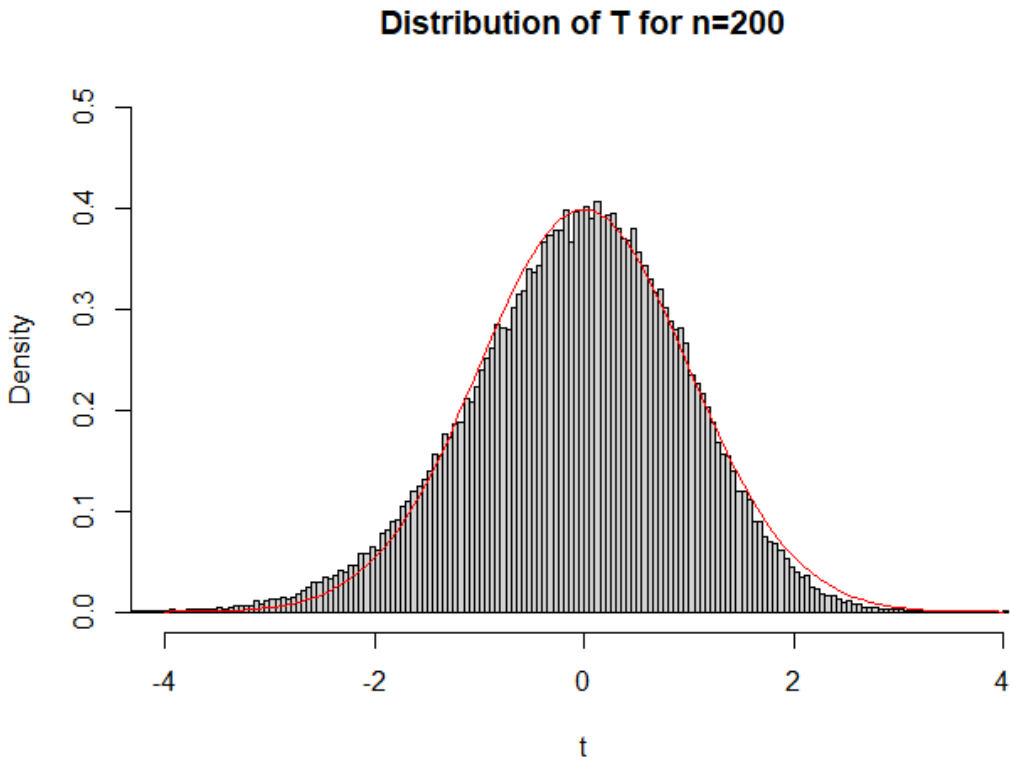
Distribution of T for skewed data

- Data: $x_1, x_2, \ldots, x_{40} \sim Exp(1)$
- $T = {}^{\bar{x}-1}/_{S/\sqrt{n}}$



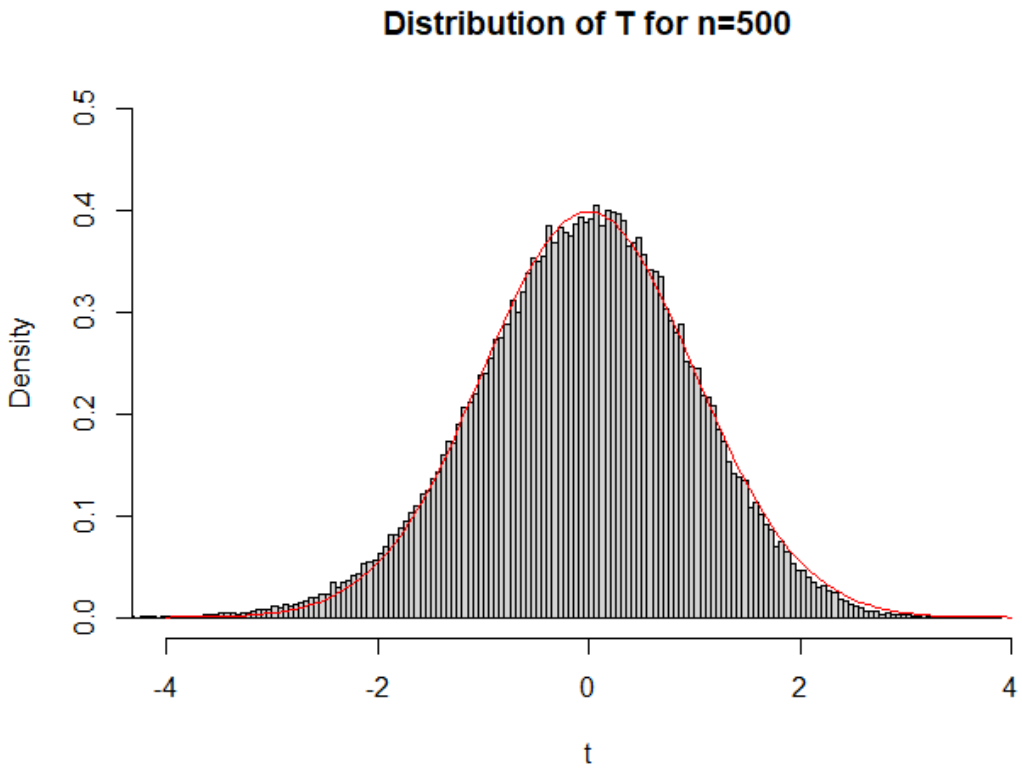
Distribution of T for skewed data

- Data: $x_1, x_2, \ldots, x_{200} \sim Exp(1)$
- $T = {}^{\bar{x}-1}/_{S/\sqrt{n}}$



Distribution of T for skewed data

- Data: $x_1, x_2, \ldots, x_{500} \sim Exp(1)$
- $T = {}^{\bar{x}-1}/_{S/\sqrt{n}}$



Recommendations on assumptions-checking

- Kolmogorov-Smirnov test, Shapiro-Wilk test for normality
- Levene's test for normality
- Visualisations / graphs! They convey more information than tests…
  - Histograms, QQ-plots, scatterplots.
- Some remedies:
  - Small data / non-normality: bootstrapping methods;
  - Use robust statistics (Larson-Hall, 2016),
  - Bayesian methods with mildly informative priors.

Checking statistical assumptions

- *"If theoretical insights and pedagogical recommendations are to be trusted, they must come as the results of the accurate use of appropriate methods"* Loewen et al (2014; 379)

- Study quality – the adherence to standards of rigor and transparency.

- Assumptions checked and reported – indicator of transparency and quality.

Most popular statistical procedures

- One-way ANOVA
- Multiple regression
- Independent samples t-test
- Chi-square test

One-way ANOVA

- Equal variances
- Normality
- Independent errors
- Outliers

Multiple regression

- Linearity
- Normality
- Homoscedasticity
- Independent errors
- Multicollinearity
- Outliers
- Sample size

Independent samples t-test

- Normality
- Equal variances
- Independence

Chi-square test

- Minimum expected frequencies
- Independent variables

Reporting of assumptions-checking

The percentage of studies that explicitly mentioned that assumptions were checked:

- **3%** Plonsky and Gass (2011): 174 published papers on L2 interaction across 14 journals and two edited volumes.
- **17%** Plonsky (2013): 606 quantitative studies published from 1990–2010 in two L2 journals.
- **16%** Liu and Brown (2015) A methodological synthesis of research on corrective feedback on L2 writing.
- **22%** Lindstromberg (2016): 96 quasi-experimental studies in 90 articles, published from 1997 to 2015 in *Language Teaching Research*.
- **15.53%** (normality) Plonsky & Ghanbar (2018): methodological synthesis of multiple regression in L2 research
- **74.7%** Al-Hoorie and Vitta (2018): 150 articles published in 2016 or later from 30 applied linguistics journals
- **17%** (stringent), **24%** (lenient) Hu and Plonsky (2019): 107 studies in two L2 journals over 2012 – 2017.
- **17%** Sesé and Palmer (2012): 623 articles published in eight psychology journals

Reporting of assumptions-checking

- Especially low for t-test.
- The highest for regression.
- Normality is the most frequently checked assumption for parametric tests.

Reporting of assumptions-checking

Causes of the lack of reporting of assumptions-checking:

- No explicit mention of checking assumptions does not necessarily mean that researchers did not check assumptions:

  self-indicated frequency of checking assumptions: 60% to 80% (Plonsky et al. (2017) survey on 364 applied linguists)

- Word limit / space limitations

- Lack of knowledge / statistical literacy:

  14% of PhD students and 30% of professors in a survey of 331 applied linguists felt that they had received adequate statistical training (Loewen et al., 2014)

Recommendations on assumptions-checking

- Do not just report that assumptions were checked – state how they were checked.
- Produce online supplementary material.
- Make the raw data set publicly available (e.g. IRIS database).
- Make the analysis code available.

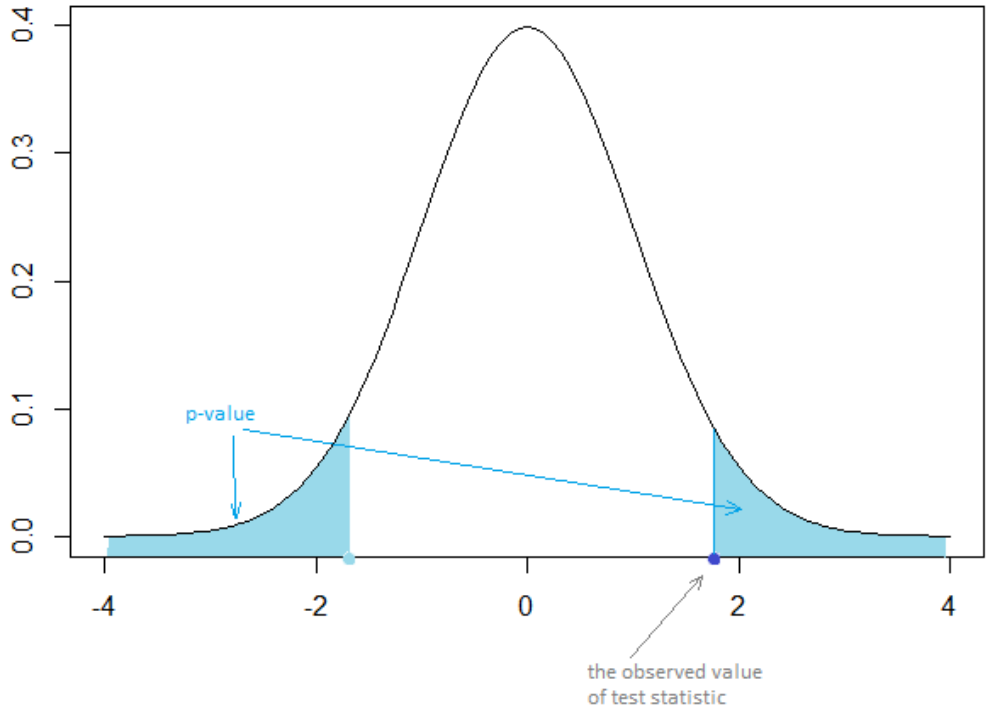
P-value

- p-value associated with the observed t-value:

  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.

- Reject the null hypothesis if this conditional probability falls below α (usually 0.05).

- P-value is a measure of evidence against H_0.



Misconceptions about p-value - 1

A p-value greater than 0.05 tells us that the null hypothesis is true.

- Suppose that $t = 0.8$. Then p-value=0.424.
- Can we say we are confident that μ =0?
- Another common example: "*There is no effect of factor X on dependent variable Y*", based on a p-value larger than 0.05 in ANOVA.
- This claim can only be made when power is high.
- We should say: we failed to find evidence against H_0 (we failed to find an effect in ANOVA).

Misconceptions about p-value - 2

The smaller the p-value, the greater the confidence in the *specific* alternative hypothesis.

- Suppose that t =4.29 with $\bar{x} = 3$ and SE=0.7. Then p-value=0.000018.
- Can we say we are confident that μ =3?
- 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 $\mu = 0$.

Misconceptions about p-value - 3

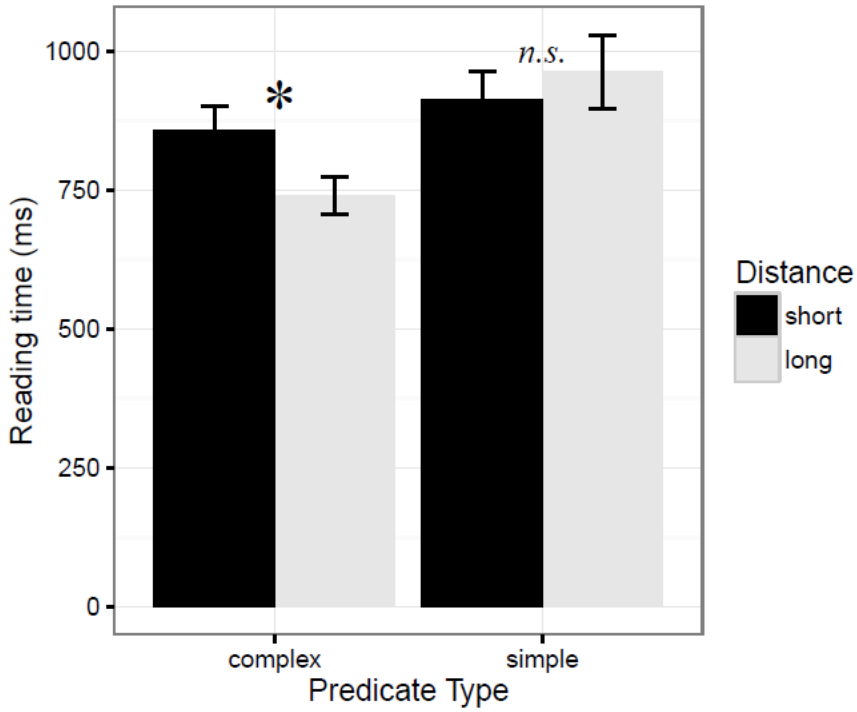
If p-value < 0.05 we have found that H_1 is in fact true.

- No absolute certainty is afforded by the p-value, no matter how low it is.
- 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.
- P-value alone should not convince us that the effect is “real”; successful replications of the effect are much more convincing.

Example 2: ANOVA

Two nested comparisons allow us to draw conclusions about interactions.

- Example: 2 × 2 factorial design with factors:
  - Predicate type: complex vs simple.
  - Distance between the verb and an argument noun: long vs short.



Source: Vasishth S and Nicenboim B (2016)

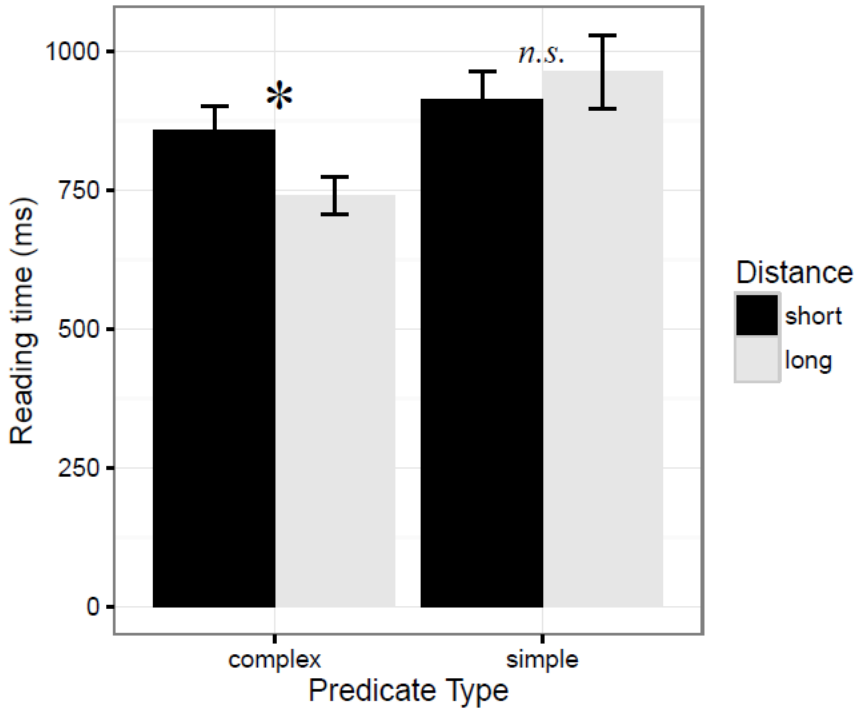
Example 2 cont'd

- One-sample t-tests for the effect of distance within complex predicates and simple predicates separately:
  - in complex predicates:

    $t(59)=-2.51$, p-value=0.02,
  - in simple predicates:

    $t(59)=0.52$, p-value=0.61.

- Can we now conclude that the interaction between the two factors exists?



Example 2 cont'd

- In the first test:

$$H_0: \ \mu_{short,complex} = \mu_{long,complex}$$
$$H_1: \ \mu_{short,complex} \neq \mu_{long,complex}$$

- In the second test:

$$H_0: \ \mu_{short,simple} = \mu_{long,simple}$$
$$H_1: \ \mu_{short,simple} \neq \mu_{long,simple}$$

- In the interaction:

$$H_0: \ \mu_{short,simple} - \mu_{long,simple} = \mu_{short,complex} - \mu_{long,complex}$$
$$H_1: \ \mu_{short,simple} - \mu_{long,simple} \neq \mu_{short,complex} - \mu_{long,complex}$$

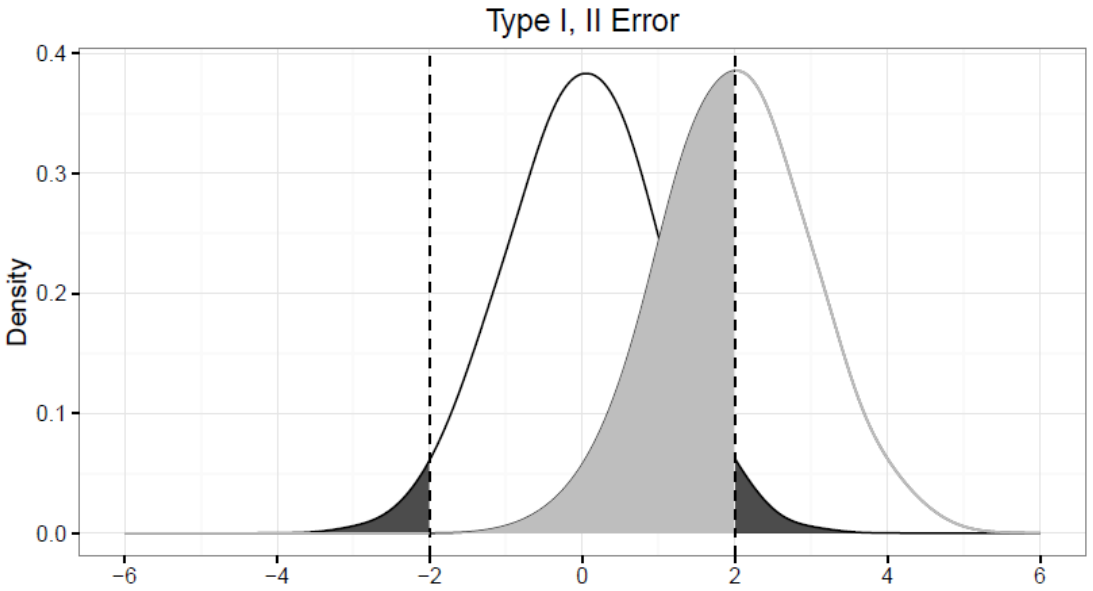
- When we do this t-test we find: t(59)=−1.68, p-value=0.1.
- Thus, one must always check whether the interaction is significant.

Example 2 cont'd

- 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.
- Example:
  - in one experiment a particular effect occurres,
  - but in a subsequent experiment with a new factor the effect disappeares,
  - the new factor in the second experiment led to the disappearance of the effect?
  - This would have been only valid if the relevant interaction had been found.
- As evidence, Nieuwenhuis, Forstmann, & Wagenmakers (2011) present a survey of published articles showing that approximately 50% of them (79 articles) draw this incorrect inference.

Type I and Type II errors

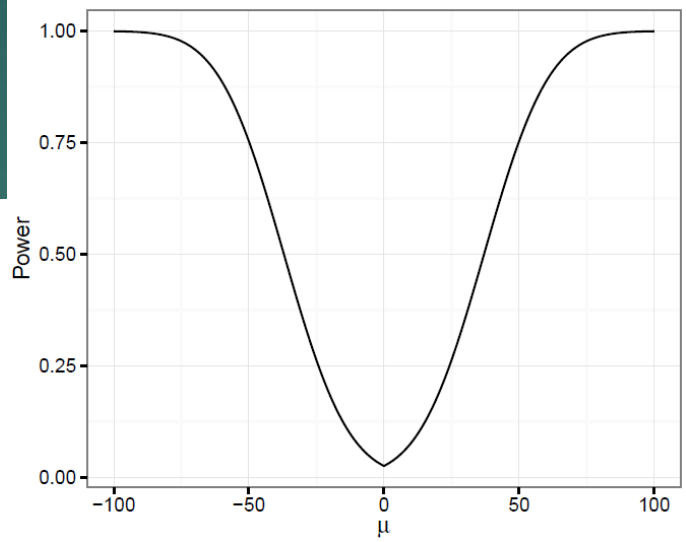
- Type I error – incorrectly rejecting H_0 when it is true.
  - The probability of type I error: α
  - Conventionally fixed at α =0.05 before we run an experiment.
- Type II error – incorrectly failing to reject H_0 when it is false.
- Power – the probability of correctly rejecting H_0 when it is false.



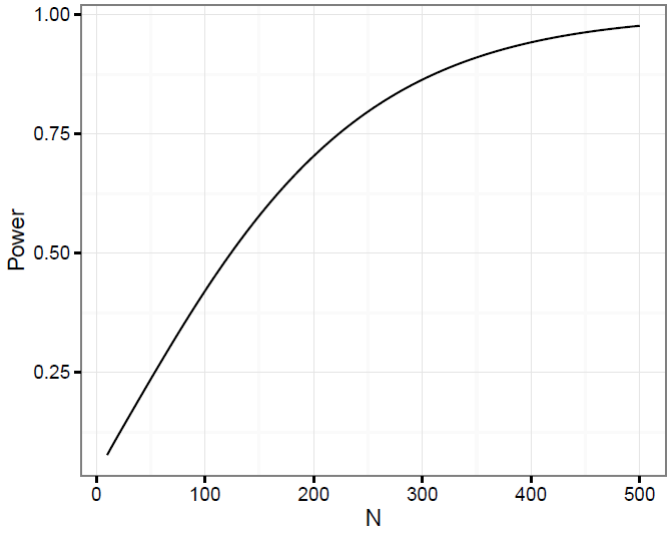
Source: Vasishth S and Nicenboim B (2016)

Power

- Power is a function: the further away μ is from 0, the larger the power.



- How to increase power and decrease Type II error?
  - design an experiment with a stronger manipulation, one which will lead to a larger effect.
  - Increase the sample size.



Source: Vasishth S and Nicenboim B (2016)

Power

- How to increase power and decrease Type II error?
  - 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.

Power

- It is especially important to do the best one can to achieve high power if we are interested in arguing for the null.
- 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.
- 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.
- Determine a realistic estimate of the true effect size for a particular phenomenon:
  - meta-analysis,
  - literature review,
  - knowledge elicited from experts on the topic.

Power

- There is no substitute for attempting to calculate power before running an experiment, using the best estimates one can obtain.
- It is a mistake to use ‘observed’ power, computed after the experiment has been run.

Example 3: LMM with low power

- Simulated example
- Simple 2-condition design
- 40 participants and 16 items
- Dependent variable: reading time (lognormal distribution)
- True effect size: 0.01 for log-transformed outcomes; a difference of 4 ms from a grand mean of 550 ms.
- 1000 simulated data sets
- For each data set a linear mixed model fitted with a log-transformed dependent variable, with varying intercepts for subjects and for items.

Source: Vasishth S and Nicenboim B (2016)

Example 3: LMM with low power

- Power: proportion of t-values>2
  - 0.09
- Type S error: proportion of models with signif. Effect but estimated effect in the opposite direction to the true effect:
  - 0.11
- Type M error: mean ratio of the estimated effect to true effect:
  - 5.08
- For low power, we are likely to get an inflated estimate.
- If power is low, the magnitude & sign of the effect may not be useful for calculating power in future experiments.

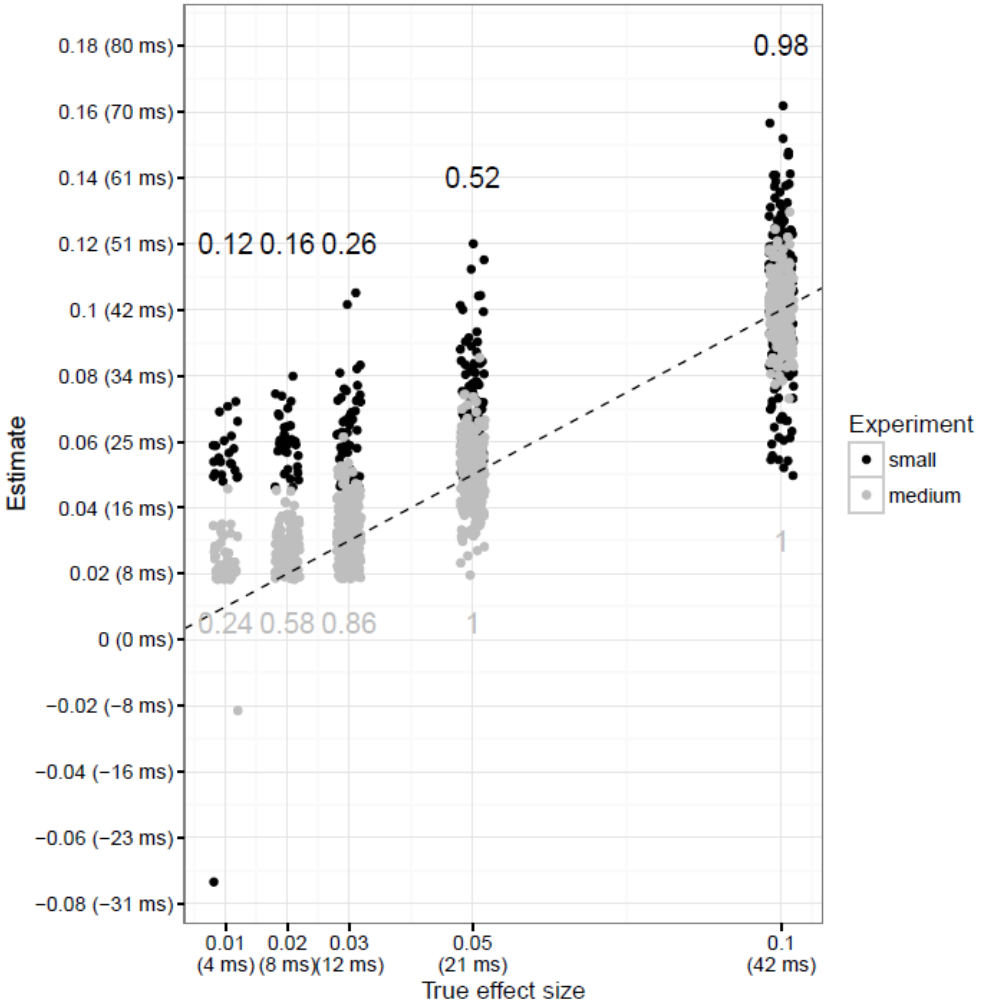
Source: Vasishth S and Nicenboim B (2016)

Example 3: LMM with low power

- The next simulation illustrates the problem of low power by showing potential differences between estimates and various true effect sizes.
- Data as before with different values of true effect sizes (0.01, 0.02, 0.03, 0.05, 0.1) and in two flavours:
  - a small sample experiment (but still publishable) with 30 subjects and 16 items,
  - a medium-sized experiment with 80 subjects and 40 items

Example 3: low power

- Plot shows 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, 200 experiments were simulated.
- The x-axis shows the true effect size on the log scale, the y-axis shows the estimate of the effect from linear mixed models with significant results.
- The power is shown within the figure.
- The dashed line shows the ideal situation where the estimate and the effects are the same.



Source: Vasishth S and Nicenboim B (2016)

Example 3: conclusions

- Exaggerated estimates (Type M errors) are more common for low-powered experiments.
- When the underlying effect is very small, some experiments will show results with the incorrect sign (Type S error).
- If researchers mistakenly believe that lower p-values give stronger evidence for the specific alternative hypothesis then journals publish larger-than-true effect sizes
- If power is very low, and if effect sizes are larger-than-true then power calculations based on published data overestimate power, and thus also overestimate the replicability of our results.

Data distribution and power

- 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)
- 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

Data distribution and power

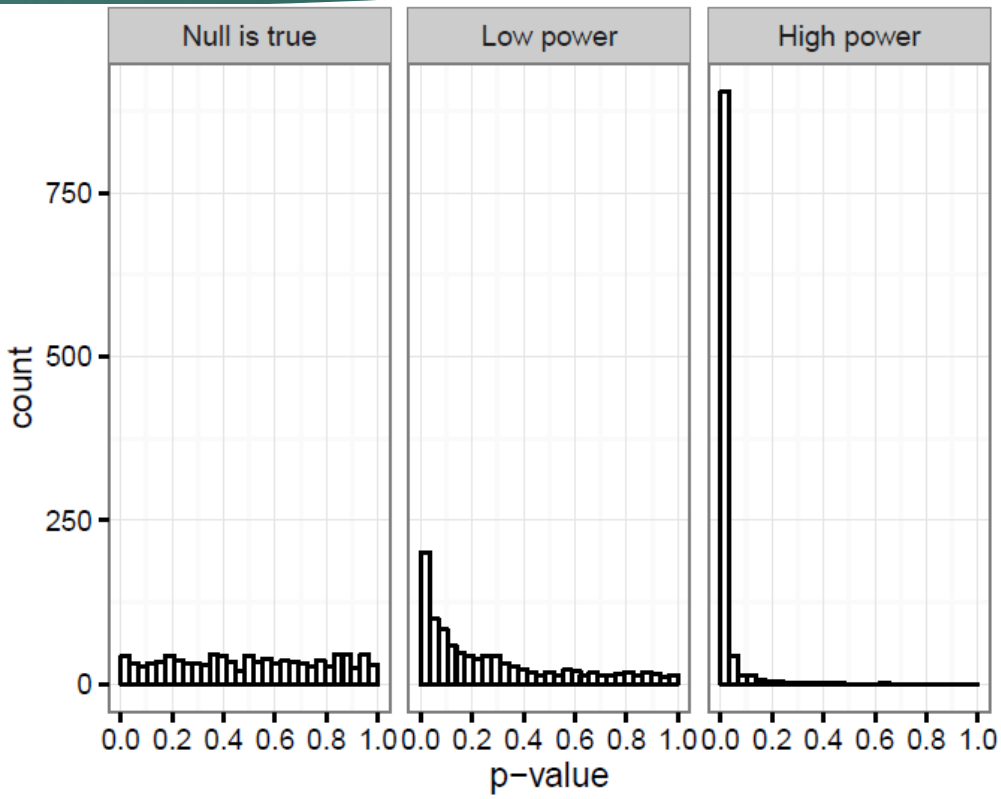
- Reading time or reaction time distributions are best fit with three parameter distributions such as:
  - **ex-Gaussian** (the convolution of a Gaussian and an exponential distribution),
  - **shifted lognormal** (the log-transformed normal distribution shifted to the right),
  - **shifted Wald** (the inverse of the normal distribution shifted to the right);
- Another way: **transform** 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 to find the right transformation;
  - for reading times, the reciprocal or the log transformation are often adequate, and easy to interpret.

Low power - summary

- 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.

P-values and power

- a true effect and high power will almost guarantee a very low p-value
- the distribution of p-values is uniform under the null hypothesis regardless of sample size.
- 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.



Simulated p-value distributions under a true $\mu = 0$ and under a true μ *not* 0 with low and high power (0.25 and 0.93 respectively). Vasishth S and Nicenboim B (2016)

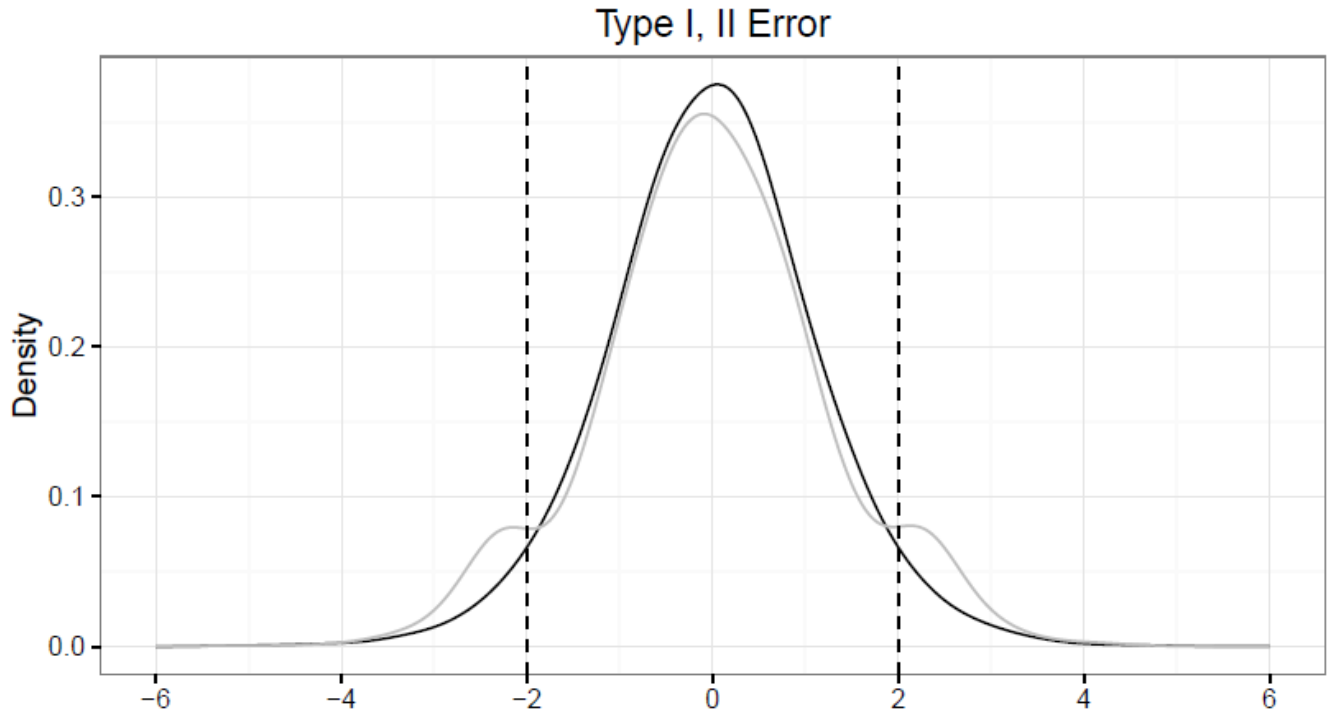
Replicability

- 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.

Running till 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.
- A typical initial *n* might be 15. This approach would give us a range of p-values under hypothetical repeated sampling.
- If we track the distribution of the t-statistic for this approach, we will find that Type I error is much higher than the assumed 5% (in our simulation, approximately 15%).

Running till significance is reached



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.

Vasishth S and Nicenboim B (2016)

Running till significance is reached

- 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.
- 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.

Degrees of freedom in analysis

- It is often necessary to compare the fits of different models.
- Problem: 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.

Variance decreases artificially due to aggregation

- When a LMM 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 LMM 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.

LMM's / Outliers

- LMM can solve some of the multiple comparisons problems if all relevant research questions can be represented as parameters in one coherent hierarchical model
- Another solution is to fit independent models but to apply some type of correction such as the Bonferroni correction
- Outliers may indicate a heavy-tailed distribution; therefore should not be removed automatically.
- transparency: authors should report failed manipulations, and whether the results depend on the removal of outliers; release the data and the analysis code.

Some remedies

- When new data can be easily gathered, an attractive solution is to take results as exploratory until being confirmed with new data.
- 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.

Conclusions

The best way to use frequentist methods is to:

- ensure appropriately powered hypothesis testing
- Check model assumptions
- Clearly separate exploratory data analysis from planned comparisons decided upon in advance
- Attempt to replicate results

Where data are sparse, Bayesian data-analysis methods should be considered…

References – part 1

- Al-Hoorie AH and Vitta JP (2018) The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*.
- Hu Y and Plonsky L (2019) Statistical assumptions in L2 research: A systematic review. *Second Language Research* 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
- Lindstromberg S (2016) Inferential statistics in Language Teaching Research: A review and ways forward. *Language Teaching Research* 20: 741–68.
- Larson-Hall J (2016) *A guide to doing statistics in second language research using SPSS and R*. 2nd edition. New York: Routledge.
- Liu Q and Brown D (2015) Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing* 30: 66–81.
- Loewen S, Lavolette E, Spino LA et al. (2014) Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly* 48: 360–88.
- 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.

References – part 2

- 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.
- 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
- Plonsky L (2013) Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition* 35: 655–87
- Plonsky L and Ghanbar H (2018) Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *Modern Language Journal* 102: 713–31.
- Plonsky L and Gass S (2011) Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning* 61: 325–66.
- Plonsky L, Brown D, Chen M et al. (2017) *Quantitative data ethics in applied linguistics: Sins of omission and commissions*. *Paper presented at the Annual Conference of American Association for Applied Linguistic*, Portland, Oregon, USA.
- Sesé A and Palmer A (2012) The current use of statistics in clinical and health psychology under review. *Clínica y Salud* 23: 97–108.
- Vasishth S and Nicenboim B (2016) Statistical methods for linguistic research: Foundational Ideas – Part I