UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale e Dipartimento di Psicologia
dello Sviluppo e della Socializzazione

Corso di laurea magistrale in Psicologia Clinica

Tesi di laurea magistrale

**The comparison of two independent proportions via
Design Analysis:
Theoretical background and an *R* implementation**

Il confronto tra due proporzioni indipendenti attraverso la Design Analysis:
Background teorico e una implementazione in *R*

Relatore

Prof. Gianmarco Altoè

Laureando: Riccardo Bettin

Matricola: 1204997

Anno Accademico 2020/2021

Table of contents

Summary.....	5
1. The replication crisis in science	7
1.1. The crisis	7
1.2. “Questionable publication practices”	9
1.3. Peculiarities about psychological sciences.....	12
1.3.1. Social consequences	13
1.3.2. Only a replication problem, or a replication misunderstanding? ...	15
1.3.3. <i>QRPs</i>	17
1.3.4. Low power	21
1.4. Aims	23
2. Design Analysis.....	25
2.1. Any solution?.....	25
2.1.1. Focusing on power	29
2.2. Design analysis and Type <i>M</i> , Type <i>S</i> errors.....	31
2.2.1. Plausible effect size	34
2.2.2. Prospective and retrospective design analysis	37
2.3. <i>PRDA</i>	41
2.4. Case study.....	43
3. Difference between independent proportions	47
3.1. Independent proportions and Design Analysis	47
3.1.1. Cohen’s <i>h</i>	50
3.2. Design analysis for difference between independent proportions	53
3.3. Case study.....	60
3.3.2. Retrospective design analysis	62
3.3.3. Prospective design analysis	68
4. Discussion and conclusions	71

4.1. General discussion	71
4.2. Limits	73
4.3. Further developments.....	74
References.....	77
Appendices.....	87
Appendix A: Code for prospective and retrospective design analysis applied to Wingen et al. (2020).	87
Appendix B: Code for the plot about the power of a difference of .05 between independent proportions with several sample sizes (<i>Figure 3.1</i>).....	89
Appendix C: Code for converting proportions to Cohen's h and vice versa.	91
Appendix D: Code of the functions for retrospective and prospective design analysis for the difference between independent proportions.....	92
Appendix E: R functions for design analysis with independent proportions.	96
Appendix F: code for retrospective design analyses and difference between H_0 and H_1 histograms on the case study.	102
Appendix G: Code for prospective design analyses on the case study.	105

Summary

In the last decade science has fallen into a replication crisis, this means that researches, when replicated, do not give the same results as the original ones. The difficulty in replicating studies can be due to several reasons, some of which regard the scientific world in general, such as the actual publication system that encourages incorrect behaviours and questionable research practices by scientists, and some that change between scientific fields. In fact, some scientific field feel this crisis more than others, and psychology is one of them. Low statistical power and misuse of statistics in psychology is reported from a long time. The first to criticize psychologist in regards of the use of power has been Cohen in 1962. This crisis can lead to the loss of trust in psychology and in science in general, for this reason it is important to find some solutions to the crisis.

Several possible solutions have been proposed. In this work we will focus on Design analysis, that is quite a new notion that can help in the process of getting out of the replication crisis. This analysis consists in calculating (power, and) new type of inferential errors to help researchers to better understand the consequences of low power, small sample sizes and studies started without appropriate planning. Design analysis can be done prospectively and retrospectively, that is different from post-hoc power calculations. The aims of this work are mainly to extend the design analysis to the case of differences between independent proportions and to provide an *R* implementation that can be used by researchers.

In the first chapter we are going to introduce the replication crisis, explaining how modern publication practices can lead scientists to try to publish at any cost. Referring to psychology, we will analyse what the social consequences of the crisis could be and how the trust loss in psychology could be tackled. An important subject that not always is addressed is the misinterpretation of replication itself and how to interpret results of replications. Low power and questionable

research practices, two things that in psychology are very common, are covered too in this chapter.

In the second chapter we will firstly introduce some proposed solutions to the replication crisis, such as the use of confidence intervals and safeguard power analysis. After this introduction, we will start explaining design analysis, what we think is a good way to try to plan better studies and make scientist reason more when planning studies rather than using mechanical reasoning through simple calculations. We are going to elaborate what the new type of errors are and what is defined as the plausible effect size, all key components for design analysis. Since design analysis can be retrospectively conducted, we are going to compare it to post-hoc power calculations and we will highlight the differences. Then, showing some examples too, an *R* package for design analysis will be introduced: *PRDA*.

In the third chapter we will do a brief introduction on the difference between independent proportion and on design analysis. We are going to show how the difference between proportion differ based on where the difference is situated on the proportion scale, and this leads us to explain the two main different approaches to study proportions. We also offer functions to convert values from one approach to the other and vice versa. The rest of the chapter is about the function we implemented for design analysis on difference between independent proportions, we explain how they work, we show some examples and we apply them to a real case study.

The last chapter is a discussion of what we have learned from this work, the limits of our considerations and functions and the further research that would be interesting to see in the next years starting from our conclusions.

First Chapter

The replication crisis in science

*“The scientific process is not ideological.
Science does not always provide comfort for
what we wish to be; it confronts us with what is.”
(Open Science Collaboration [OSC], 2015)*

In this chapter we will be introducing the replication crisis in science and some factors that can have a big role in the crisis of the scientific world. Then we will discuss the crisis in psychology facing a topic not widely tackled: social consequences of the crisis. This leads to trying to understand whether the problem is only about the low replication rates or about what people (and sometimes scientists) think replication is. Furthermore, questionable research practices will be considered, as in the recent years they have frequently been studied. Opposed to this, the long known problem of low statistical power in psychology will be briefly discussed, taking to the second chapter where it will be further developed.

1.1. The crisis

It is more than a decade that a replication crisis has been recognized in science and the first to bring awareness to this problem was Ioannidis (2005). In 2005 Ioannidis published “Why most published research findings are false”, he explained that sometimes studies do not replicate, that most published research claims could be

false and that this creates confusion and disappointment between scientists. This crisis has, since then, been recognized across many scientific fields such as medicine, biology, economics, experimental social sciences, and psychology (Open Science Collaboration, 2015; Perugini, 2014; Ioannidis, 2005; Button et al., 2013; Szucs and Ioannidis, 2017; Freedman, Cockburn and Simcoe 2015; Camerer et al., 2016; Baker and Dolgin, 2016; Camerer et al., 2018; Mede, Schafer, Ziegler and Weisskopf, 2020). Poor reproducibility rates led to a metascientific field that studies scientific methodology itself trying to improve it. Now, thanks to metascience, there is cumulated evidence that research practices can be enhanced to make a more efficient use of public financial investments in the scientific community (Munafò et al., 2017).

“There are two important aspects to these insights [of Popper’s] that inform scientific thinking. First, a finding needs to be repeatable to count as a scientific discovery. Second, research needs to be reported in such a manner that others can reproduce the procedures” (Zwaan, Etz, Lucas and Donnellan, 2018). But what does it mean to replicate a result in science? We will investigate further this topic in a few pages, for now let’s consider replication as “testing if empirical studies yield the same results when repeated under the same or similar conditions” (Mede et al., 2020). Being able to repeat a study is a basic requirement for science and it’s necessary for cumulative knowledge, but when results are not replicated, not only the credibility of the results but of the entire discipline is questioned (Perugini, 2015), “a field known for publishing false positives risks losing its credibility” (Simmons, Nelson and Simonsohn, 2011). Thus, replicated results are more reliable and solid than non-replicated ones (Hendriks, Kienhues and Bromme, 2020), still “more than 70% of researchers have tried and failed to reproduce another scientist’s experiments, and more than half have failed to reproduce their own experiments” (Baker, 2016).

Ioannidis (2005) considers the widespread but wrong idea that one study with statistically significant results is enough to claim conclusive research findings one of the key factors for nonreplicable results. Along with this, he includes biases, power (small samples and small effect sizes), flexibility in the research design,

financial interests, prejudices in a specific field and the numbers of independent teams involved in a specific topic to be aspects of the crisis that influences one another. Other factors such as measurement errors and questionable research practices that facilitate the publication of statistically significant results are considered as the main reasons of the replication crisis too (Szucs and Ioannidis, 2017).

An example of how measurement errors can influence replicability is explained by Loken and Gelman (2017). Scientists usually think that finding strong and robust evidence with noisy data is proof of how strong the underlying effect is, but this is not the case. This is called the “what does not kill statistical significance makes it stronger” fallacy. Noise is everything that influences the measure of a variable, increasing the uncertainty around the parameter estimates and producing a difference between observed and true values. Since measurement errors are defined as a random variation and are part of the noise, poor measurements can not only reduce the estimate of an effect size but exaggerate it too. In noisy settings, estimates can be an order-of-magnitude larger than the plausible effect, thus, statistical significance is no strong evidence for the magnitude or even the sign of an effect (Loken and Gelman, 2017). “The consequences for scientific replication are obvious. Many published effects are overstated and future studies, powered by the expectation that the effects can be replicated, might be destined to fail before they even begin.” (Loken and Gelman, 2017).

Many causes can be considered at the base of the replication crisis, some of them can be considered part of the scientific nature (i.e.: uncertainty, that cannot be solved) and others may differ between scientific fields, but there is one that has a significant role in every field: publication practices.

1.2. “Questionable publication practices”

When you think about publication practices you probably think about publication bias; this is one of the “questionable publication practices” that will be discussed in

this paragraph, but first we will analyse other ways that publication can distort science with.

Young, Ioannidis and Al-Ubaydli (2008) compared scientific information to economic commodity to explain how some economic terms have analogies in the scientific world and how these analogies can have an impact in publication and thus distort the reality of scientific data. Since journals create revenue and are important for scientific development, knowledge can be considered as the commodity that is transferred from scientists (it's producers) to other scientists (or patients, agencies, etc., that are the consumers). Now we will analyse what Young et al. (2008) came up with (summary in *Table 1.1*).

Table 1.1: Analogies between economic terms and scientific publication.

Economic Term	Meaning	Analogy in Scientific Publication
Winner's curse	The winner in an auction tends on average to have overpaid, especially when no participant is sure exactly how valuable the item is.	Scientific studies try to find true relationships, but none are certain of what these relationships are exactly. Published articles, especially in very competitive journals, have on average exaggerated results.
Oligopoly	A market where a few traders have the major share and each oligopolist has significant power to influence the market.	Very few journals with limited publication slots (compared with geometrically increasing scientific data that seek publication) determine highly visible science.
Herding	"Follow-the-leader" behaviour: the actions of the first or dominant player supersede the individual information and actions of all the players in a market.	Scientists may uncritically follow paths of investigation that are popularised in prestigious publications, neglecting novel ideas and truly independent investigative paths.
Artificial scarcity	Restrictions on the provision of a commodity above that expected from its production cost.	Print page limits are an obvious excuse for failure to accept articles, and further the small number of major "high-impact" journals have limited slots; extremely low acceptance rates provide status signals to successful publications and their authors.
Uncertainty	Situation where the real long-term value of a commodity is largely unpredictable.	For much (most?) scientific work, it is difficult or impossible to immediately predict future value, extensions, and practical applications.
Branding	Marking a product as valuable; of key importance when it is difficult to determine a product's value prior to consuming it.	Publishing in selective journals provides evidence of value of a research result and its authors, independent of the manuscript's content.

Notes: Reprinted content from Young et al. (2008).

Between 1997 and 2007 in Scopus-indexed journals the number of published findings had a 50% increase (from 590,807 to 883,853) but the number of journal slots available has not increased that much, causing only a limited

number of conducted studies published in journals. This creates an oligopoly of a few journals that can have a high impact on what will be published. Since these few journals use impact factors as criteria for publication, the exaggeration of result is even worsened (Young et al., 2008).

Furthermore, being published in such few prominent journals gives scientists a status signal that makes them advance in their career, but this makes herding easier. When few successful individuals choose an area of investigation, many teams undertake the same trajectory leading to investigation in a path that, not only could be incorrect, but completely wrong (Hirshleifer and Teoh, 2003).

If investigations are not in the wrong direction, it does not mean that we can be certain about the immediate or long-term value of the findings. Thus, uncertainty plays a big role in publishing since there is no way to predict when, why or what will be valuable in future. Branding is used to assign value and “overcome” uncertainty and “serves a similar and complementary function to the status signal for scientists” (Young et al., 2008).

Journals’ selectivity gives them a lot of authority; one result of this is artificial scarcity. Nowadays, with online papers, there could possibly be no limit to “printed page space”. Making available only some studies on paper printed journals and all the others on online journals would be quite easy. Journals are using the excuse of limited printed page space to keep low acceptance rate and to give the idea “of exclusivity based on merit and more frenzied competition among scientists “selling” manuscripts.” (Young et al., 2008).

Last but not least: winner’s curse. This is the case of scientists who find significant results (usually) with low power and noisy data, the results are exaggerated and the scientist’s “luck” for finding them becomes a curse. Winner’s curse has a significant role in publication bias, since the latter occurs when a study with a significant result has greater probabilities to be published than one with non-significant results, even if the research quality is the same. The more the results are extreme, the more probability there is that they are published (Young et al., 2008). This means that a scientist that falls in the winner’s curse trap finding extremely exaggerated data has more probabilities to be published, but, when his study is

replicated, no effect or a largely smaller effect is found (knowns as “the decline effect”) and consequently we have low replication rates.

“Even if a researcher conducts studies competently, analyses the data effectively, and writes up the results beautifully, there is no guarantee that the report will be published. [...] success in publishing is partly a function of social savvy of knowing what is publishable and empirical savvy in obtaining publishable results.” (Nosek, Spies and Motyl, 2012). Publication bias is not due only to journals and the way articles to publish are chosen, but to scientists too: most null result studies are never even written because they are seen as if there is no publication potential; this is called the “file drawer” problem (Franco, Malhotra and Simonovits, 2014). This is troubling for two reasons: efforts and resources are wasted because of researchers repeating the same study to find no effect at all, and, if it is found a significant result by chance conducting a similar study, then the absence of literature of null results on the topic will affect knowledge by giving the impression that there could be stronger effects (Franco et al., 2014). In a way, this can lead to low replication rates too, since by trying to replicate the result, you would find the same results as the non-published null result findings on the topic.

So far, we have introduced the replication crisis in science and some problems in research publication in general. Every scientific field has different approaches and different difficulties and weaknesses; in each of these fields several characteristics of each had a significant role in the replication crisis. In the next paragraph we will analyse some peculiarities about the crisis in psychological sciences.

1.3. Peculiarities about psychological sciences

One of the most known replication studies is the Replication Project: Psychology (RP:P) conducted in 2015 by Open Science Collaboration. Several teams worked on the replication of 100 published studies: only about one third reported similar result to the original and the ones that replicated showed smaller effect sizes. A similar work has been done in 2018 by Camerer et al. and found similar results,

concluding that effects in literature are exaggerated because of low statistical power and bias selecting significant results for publication.

This low replication rates, as much as other complaints because of small samples, low power, questionable research and measurement practices and null-hypothesis significance testing (NHST) are part of the crisis and some doubts should automatically rise: what about the trust that society lies in psychological research and findings? Can this trust be lost because of low replication? And what can we do to get it back?

1.3.1. Social consequences

Some scientists had these doubts about science in general (Hendriks et al., 2020; Broombell and Kane, 2017). Broombell and Kane (2017) claim that non-scientists that perceive scientific evidence as uncertain will also perceive the respective scientific field of lesser value. Hendriks et al. (2020) found that knowing about the non-replication of a result modified people's judgment on the credibility and strength of the original research. The trustworthiness of the scientist was influenced too by the success or fail of the replication, even though the effect was small. "We found in our study that laypeople judge the credibility and evidence strength of research results and the trustworthiness of researchers differently after learning about a successful or failed research study. This could be because participants recognized and evaluated the replication as a scientific practice used to ascertain the certainty of results" (Hendriks et al., 2020) and they added that, in a public controversy about a socio-scientific issue, the success or failure of a replication would have an even greater impact on the credibility of scientific claims and on the trustworthiness of the researchers of the original study.

Psychological research plays a big role in society and at the same time it relies on society for participation and for the necessary public funding, so having a trustworthy reputation as scientific field is a must. Moreover, policy makers and practitioners rely on it (Wingen, Berkessel and Englich, 2020). Wingen et al. (2020; see Anvari and Lakens, 2018 too) tried to answer to the questions above relatively

to psychology and hypothesized some possible solution to get back people's trust. They conducted 5 studies to assess if in psychology the replication rate influences people's trust and if solutions as science transparency, explanations about replications and restoring replicability could work to increase trust.

In the first study they examined whether the replication rates correlate with the trust and the perceived value that people assign to psychology by making people estimate psychology's replication rate and how much they trust psychology (for better information on how the research was conducted, please refer to the original research: Wingen et al., 2020). In the second study they assigned people to three conditions and to each was taught about a low or medium or high replication rate in psychology and then measured people's trust in psychology in each condition. In both studies people's trust correlates with their beliefs about psychology's replication and the replication rates they were taught about: the higher the replication rate, the higher the trust.

In the third study Wingen et al. (2020) evaluated if informing people about increased transparency and the open science movement (preregistrations, open data, open materials) would help to gain back trust. Trust in psychology did not vary learning about open science and transparency.

In study number 4 the aim was to evaluate the effectiveness of explanatory strategies such as *QRPs* and hidden moderator as reasons of the low replication. The hidden moderators' explanation did not have a significant effect nor the *QRPs* one, but the latter had lower trust than the former.

In the last study it was tested whether the public's trust could be repaired if replication rates got higher, but participants did not show higher trust either.

Even though the last three studies did not have enough power to find small effect sizes, it seems that these solutions would not have large trust-repairing effects. Wingen et al. (2020) suggested that probably, if replication rates stay high for enough time, people's trust would start to increase. Since regaining trust is not easy, having stronger evidence from the beginning, that would mean being able to have higher replication rates too, is even better and it can be considered part of the aims of this work.

In a similar study, Hendriks et al. (2020) realised that people don't necessarily have the right idea about what replication is, how it works and what it means to replicate or not a result, so is it...

1.3.2. Only a replication problem, or a replication misunderstanding?

Hendriks et al. (2020) wrote that their experiment could be considered an “empirical analysis of folk theories about replication” because they noticed that participants, when a study failed to replicate, did not know what exactly was called into question, and that their judgement on the original study scientist's trustworthiness diminished relatively to the failed replication, independently on the fact that it was a self-replication or not. Hence, in this paragraph we will try to clarify some wrong, or at least not accurate, beliefs about replication.

To explicate the doubts mentioned right above, from a scientific point of view it's wrong to blame the author of the original study for a failed replication or a negative outcome of a study, but the difference in trustworthiness attributed from people to the author in Hendriks et al. (2020) could be due to the mistrust because of questionable research practices. It must be specified too, that single successful or failed replication are difficult to interpret (Maxwell, Lau and Howard, 2015): a single failed replication gives no answers on the reasons it did not replicate, whether it was because the original study or the replication itself, and those results can only be used to evaluate the results of the original research, but not the underlying claims (Hendriks et al., 2020).

In scientific context, regardless “folk theories”, replication has different functions depending on what is the claim of the original research you want to verify (Fletcher, 2021):

1. There were no mistakes in the data analysis that led to the results: statistics are recalculated to verify if there are no technical errors, and in some cases several types of alternative analysis are verified to see if results are the same as in the original study (Nuijten, Bakker, Maassen and Wichert, 2018). These supplementary analyses are conducted on the same data of the

original research and, in a way, it can be considered part of the fourth point too.

2. Sampling error did not have a role: low sample sizes and chance could produce misleading data, and this could be prevented by collecting more data, the more, the better.
3. Results are not related to contextual factors: during data analysis not all variables and their interaction with independent variables are accounted for, doing so could change the results of the original study.
4. Exclude fraud or questionable research practices: the methods of a study should, as already cited “be reported in such a manner that others can reproduce the procedures” (Zwaan et al., 2018), they should be reported very precisely to assure the research’s validity; using questionable research practices and fraud distorts the methods described. As a matter of fact, Hendriks et al. (2020) noted that their participants supposed that replication are good scientific practice regardless of who did them, but a replication from an independent lab has more scientific soundness because it can rule out questionable research practices (and even fraud). In fact, Makel, Plucker and Hegarty (2012) noted that when replications were totally independent from the original authors, they were less likely to be successful too.
5. Generalization: this means to assess if a hypothesis applies to a larger (or different) population than the one it was studied from in the original research.
6. Operationalising or testing the hypothesis in a different way from the original study confirms the results: if the original hypothesis is right, using different ways to test it or operationalise it means to explore and test the implication about the hypothesis, thus, make it sounder.

Ebersole, Axt and Nosek (2016) claimed that:

“Taking risks with what could be true seeds the marketplace of ideas with new directions. If we are not wrong frequently, we are failing to push on the frontiers

of knowledge hard enough. At the same time, because true innovations are rare, valuing replication will foster efficient filtering of interesting findings and accelerate knowledge building.”

Whilst this can be considered true in the one hand, we must take into account other matters: false positives can lead to investment in fields that turn out to be unproductive and, in the scientific practice, false positives can be very ongoing and failed replications are never conclusive about them (Simmons et al., 2011) and several replications are needed to clarify these inconsistencies (Maxwell et al., 2015). Hence, in this case too can be better having stronger results from the beginning and finding a way to interpret how much the results of a study are accurate. For this reason, in the second chapter will be introduced Type *M* and Type *S* errors.

1.3.3. *QRPs*

Questionable research practices (*QRPs*) have been already mentioned several times in regard to replication; in this paragraph we will see what they are and how frequent they are in the psychological field.

QRPs are “the grey area of acceptable practices” that increase the likelihood of finding significant result and strong evidence, they are more frequent and could be more detrimental for scientific literature than fraud (John, Loewenstein and Prelec, 2012). The term is used to not only to describe overtly misleading practices, but for ambiguous practices too; even though in this case they could be completely unintentional by the researcher (i.e.: excluding participants for good reasons but not being totally transparent about it; Button et al. (2013) claimed that excluding 3 observations in a sample of 12 can significantly change the results) and they can go from simple omission of information to fraud (Flake and Fried, 2019).

Agnoli, Wicherts, Veldkamp, Albiero and Cubelli (2017), John et al. (2012) and Fiedler and Schwarz (2016) conducted a survey respectively in Italy, in USA

Table 1.2: Prevalence of scientists (self-admission) having adopted at least one questionable research practice in Italy, USA and Germany with 95% confidence intervals in brackets.

Questionable research practices	Italy	USA	Germany*
Deciding whether to collect more data after looking to see whether the results were significant.	53.2 (46.6 – 59.7)	55.9 (51.5 – 60.3)	~ 35%
In a paper, failing to report all of a study's dependent measures.	47.9 (41.3 – 54.6)	63.4 (59.1 – 67.7)	~ 35%
In a paper, selectively reporting studies that “worked”.	40.1 (33.6 – 46.6)	45.8 (41.3 – 50.2)	~ 40%
Deciding whether to exclude data after looking at the impact of doing so on the results.	39.7 (33.3 – 46.2)	38.2 (33.9 – 42.6)	~ 40%
In a paper, reporting an unexpected finding as having been predicted from the start.	37.4 (31.0 – 43.9)	27.0 (23.1 – 30.9)	~ 45%
In a paper, “rounding off” a <i>p-value</i> (e.g., reporting that a <i>p-value</i> of .054 is less than .05).	22.2 (16.7 – 27.7)	22.0 (18.4 – 25.7)	~ 20%
In a paper, failing to report all of a study's conditions.	16.4 (11.5 – 21.4)	27.7 (23.7 – 31.7)	~ 35%
Stopping collecting data earlier than planned because one found the result that one had been looking for.	10.4 (6.4 – 14.4)	15.6 (12.4 – 18.8)	~ 5%
In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do).	3.1 (0.9 – 5.4)	3.0 (1.5 – 4.5)	< 5%
Falsifying data.	2.3 (0.3 – 4.2)	0.6 (0.0 – 1.3)	< 5%

Notes: Adapted from Bertoldi, G. (2019). Italian data from Agnoli et al. (2017). USA data from John et al. (2012). German data from Fiedler and Schwarz (2016).

*Data from Fiedler and Schwarz (2016) are approximated from a graph, since on the paper and in supplemental material no exact data were reported.

and in Germany to find out how widespread these practices are between psychologists. The results are quite similar between the three different contexts, showing that *QRPs* are quite common with a self-admission prevalence of 27.3% in Italy, 29.9% in USA and 25.5% in Germany (prevalence based on self-admission in *Table 1.2*). On average, more than a quarter of the *QRPs* considered in these three studies were used by scientists, with Italian and USA scientists considering them defensible or at least possibly defensible (Agnoli et al., 2017). “These consistent results are substantial reasons for concern about the integrity of the research literature.” (Agnoli et al., 2017).

Table 1.3: The percentage of 15,000 simulations in which at least one of a set of analysis was significant.

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A	17.8%	9.5%	2,2%
Situation B	14.5%	7.7%	1.6%
Situation C	21.6%	11.7%	2.7%
Situation D	23.2%	12.6%	2.8%

Notes: Situation A: data collection of 2 dependent variables ($r = .5$), testing whether the manipulation affected one, the other, or both. Situation B: collecting 20 observations then testing for significance, if the result is significant the data collection is stopped, if not, 10 more observations are collected, and the data is tested again for significance. Situation C: controlling for gender or for an interaction between gender and the independent variable. Situation D: running three conditions and reporting the results for any two or all three (i.e.: low vs. medium, low vs. high, medium vs. high, low vs. medium vs. high). Adapted from Simmons et al. (2011).

A well-known *QRP* is *P-hacking* (also known as fishing; Gelman, 2013a), it means to implement several data analysis only after the data has been collected without them being decided a priori. This does not mean that scientists collect the data and then decides what analysis apply to find a significant result, but in most cases that they have an idea at the beginning and only seeing the data they definitely decide what to do (Gelman and Loken, 2013). *P-hacking* combined with confirmation and hindsight bias can lead to consider appropriate expected and desired result that can consequently lead to post hoc hypothesis without realizing or indicating their nature (Munafò et al., 2017). This is called HARKing (Hypothesizing After the Results are Known) and can be defined as “presenting a post hoc hypothesis in the introduction of a research as if it were an a priori hypothesis” (Kerr, 1998).

“It's easy to find a $p < .05$ comparison even if nothing is going on, if you look hard enough - and good scientists are skilled at looking hard enough and subsequently coming up with good stories [...] to back up any statistically-significant comparisons they happen to come up with” (Gelman and Loken, 2013). In fact, a prevalence of *p-values* just under .05 was found from Masicampo and Lalande (2012), suggesting the presence of publication bias and P-hacking too. Simmons et al. (2011) showed that these practices strongly increase the probability to find false positive results by simulating situations with various researcher degrees of freedom (the flexibility that researchers have in taking decisions about data analysis, sample size, variables, data exclusion, exc.) and calculating the real probability of a false positive result with a fixed 5% Type I error (α ; results in *Table I.3*).

It is important for this work to report their thoughts about flexibility in sample size too; in fact, sample size is not always set a priori but researcher often stop the data collection based on the interim data analysis. Researchers are not aware of the impact this can have, most of them think that it has only little effect on finding a false positive result, but it is not the case at all (Simmons et al., 2011).

1.3.4. Low power

Statistical power is the probability of finding an effect if the effect really exists. It mainly depends on 3 factors: sample size, effect size and the significance level (α ; Sedlmeier e Gigerenzer, 1989) and the bigger these 3 factors, the higher is the power. If a study has low power, it means that, even if an effect exists, it is very difficult for it to find and to estimate the real value. Contrarily, if a statistically significant result is found with low power, it does not mean, as we will see in Chapter 2, that the finding is even more remarkable (Gelman and Loken, 2014).

Low power is a distinctive feature about psychological research. Already in 1962 Cohen criticized this aspect of psychology in “The statistical power of abnormal-social psychological research: A review”. He claimed that most researchers base the sample size on weak basis, and this could lead to low success in finding effects that exist and/or false positives, thus, a waste of resources. Mean power he found in psychological research is of .18 for small effects, .48 for medium effects and .83 for large effects. Even though Maxwell et al. (2015) “believe that values in the neighbourhood of .90 and .95 may be more appropriate than .80”, ideal power is considered to be around .80 (Cohen, 1988), so the studies had enough power only to detect large effect size but not medium or small ones.

After Cohen, many conducted similar researches finding similar results. Twenty-seven years later, Sedlmeier e Gigerenzer (1989) found that power did not increase from what Cohen found, in fact the mean power was .21 for small effects, .50 for medium effects and .84 for large effects. Nothing has change in nearly 30 years and “researchers paradoxically seem to prefer probable waste of time, money, and energy to the explicit calculation of power.” (Sedlmeier e Gigerenzer, 1989).

What about after another 15 years? “Despite ever-increasing attention, more recent reviews of published literature in psychology have largely continued to show that the majority of these studies lack appropriate statistical power.” (Maxwell, 2004). Only in 2017 Szucs and Ioannidis found that power was slightly higher for medium effect sizes, but for small and large it was quite the same as before.

Often researchers do not take into consideration power because they claim it is useless. This happens because they don't always know effect size, they argue that they don't have enough data to estimate them and to do it the study has to be conducted, so their conclusion is that power analysis is needless (Westfall, 2015). A commonly used excuse is that the data that have been collected are extremely time and labour expensive (thus, small sample size), but "the numbers don't know that. [...] Since the numbers don't remember where they came from, they always behave just the same way, regardless" (Lord, 1953) hence, a small sample size has low power regardless how the data was collected (Lilienfeld and Strother, 2020). Moreover, small studies can end in the file drawer more easily than big studies (Button et al., 2013).

Power is directly related with replication because the more power a study has, the more probabilities there are that the study can find the underlying effect (if it exists) and estimate the real effect size, without there being magnitude (Type M) or sign (Type S) errors. Furthermore, "these so-called failures to replicate may not be failures at all, but rather are the result of low statistical power in single replication studies, and the result of failure to appreciate the need for multiple replications in order to have enough power to identify true effects." (Maxwell et al., 2015). This means that not only the power of the original study may affect the probability of it replicating, but even the power of the replication itself could make the replication end in failure if the effect size of the original study was exaggerated enough.

Without taking power in consideration, researches could not replicate due to false positive, or exaggerated (Button et al., 2013;) and/or in the wrong direction estimates of effect sizes (Gelman and Carlin, 2014). These errors are part of the design analysis first suggested by Gelman and Carlin (2014); a way to integrate power analysis to design stronger studies and to make results more interpretable.

The concept of power is not always clear and there are many ways it can be misused and misunderstood, some of which will be addressed in the next chapter explaining what design analysis and plausible effects size are and how post-hoc power calculations are misleading.

1.4. Aims

We have analysed when the crisis started and some factors of the publication system that have contributed to worsen it, learning that trust in science can be lost. In particular we noticed that in psychology trust not only can be lost, but it's difficult to regain also because replication functions are not clear in non-scientists. If we consider *QRPs*, the situation can get worse to the eyes of scientists too, since for some of them some *QRPs* are defensible, but this does not change the effects these practices have in literature. Low power had a significant role too, but more of it will be discussed in the next chapter.

After analysing the replication crisis and the peculiarities about psychological sciences and their roles in the crisis we can clearly say that the aims of this work are, mostly, to present a possible solution: design analysis. This could be a solution not necessarily only to low power, but to interpret in a better way the results and the consequences of low power and small sample sizes (relating both to the effect size) of already conducted studies and to better understand how to design new studies (referring to sample size, power and estimating the effect size). Moreover, we provide a function to conduct design analysis on difference between proportions on the *R* cran statistical software, as for Cohen's *d* and Pearson's *r* they already exist (Altoè et al., 2020; Bertoldo, Callagher and Altoè; 2021; Callagher et al., 2021).

As a broader purpose we would appreciate to bring more awareness in general not only at the replication crisis in psychology and to the methodological aspects (already largely addressed in literature) but also at the less discussed problem of social consequences that a crisis in psychology can lead to if not addressed properly, asking ourselves what the best path is to get back on track.

Second Chapter

Design Analysis

“[...] Even though a design analysis requires significant effort, we believe that it has the potential to contribute to planning more robust studies and promoting better interpretation of research findings.”
(Altoè et al., 2020)

In this chapter we will introduce some possible solutions to the replication crisis. After this we will start to introduce type M and type S errors for design analysis, explaining what they are, how they are calculated and what is their meaning, leading to define what a plausible effect is and how to estimate it. We will then explain prospective design analysis and retrospective design analysis and what is the difference with retrospective power analysis (post-hoc power). At the end of the chapter, we will introduce an R package to conduct design analyses and in the last paragraph there will be an example of design analysis applied to a real case study with the presented R package.

2.1. Any solution?

In this paragraph we will briefly address some proposed solution for the replication crisis itself, such as open science and increased transparency, and other possible solutions that has been proposed to solve the low power and small sample size

problems, and we will introduce the problem in estimating effect size. There are a lot of proposed solutions, even though the aims of this work are not to provide an exhaustive list of all possible solutions in depth, we find quite interesting various suggested solutions and we will try to mention the most known ones on different aspects of the topic, even though we will concentrate on design analysis the most.

As a large-scale solution to avoid *QRPs*, biases or fraud and thus try to make result and researches more replicable and evaluable it has been suggested to make science more “open” and transparent. This means to make publicly available all the collected data and the materials and to pre-register studies’ hypotheses and analyses. To incentive these procedures, badges are assigned to researches that follow these steps: Open Data Badge, Open Materials Badge and Preregistered Badge (Perugini, 2014).

For several reasons Button et al. (2013) and Munafò et al. (2017) suggest researchers to pre-register their studies, improving reporting and disclose methods and findings transparently and make data and materials available. Making data and materials publicly available helps to replicate and extend one’s findings and to have the analyses and results checked by others.

Pre-registration means to register a study design and, even better, a specification of the research’s procedures, outcomes and statistical analyses planned to be conducted (all of this, clearly, before conducting the study and knowing the results). This helps other researchers to understand whether a study is confirmatory or exploratory, since presenting a confirmatory analysis as exploratory makes false positive easier. Publication bias and the file drawer problem are directly tackled too because, even though a research is not published (or not written at all), the presence of the registration makes it discoverable. In some cases (i.e.: Registered Reports) study designs are reviewed before starting to collect the data. “Preregistration has the goal to allow others to transparently evaluate the capacity of a test to falsify a prediction. [...] (and) to prevent researchers from non-transparently reducing the capacity of the test to falsify a prediction” (Lakens, 2019). Lakens (2019) adds that is not pre-registration itself to make a study better than a non-registered one, but rather the fact that a pre-registered study is more

evaluable from other researchers and because pre-registering a study makes scientists carefully think to what they are doing so that they address issues that otherwise wouldn't have thought about. Sometimes, by going through the process, scientists might even realise that they are not ready to test a given hypothesis.

Improving reporting and disclosing methods and findings transparently is a must if we want researches to be more replicable, since without specific and clear information there is no way to understand how a study was conducted in the first place. Furthermore, explanations and justification for certain decisions help readers to better understand and interpret findings. “Without transparency, claims only achieve credibility based on trust in the confidence or authority of the originator. Transparency is superior to trust.” (Munafò et al., 2017).

While pre-registration is a way to address the replication crisis directly, other options has been brought up to test for significance and to estimate the effect size, in particular we refer to the use of confidence intervals. Confidence intervals create a range of plausible values of an effect size with a certain probability that the true value of the population is included based on the collected data. The lower and upper limits of the interval suggest how small or how big the effect might be; a wide range of values shows uncertainty around the effect estimate even if the null value is not included in the range (that means the result is statistically significant; Flora, 2020). But confidence intervals should not be considered as a significance test considering a result “significant” if the null value is not in the interval and “non-significant” if the null value is inside the interval (Poole, 1987) and researchers should pay attention to the whole range of the interval (Flora, 2020).

Hoenig and Haisey (2001) reported two interpretation of confidence intervals, one being the most known that they cover the true value with a certain level of probability, and the second interpretation is that confidence intervals are equivalent to hypotheses tests. In both cases the width of the interval shows how confident we can be about the estimate of the effect size, but this is not straight forward. In fact, in the second interpretation in which the excluded-from-the-interval values are refuted by the observed data and the included values cannot be refuted, if the range of non-refuted values is very limited and near to the null value

we can be confident that the measured effect is near to zero, but if the range is wide and contains the null value it is not a sign that the effect is null even though the results are not statistically significant.

In this regard, multiple examples have been reported (i.e.: Goodman and Berlin, 1987; Maxwell et al., 2015; Dziak, Dierker and Abar, 2018). If we have a new therapy that to be considered efficient has to be at least 20% better than the already in use therapy, and the difference we find is of $11\% \pm 13\%$, the confidence interval is from -2% to 24% . The interval does not rule out the 0% neither 20%, this means that 0% is a plausible value, but it is wrong to conclude that the effect is equal to zero and that the therapy is not efficient since 20% is a plausible value too, so these two values are both equally supported. A different situation would be in the case of a $2\% \pm 13\%$ and in a $1\% \pm 2\%$ difference. In the former case the confidence interval would be from -11% to 15% , that excludes 20% but it would be wrong to conclude that the effect is null, in the latter case the confidence interval would be from -1% to 3% and we can be quite confident that the effect is near to null. In both cases the new therapy would not be considered efficient, but practical implications would be different. Note that applying this in replication studies would mean that, even if the replication does not find significant results as the original study, if the confidence interval includes the values found from the original study then the replication has failed but has not overturned the original results nor confirmed them, thus concluding that it is not evidence of a null effect (Maxwell et al., 2015).

The previous example introduces another way to test hypotheses that is suggested to be another solution: equivalence testing. In equivalence testing it has to be chosen a range of values in which the effect can be considered null or of non-relevance, this is called equivalence region. Researchers must decide what that range is based on literature and on their experience. The equivalence region will be referred to the population, so when collecting the data, it is not enough to estimate the effect size of the sample, but it has to be calculated the confidence interval to take sample variability into account and to know the possible values in the

population. The calculated confidence interval can fall into one of three categories (Maxwell et al., 2015):

- It falls all in the equivalence region: in this case it is quite sure that the true effect size of the population is not above the equivalence region so the effect can be considered null or of non-relevance.
- It falls entirely outside the equivalence region: in this case it is quite sure that the effect is not near to zero since the true effect of the population probably is outside the equivalence region. This result is statistically significant, and it is considered to have a practical and theoretical relevance.
- It overlaps the equivalence region: in this case the effect could be considered as null or as non-null, so the result is ambiguous because it does not support any of the options.

A big downside of equivalence testing is that to be adequately powered the sample size has to be much larger than the one in an adequately powered significance test (Maxwell et al., 2015). Another solution that has the same downside (Anderson, 2019) but allows to calculate effectively a population effect size is Accuracy In Parameter Estimation (AIPE; Kelley and Rausch, 2006; Maxwell, Kelley and Rausch, 2008). In AIPE researchers plan the sample size to estimate a population effect size with a predefined precision, in other words: to have an estimated sample effect size with a confidence interval of a certain width.

2.1.1. Focusing on power

Despite the fact that suggested solutions don't consider power explicitly, it does not mean that it will not be considered anymore; in fact, there are proposed solution in regard to power calculation and ways to try to avoid some misinterpretations and misuses of this method.

One of the simplest things to make power calculations better and more precise is distinguishing between conditional power and predictive power. The former is calculated with an effect size that is presumed to be already known with

certainty, so it does not take into account uncertainty around that value, the latter is calculated taking into account the uncertainty around the value and averages the power around the plausible values of the effect size, and it also requires much larger samples (Maxwell et al., 2015).

A really interesting and simple way to address mistakes in power calculations due to uncertainty around the estimated effect size has been proposed by Perugini, Gallucci and Costantini (2014). This method has been called “Safeguard power” and they defined it as: “the power calculated from an estimate such that it provides a given likelihood (i.e., 95%, 80%) of being correct—that is, a given likelihood that the true effect size is still compatible with the sample effect size, even assuming a worst-case scenario.” Their aim is to include uncertainty in the calculations of power in replications to protect from overestimation of effect size in original studies, that is more frequent and dangerous than underestimation.

Table 2.1: Safeguard power applied to some hypothetical data.

Study	d_0	d_{s80}	N_0	N_p	N_{s80}	$N_{2.5}$	SSR ₈₀
A	0.4	0.21	80	156	570	200	7.12
B	0.7	0.45	50	52	124	125	2.48
C	0.7	0.58	200	52	76	500	0.38
D	1.0	0.66	30	28	58	75	1.93
E	1.5	1.05	20	14	24	50	1.2

Notes: d_0 is the effect size of the original study; d_{s80} is the considered effect size for safeguard power calculations at an 80% level of protection; N_0 is the sample size of the original study; N_p is the sample size calculated from standard power analysis; N_{s80} is the sample size calculated from a safeguard power analysis at an 80% level of protection; $N_{2.5}$ is the sample size calculated from the 2.5 rule (2.5 times the sample of the original study); SSR₈₀ is the ratio between N_{s80} and N_0 (N_{s80}/N_0). Adapted from Perugini et al. (2015).

To calculate safeguard power it is used the lower boundary of the confidence interval of the sample effect size and the sample size of the prior study (refer to the original study for the table for calculations). Since larger sample size means narrower (thus, more accurate) confidence intervals of the effect size, this method considers the precision of the original study to adjust the sample size for the replication study; so, if the original study has small samples and small effect size, in the replication is needed a larger sample size, and if the original study has large samples and large effect size, the sample size for the replication would be similar to the one suggested by the standard power analysis and could even be lower than the original study's sample (*Table 2.1*; Perugini et al., 2014).

As you can see in *table 2.1*, standard power analysis calculates smaller samples than safeguard power analysis because it is not sensitive to the precision of the estimate of the effect size. The situation is even worse for the 2.5 rule that is dependant only on the sample size of the original study but not to effect size, leading to extremely large sample size in some cases (Perugini et al., 2014). This method can be really useful when designing a replication study, but it does not give any information about the original study, whether it was underpowered or not and the consequences on the effect size estimates, and it is not applicable to the replication study itself. For these reasons and because it can be adjusted and adapted based on everyone's needs, we suggest and further investigate design analysis (Gelman and Carlin, 2014).

2.2. Design analysis and Type *M*, Type *S* errors

Design analysis is considered an extension of standard power analysis that adds quantities of direction and size of the effect's estimate to type I and type II errors and power (Gelman and Carlin, 2014). In the NHST framework Type I error (α) is the probability to refuse the null hypothesis (H_0) when it is true, this is considered a false positive since an effect is found where there is no real effect, and it is the most considered error (Cohen, 1962). Consequently, the opposite of type I error ($1 - \alpha$) is the probability of making the correct decision of non-refuting H_0 if it is true.

Type II error (β) is the probability of non-refuting the null hypothesis when it is false, this is considered a false negative since no effect is found even if there is actually one, and this error is not taken much into consideration. Its opposite ($1 - \beta$) is the statistical power, hence, as already mentioned, the probability of refuting H_0 when it is false. The next table (*Table 2.2*) is a visual representation of what just explained, and it may makes it easier to understand.

Table 2.2: Type I and II errors and power.

		True state of the world	
		H_0 is true	H_0 is false
Decision	Non-refuting H_0	(1 - α)	Type II error (β)
	Refuting H_0	Type I error (α)	(1 - β) Statistical Power

Notes: for simplicity this table is adapted to the NHST framework since the original would be Neyman-Pearson Framework where the names of the hypothesis and the decisions would be defined in a different way. For further details refer to Bertoldo (2019) and to Gigerenzer, Krauss and Vitouch (2004).

The aim of design analysis is to provide clearer information on “the dangers of erroneous findings in small studies, [...] (since it) is not sufficiently well understood that “significant” findings from studies that are underpowered (with respect to the true effect size) are likely to produce wrong answers, both in terms of the direction and magnitude of the effect.” (Gelman and Carlin, 2014). The main differences with traditional power calculations are the calculation of inferential risks Type *M* (magnitude) and Type *S* (sign) errors and the so called “plausible effect size”. This method can be used prospectively (to plan a new study or a replication study) and retrospectively (to interpret in a better way already conducted studies’ results). Plausible effect size and prospective and retrospective analyses

will be explained in the next paragraphs, for now we will concentrate on type M and type S errors.

Type M error, or Exaggeration Ratio, indicates the expected overestimation of the plausible effect size by a statistically significant effect (Gelman and Carlin, 2014; Altoè et al., 2020). Type S error is the probability of finding a significant effect in the wrong direction, in other word, an effect with the opposite sign to the plausible effect size (Gelman and Carlin, 2014; Altoè et al., 2020; Bertoldo, Callagher and Altoè, 2021). To estimate these parameters, scientists need to use a plausible effect size, but let's see first how they are computed (Bertoldo et al., 2021).

Figure 2.1: Type M (exaggeration ratio) and type S errors as functions of power. Reprinted from Gelman and Carlin, 2004).

Imagine conducting the same study thousands of times with different samples from the same population, and for each sample register the effect size and the p -value. The type M error is the ratio between the mean of all significant effect sizes in absolute value and the plausible effect size. The type S error is the ratio between the number of significant effect size that have the opposite sign of the plausible effect and the number of total significant results, in other words it is a

simple proportion. Using statistical significance with $\alpha = 0.05$ can lead to a 40% probability to find a result in the wrong direction (Gelman and Carlin, 2014). These errors are additional information to statistical power, and both increase when sample size and/or effect size decrease (in *Figure 2.1* you can see how type *M* and *S* errors change as a function of power). If we consider type I error, the situation might change: if you decrease α (i.e.: 0.01 instead of 0.05), even if you are making your significance test more conservative, in fact type *S* error decreases because you are making it more difficult to find a result in the wrong direction, type *M* error increases because to find a significant effect with a lower *p-value* (maintaining sample size) the effect size has to be exaggerated even more (Bertoldo et al., 2021).

In this whole process of design calculations, the most important (and difficult) thing to do is to estimate the plausible effect size.

2.2.1. Plausible effect size

Gelman and Carlin (2014) claim that researchers usually specify the effect size for power calculations basing on empirical estimates of previous studies or from data at hand, but this is not a reliable way to do it since, as we already mentioned, effect size is often overestimated because of publication bias and underpowered studies. Underestimating an effect size leads to a power gain that is lower than the power loss of an overestimated effect, and this happens because the relation between power and effect size is not linear (Maxwell et al., 2015). Effect sizes will be inaccurate and grossly misleading if statistical significance is the only way to value the importance of a study (Flora, 2020). In fact, as we can see from the following example (*Figure 2.2*), *p-values* found in studies with small sample sizes are smaller than those found with bigger sample sizes, and this goes against the belief that a significant effect found with a small sample would be significant or even more relevant with a bigger sample (Simmons et al., 2011), in other words: the “what does not kill statistical significance makes it stronger” fallacy. For Asendorpf et al. (2013) and Flora (2020) research should concentrate on effect size as the primary

result rather than the existence or not of the effect, and the effect should be contextualised and interpreted carefully.

Figure 2.2: Simulation of p -values under H_0 found by adding an observation to each of two conditions and testing for significance after every addition, demonstrating that what would be significant with small sample sizes would not necessarily be significant with bigger sample sizes. The red dotted line corresponds to a p -value equal to 0.05. Reprinted from Simmons et al. (2011).

Sometimes researchers use pilot studies to have an estimate of the effect size to do power calculations, but according to Westfall (2015), Leon, Davis and Kreamer (2011) and Altoè et al. (2020) these kinds of studies are often underpowered; thus, their estimate of the effect size is not precise at all, leading to erroneous levels of power. To some it may seem paradoxical to need an estimate of the effect size to calculate power before the study is conducted, since conducting the study is necessary to have that estimate (Altoè et al., 2020) but “any procedure

that does not base sample size on an estimated or theoretical value of population effect size may result in either an underpowered or overpowered study.” (Anderson, 2019).

The effect size is a measure that indicates the magnitude of the association between two or more variables, and it is an estimation of population effect from sample data (Kelley and Preacher, 2012). “The plausible effect size refers to what could be approximately the true value of the parameter in the population” (Altoè et al., 2020) or “that which would be observed in a hypothetical infinitely large sample” (Gelman and Carlin, 2014). Clearly researchers don’t know this value, but they can estimate it in different ways. Usually, researcher estimate effect size for their study in two ways; one of them is empirically by using the same effect size of a similar study previously conducted or by the data at hand (Gelman and Carlin, 2014). As mentioned several times in this work and claimed by Gelman and Carlin (2014) too, studies or preliminary data are often underpowered and effect sizes estimated with small sample size are often misleading and overestimated (*Figure 2.3*).

So how do you define a plausible effect size if “published effect size estimates may often be overestimated to the point of providing no guidance to true effect size” (Gelman and Carlin, 2014)? There is not an exact answer neither a right one, but there are several recommendations and some new methods that help researchers on this. Meta-analysis, systematic literature review and all the evidence external to your study and data are some of the ways Gelman and Carlin (2014) suggest, furthermore, they recommend researchers to consider a range of plausible effect size and to be aware that published estimates are fallacious. Altoè et al. (2020) added that researchers should be transparent about any decision on the choice of a plausible effect size so that replications can be planned solidly, and results of the original study can be evaluated in a better way. There are other ways to find a plausible effect size, such as setting a smallest effect size of interest (SESOI; Lakens, Scheel and Isager, 2018), bias- and uncertainty-corrected sample size (BUCSS; Anderson, Kelley and Maxwell, 2017) and expert knowledge elicitation techniques (O’Hagan, 2019).

Figure 2.3: An example of “the winner’s curse”. A researcher studying a plausible effect size of $d = .20$ (vertical dashed line) with 33 participants, a two-tailed independent sample t -test, a $\alpha = .05$ (vertical continuous lines) and a 13% power. In this case it is not possible to find a significant result without overestimating the effect size, since the p -value would be $< .05$ only with effects bigger than $d = |0.49|$. Reprinted from Altoè et al. (2020).

2.2.2. Prospective and retrospective design analysis

As you may know, retrospective power calculations have been widely criticised because using collected data can lead to bad estimates of power, and in some cases a huge overestimation of it (Hoenig and Haisey, 2001; Gelman, 2019). This “method” is known as “post-hoc power” and it consists in calculating the power of a study using the observed effect size. Post-hoc power calculations are often used as an excuse if the results of a study are not significant and the researcher expected them to be significant (Hoenig and Haisey, 2001; Anderson, 2019; Dziak et al., 2018), but for a nonsignificant result power will always be less than 50% (*Figure*

2.4 on the right; Goodman, 1992; Hoenig and Haisey, 2001). Calculating post-hoc power adds nothing to the results of the study because the power is determined by the *p-value* (Figure 2.4), thus, it can't lead to the conclusion that a study had enough power but did not find a significant result: the observed power should not change the interpretation of the results or the opinion whether the topic was worthwhile or not (Hoenig and Haisey, 2001). The same is done to demonstrate that a study has enough power and has been design appropriately when a study's results are significant, but in both cases (a significant or a non-significant result) such explanation is simply circular reasoning (Dziak et al., 2018; Anderson, 2019). Dziak et al. (2018) gave a great explanation of the heuristic behind this practice (*PHP* stands for post-hoc power):

“Heuristically, power is a characteristic of the sampling distribution (i.e., distribution of possible samples), rather than of a particular sample. One could think of each sample of size n as a marble in a jar, marked with the *p-value* that would be calculated from this particular sample. The difficulty is that, regardless of the number n of participants inside the sample, the number of samples an investigator has is generally 1. Thus, the user of *PHP* is in the unenviable position of having to make an inference about the jar of marbles, using information from only a single marble. In such a situation, the best the investigator could do would be to assume that whatever kind of marble was drawn must have been from the majority. Thus, if the observed *p-value* is less than α , then a good provisional guess is that most of the *p-values* are less than α . If the observed *p-value* is greater than α , then a good provisional guess is that most of the *p-values* are greater. Following this logic, if the observed *p-value* is equal to α , the best guess would be that half of the *p-values* are larger than α and half are smaller than α . Thus, a *p-value* equal to α (e.g., .05) should lead to *PHP* of exactly 50%. Because the *PHP* is really the *p-value* in disguise, measuring the importance of a finding via the *PHP* is equivalent to measuring its importance via the *p-value*, a practice which is widely criticized [...].”

Gelman and Carlin (2014) explain that they agree with such critiques but still consider retrospective design analysis a good way to evaluate statistically

significant result from conducted studies. This because their interest is not on statistical significance but on magnitude and sign of the estimated effect size and, even more important, the effect size used for these calculations is not estimated from the data at hand but is an estimate from external data and systematic literature reviews. In other words, retrospective design analysis is a way to find out the power and type M and type S errors of a conducted research by using a plausible effect size that you consider to be true, so that you can evaluate more efficiently the result of that study and make your own considerations about it.

Figure 2.4: Observed (post-hoc) power as a function of p -value, showing that a p -value of .05 equals to about 50% power. On the left: simulation of 500 t -tests with 79.5% power to detect a 0.5 effect size, reprinted from Gelman (2019). On the right: one-tailed z test, $\alpha = 0.05$, reprinted from Hoenig and Haisey (2001).

So, if you need to evaluate results of a conducted study, the first thing to do is define a plausible effect size, then power, type M and type S errors are calculated;

it is always better to evaluate a study using more than one plausible effect size so that you consider more than one scenery to have a better idea. Retrospective design analysis can be used on your own study after you have conducted it, to evaluate retrospectively your results too.

While retrospective design analysis has some differences with retrospective power analysis, prospective design analysis is quite similar to prospective power analysis. In fact, it has the same purpose of suggesting a sample size with a given power, but adding some information, such as type M and type S errors, that help researchers to evaluate better the consequences of variations of the sample size on their research. In this case too, is better to consider different scenarios.

Table 2.3: Steps to conduct design analyses.

Prospective design analysis	Retrospective design analysis
Hypothesize a plausible effect size	Hypothesize a plausible effect size
Choose α level and desired power	Find out α level and sample size
Calculate type M and S errors and power	Calculate type M and S errors and power
Consider different scenarios and choose your sample size	Consider different scenarios
Collect data and run statistical analyses	Evaluate and interpret results
Retrospective design analysis	

Plausible effect size is nothing more than an assumption, especially when the object of study is new and not much research has been done, but Gelman and Carlin (2014) stated that design analysis is still very useful because, first of all, “even a rough prior guess can provide guidance” (Gelman and Carlin, 2014), secondly, design analysis leads researchers to engage more with the literature on the subject they are studying, thirdly, design analysis forces researches to find a plausible effect size, helping to specify the research questions. These kinds of

benefits are really useful in psychology, since psychologic research has usually small samples with a lot of variability and uncontrolled variation and effect sizes not well defined (Gelman and Carlin, 2014). Anyways, design analysis should not be an automatic problem solver, but a way to make researchers think about what they are doing and to explore different case scenarios to better understand what they are studying, and to have more collaboration between psychologists and statisticians (Altoè et al., 2020).

Before applying design analysis on a case study, we will introduce a way to do all the calculations.

2.3. *PRDA*

PRDA is an *R* package for Prospective and Retrospective Design Analysis (Callagher et al., 2021). This package includes calculation for standardised effect sizes (Cohen's d and Pearson's correlation coefficient). In this package there is also the possibility of defining the effect size not as a single value but as a distribution of plausible values, so that you can take into account uncertainty (refer to Altoè et al., 2020 for further details).

PRDA contains two functions: `retrospective()` and `prospective()`. In both functions you can modify the α level, the test you want to apply and other more technical values to fit your needs (for a complete explanation refer to: <https://cran.r-project.org/web/packages/PRDA/PRDA.pdf>).

Before applying the functions on the case study, let's see how these two functions work with a simple example. This first example is a retrospective design analysis. Let's assume we are trying to evaluate a study with a plausible Cohen's d of .30, the sample sizes are 25 subjects for the first sample and 30 for the second one, the test is a two-tailed test with an α of .10. In the output there are some information about the input, the power and the inferential risks (type M and type S errors), and the critical d values for a significant effect. The function and the output would look like this:

```
retrospective(effect_size = .30, sample_n1 = 25, sample_n2 = 30,
              test_method = "two_sample", alternative = "two_sided",
              sig_level = .10)

##      Design Analysis
##
## Hypothesized effect:  cohen_d = 0.3
##
## Study characteristics:
##   test_method  sample_n1  sample_n2  alternative  sig_level  df
##   two_sample    25        30        two_sided    0.1        53
##
## Inferential risks:
##   power  typeM  typeS
##   0.289  2.072  0.011
##
## Critical value(s): cohen_d = ± 0.453
```

As you can see, this can also be an example of the winner's curse represented in *Figure 2.3*. The arguments required from these functions depend on what test you are going to apply, other arguments are set by default to a certain value and it is optional to change them. For `retrospective()` effect size and the first sample size are always required, and for `prospective()` effect size and power are always required. In the next example we will see a prospective design analysis.

```
prospective(effect_size = .30, power = .80, test_method = "pearson",
            display_message = FALSE)

##      Design Analysis
##
## Hypothesized effect:  rho = 0.3
##
## Study characteristics:
##   test_method  sample_n1  sample_n2  alternative  sig_level  df
##   pearson      84        NULL        two_sided    0.05        82
##
## Inferential risks:
##   power  typeM  typeS
##   0.804  1.114  0
##
## Critical value(s): rho = ± 0.215
```

In this case we assume we are planning a study with a plausible correlation of $r = .30$ and a desired power of .80 (the α is considered to be .05; the default

setting). As in the retrospective case, the output includes information about the input, the power and the inferential risks, and the critical r values.

Applying these functions, you are able to perform design analysis to plan your research or to evaluate results of researches in literature. Now we will apply them to a real case study.

2.4. Case study

The study on which we will do a design analysis is a study that has already been cited in the First Chapter. Rather than taking a research without power analysis, as it has already been done several times (Gelman and Carlin, 2014; Bertoldo, 2019; Altoè et al., 2020; Bertoldo et al., 2021) we will apply design analysis to see if the results are comparable to power calculations conducted in the original study, to add information about type M and type S errors and to evaluate if retrospective design analysis on results is coherent with prospective design analysis. The research we will look at is “No Replication, No Trust? How Low Replicability Influences Trust in Psychology” by Wingen et al. (2020). In particular, we will consider study 1 and study 5 of the five carried out studies. In both studies, the hypothesis is one-sided, so type S error will always be 0 since there is no way to find results with the wrong sign.

In study 1 they wanted to see if the belief about replication rate in psychology in non-scientists is correlated to the trust they place in the field. This was done by giving people a brief introduction about the replication crisis in psychology referring to the “Reproducibility Project: Psychology” and they had to take a guess on how many researches out of 100 could be replicated. Wingen et al. (2020) conducted an a priori power analysis for a 95% power considering an effect of $r = .2$ (considered typical for similar social psychological research; Richard, Bond, and Stokes-Zoota, 2003) and a one-sided α of .05 leading to a suggested sample of 266 participants. It was selected a sample of 271, and after a sensitivity analysis was found an 80% power (one-sided $\alpha = .05$) to find an $r = .15$ and a 95%

power (one-sided $\alpha = .05$) to find an $r = .2$. Results confirmed the hypothesis with an $r(268) = .329, p < .001$ and confidence interval $[.218, .431]$.

We will test the a priori power analysis, the sensitivity analyses, and the results with the *PRDA* package¹. Conducting a prospective design analysis, the required sample to get a 95% power is exactly the 266 they calculated²; the type *M* error was only 1.04, indicating only a 4% exaggeration ratio of the effect size. A retrospective design analysis with their sample of 271 showed effectively a 95% power for a $r = .2$ with and a type *M* error of only 1.03, and an 80% power for an $r = .15$ with a type *M* error of 1.14, that is still an acceptable exaggeration ratio. In all the cases the critical value was clearly much lower than the effect they were considering. These design analyses suggest that Wingen et al.'s (2020) results of study 1 are robust.

In study 5 Wingen et al. (2020) assigned participants randomly to three conditions: a low replicability condition, a “now high” replicability condition and a “still low” replicability condition. The aim was to understand if public trust could be repaired by solving the low replication rate in psychology. In all the conditions participants read about the “Reproducibility Project: Psychology” learning the replication rate and that it was written in 2015. In the low replicability condition, no further information was given, in the other condition it was taught that *QRPs* were the cause; in the “still low” condition the open science movement and increased transparency in psychology did not solve the problem, and in the “now high” condition the problem has been addressed successfully.

Apparently, there is no prospective power analysis in study 5, but a sensitivity analysis has been conducted based on the sample they selected. Let's assume that the plausible effect size they are looking for can be estimated as $d = .2$ and a one-sided $\alpha = .05$. The `prospective()` function calculates a sample size of 942 subjects (314 per condition if considering them 2 by 2) for an 80% power, a

¹ For the code used for the calculations, please refer to Appendix A.

² `set.seed(122021)` in the code makes reproducible my exact results, without it results may vary due to approximation but remain comparable to the original ones.

1.14 type M error, and a 1692 sample size (564 per condition considering them 2 by 2) for a 95% power, a 1.02 type M error. These are quite large sample sizes, but the way they conducted the research could make it possible to reach these numbers. The sample they used was a total of 278 people (98 people in the low replicability condition, and 90 in the “still low” and “now high” conditions³), the sensitivity analyses they conducted are confirmed based on their sample, adding only a type M error of 1.15 in the 80% power case with a $d = .36$ and 1.04 in the 95% power case with a $d = .48$. Results in this study were not significant and the effect sizes they found are much smaller than they expected, and, as Wingen et al. (2020) themselves claimed: “Although sensitivity analyses showed that it is unlikely that the tested strategies have large trust-repairing effects, they also suggest that we had no sufficient power to rule out small, but potentially meaningful effects, which could only be detected with larger samples”. Applying a retrospective design analysis with the plausible effect size means finding exactly underpowered studies for such small effect sizes and with large exaggeration ratios.

Design analysis and *PRDA* gave comparable results that other power analyses gave regarding power, added information about the exaggeration ratio, and would have added information on the type S error too if the hypotheses were bi-directional. This gives the idea that design analysis can be reliable and can integrate standard power analysis, even though our example with a sample size of 2 studies is a bit too small to claim this...no, we don’t need a design analysis to know this. As already mentioned, this should make easier to researchers to evaluate researches and to plan new ones, having the possibility to double check too the results. In the next chapter we will further discuss design analysis presenting a new function to apply these analyses in the case of differences between proportions in a sample.

³ Data retrieved from the *R* data Wingen et al. (2020) uploaded on <https://osf.io/7vzmy/>.

Third Chapter

Difference between independent proportions

“A proportion is a special case of an arithmetic mean, one in which the measurement scale has only two possible values, zero for the absence of a characteristic and one for its presence.”
(Cohen, 1988)

In this chapter we will start with a little introduction on the difference between independent proportions and design analysis. Then we explain Cohen's h , an index to study and compare proportions in an analytical way, we will explain how he defined it and how to interpret it as an effect size and compare it to other effect sizes. We offer two functions that convert Cohen's h to normal proportions and vice versa. Then we will introduce our design analysis functions for independent proportion, showing an example and explaining briefly how they work and calculate the results. Lastly, we will apply them to a real case study.

3.1. Independent proportions and Design Analysis

Comparing independent proportions is quite common, since it is similar as comparing means. In fact Cohen (1988), as the citation at the beginning of the chapter says, defines proportions as special kind of means in which the

measurement scale has only two possible values; these values depend on the presence or absence of the studied characteristic and are 1 and 0, respectively. When you think about it as the difference between independent samples, you can easily use proportions as a correlational measure between variables. As it is for means, in the case of proportions is important too to understand whether the difference between them is because of a real difference between the two populations or it is simply due to chance in sampling.

Design analysis is a very useful way to evaluate conducted studies and to plan new ones. It helps researchers to understand the consequences of small sample sizes, low power and bad planning, thus, to understand whether the difference between means is effectively because of a difference between populations or not. One of the aims of this work is to extend design analysis' ability of helping evaluate and plan studies that compare means to studies that compare independent proportions.

Nowadays no specific function has been developed for this purpose. In our case, as we have done in the *PRDA* package, we chose a simulative approach and not an analytical one because, as it usually happens with inferential tests, the distribution under H_0 can be easy to identify throughout an analytical method, but the distribution under H_1 is more difficult to obtain. This is the analytical method to calculate the distribution under H_0 for the difference between independent proportions (Oakley, 2021):

$$Z = \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{P^*(1 - P^*) \left(\frac{1}{n} + \frac{1}{m} \right)}}$$

where

$$P^* = \frac{X + Y}{n + m}$$

Assuming X and Y to be the number of successes and n and m the sample size in the first and second sample respectively, and H_0 to be true, the Z-test statistics approximates to a normal distribution with mean equal to 0 and standard deviation equal to 1.

Since we are taking power into consideration, we need the distribution under H_1 , and the simulative approach makes this easier. Furthermore, we preferred this method so that our function could be modified and adapted for cases and indexes for which it is not possible to use an analytical approach.

Figure 3.1: The power of a difference of .05 between independent proportions with several sample sizes. The power clearly varies depending on where the difference is along the proportion scale. This has been calculated applying our retrospective function for difference between independent proportions to a million simulations every 5% difference. The R code for this plot is available in Appendix B.

In the case of proportions, unlike means, the same mathematical difference between two proportions has different power if it is situated in different points of the proportion scale, thus this unit does not provide a fixed effect size index. In other words, the difference between two independent proportions has different power depending on where along the scale it occurred (*Figure 3.1*). For this reason, different approaches exist to make evaluations on proportions. One of the approaches is Cohen's h (1988). As we will see, this index solves the problem of the fixed effect size index, but it is difficult to interpret without converting it back to raw proportions or without using other indexes. The other approach uses raw proportions, and the effect size is simply the difference between the two proportions taken into consideration and its position in the proportion scale. We chose the second approach to keep things more transparent and simpler for researchers to interpret. We also preferred this approach so that, in similar cases our functions can be easily modified to fit scenarios in which we know the distribution curve under H_0 that are not exactly the ones we thought of.

3.1.1. Cohen's h

Cohen (1988) claims that testing hypothesis about proportions by means of a normal curve test applied to the arcsine transformation of the proportions has its advantages when referring to power analysis, this will include the arcsine transformation and the effect size index: h . In other words, he applies a transformation to make equal differences on different positions of the proportion scale effectively equal. In even simpler words, the transformation flattens the curve in *Figure 3.1*.

One of the advantages of Cohen's h is that this index is independent on where the proportions are situated in the scale between 0 and 1. The arcsine transformation is a nonlinear transformation of the proportion that makes equal differences equally detectable and solves the problem of a solid effect size index that does not depend on its location in the range between 0 and 1 (Cohen, 1988).

To make thing clearer, Cohen (1988) defined what small, medium, and large effects are using this index and explained how to interpret it relating to other measures:

- Small effect size: $h = .2$ and it is a raw difference between proportions of .05 near the extremes and of .1 around the middle of the proportion scale. The difference reaches its maximum of .10 between .45 and .55 and it approaches 0 when one of the proportions approaches 0 or 1, in the .05 - .95 range its minimum value is .05. This value is equivalent to a correlation (Pearson's r) of about .10.
- Medium effect size: $h = .5$ and it is a raw difference between proportions of .20 to .25. The difference reaches its maximum of .248 between .376 and .624, within the .05 - .95 range reaches its minimum of .16 and between the .20 - .80 range it varies between .23 and .25. This value is equivalent to a correlation of about .25.
- Large effect size: $h = .80$ and it is a raw difference between proportions of .35 to .39 (between the .12 - .88 range). The difference reaches its maximum of .39 between .305 and .695, within the .05 - .95 range reaches its minimum of .293. This value is equivalent to a correlation between .37 and .39.

As already stated before, researchers must specify the effect size, not just define the effect they are studying as small, medium, or large. Anyways, as you can see and as already mentioned, it is not necessarily easy to interpret this index, for this reason in Appendix C we offer an *R* function to convert h to raw proportion difference and vice versa.

Cohen (1988) also offers table using this index to calculate power with given sample sizes and to calculate sample sizes given the desired power. In both calculation he defines different cases, one in which the size of the two samples is the same, the second in which the size of the two samples is different, and the third, that we will not discuss, in which we test if the proportion of a sample is equal to a specific value.

To calculate power, you have to refer to Cohen's tables (1988, p. 186-197) with the significance level (and if it is a one-tailed or a two-tailed test), the effect size (h), and the sample size of each sample. In the case of equal samples, the sample size in the table is the same of the number of both samples taken singularly, in the case of different sample sizes, the harmonic mean of the two sample sizes has to be calculated.

To calculate sample size, you still have to refer to Cohen's tables (1988, pp. 205-206) with the significance criterion (and if it is a one-tailed or a two-tailed test), the effect size (h), and desired power. With these data you can easily find the number of subjects for each sample in the case of equal sample sizes, otherwise, for samples with different sizes you have to apply a formula (Cohen, 1988, pp. 207) using the number of the fixed sample size and the suggested sample size from the table for the given power. There is also a method to use Cohen's (1988) tables for significance testing, but we will not look into it since presenting in a comprehensive way Cohen's h is not the aim of this work.

Cohen's h is analytic method to calculate power and sample size, and to compare independent proportions. As Cohen (1988), Holmes, Illowsky, and Dean (2017) and Oakley (2021) explain, there are other analytical ways to compare independent proportions, and these ways have all the same result (except for approximation) if the sample sizes are large enough; there are also simulative ways to compare proportions. Referring to design analysis, there is still no simulative approach for differences between independent proportions, so this is what we did, writing two functions on R that allow to conduct prospective and retrospective design analysis on independent proportions. These functions and this work will be available on the OSF (Open Science Framework) repository after graduation discussion. At the time, the functions' purpose is mainly didactic, once optimised and extended they will be included in the *PRDA* package.

3.2. Design analysis for difference between independent proportions

As just mentioned, there is no way to calculate design analysis with a simulative approach using proportions as index and effect size measure. The functions we offer make this possible without using Cohen's h but using directly proportions and sample sizes. This should help researchers to plan better studies and to evaluate already conducted studies in a simpler way than using analytic methods.

Using an analytical approach means to apply formulas and calculations to find out, for example, what your sample size should be to have the desired power, or what the power is if your sample has a fixed size. The difference with a simulative approach is that statistical software (in this case *R*) simulates the needed kind of data hundreds of thousands or millions of times, calculating all the required information not through classic formulas but using the simulated data. This is like running an experiment hundreds of thousands or millions of times and calculating the ratio of exaggerated effect size, of effect size in the wrong direction, power, all through the huge amount of data collected.

How we did for the *PRDA* package, let's see an example of a prospective and a retrospective design analysis function for the difference between independent proportions before explaining how they work and applying these functions to a real case study. The first example is a retrospective design analysis. As you can see, the function is quite simple: the `props` are the plausible proportions in the samples, and consequently the plausible effect size (`diff.prop.plau`), and `n` is the sample size of the samples. For `retrop()` the two proportions and sample sizes are always required, while the significance level is set by default to .05 but can be changed at any time.

The output includes the plausible proportions difference, that is simply the difference between the proportions in the input, the calculated power, and the inferential errors. In the following example let's assume we are considering a 10% plausible difference (.7 the first sample and .6 the second one) with a sample size respectively of 80 and 97 and an α of .01. In this case we find a study with low power (only 12%) and a high type *M* error, showing that the effect can be

overestimated up to 120%, but it is implausible for the effect to be estimated with the wrong sign.

```
retrop(props = c(.7, .6), n = c(80, 97), alpha = .01)
$diff.prop.plau
[1] 0.1

$power
[1] 0.118

$typeM
[1] 2.192

$typeS
[1] 0
```

The prospective design analysis is quite simple too. The function includes the `props` that are the same thing as in the retrospective function. For `propssp()` the two proportions are always required, while the significance level, the desired power, the sample size range, the tolerance and the number of iterations have a default setting⁴ but can be changed at any time, and if “`verbose`” is `TRUE` the output includes information about the calculations that are being done.

The output also includes the desired power we set in the input, the estimated power based on the sample size that the function calculated, the sample size for one group and the inferential errors. In the next example we assume a plausible proportion difference of .10 (.45 - .35), with an α of .1 and a desired power of 95%. We have also modified the range of possible sample sizes, the verbosity and the iterations for illustrative purposes, setting a higher tolerance to keep the example short. The function calculated 450 subjects per group, with an estimated power of 92.8% with such a plausible effect size. The type *S* error is 0 and the type *M* error is low, showing that in such cases the effect size would not be significantly overestimated.

```
propssp(props = c (.45, .35), alpha = .1, power = .95,
```

⁴ `alpha = .05`, `desired power = .80`, `sample range = 15 to 1000`, `tolerance = .01`, `iterations = 20` and `verbose = FALSE`.

```

    rangen = c (100, 1500), tol = .025, iter = 4, verbose = TRUE)
Estimating power with n per group = 800
Estimated power est_power = 0.993
Estimating power with n per group = 450
Estimated power est_power = 0.928
$power
[1] 0.95

$est_power
[1] 0.928

$n_group
[1] 450

$typeM
[1] 1.048

$typeS
[1] 0

```

In Appendix D you can find the whole code of the functions, and in Appendix E you can find the explanation on how the functions work and all their arguments.

These two functions work together, in fact the `propssp()` function needs the `retrop()` to work, and have 2 internal functions to help them work properly. Now we will explain how these functions work. Let's start from the first internal function, that is simply a function to approximate numbers in the correct way, rounding, for example, 2.5 to 3, since R does not approximate this way but would approximate 2.5 to 2. The second internal function will be explained within the retrospective one. Let's now see in a few steps how the retrospective function works:

1. The first thing that the function does is to find out the number of “successes” based on the proportion and on the sample size, to then verify if the first proportion is the greatest, if it is not, the function would not work so it notifies the user about it.
2. It then calculates the plausible difference and simulates data based on the proportions and sample sizes set by the user (under H_1), after the simulation it calculates the difference between the two simulated independent proportions.

3. In this third step, the second internal function comes into play. This function simulates the data under H_0 as if there was no difference between the two proportions set by the user, it does this by adding both “successes” and dividing by the sum of the sample sizes. The data are simulated with approximation to the normal and the upper and lower critical values are calculated (*Figure 3.2*).

Figure 3.2: 100,000 simulations of the difference between two samples ($n = n_1 = n_2 = 400$) with difference in proportions of .1 under H_0 (there is no difference between the two proportions); .5 for the first sample ($x_1 = 200$) and .4 for the second sample ($x_2 = 160$). The function calculates the exact critical values of the data and the p -value of the proportions set by the user. This is a visual representation of what the internal function in Step 3 calculates, the red lines are the critical values.

4. The upper and lower critical values are then used on the approximation to the normal of the simulated data under H_1 (Figure 3.3) to verify how the results compare to the data simulated under H_0 .

Figure 3.3: 100,000 simulations of the difference between two samples ($n = n_1 = n_2 = 400$) with difference in proportions of .1 under H_1 ; .5 for the first sample ($x_1 = 200$) and .4 for the second sample ($x_2 = 160$). This is a visual representation of what is simulated by the retrospective function, the red lines are the same that the internal function calculated, and all the data are then used in step 5.

5. With all these simulated data (1 million simulations set by default) it is now possible to calculate the mean power, the type M error, and the type S error in the way explained in the Second Chapter:
 - The power is calculated with a proportion between every time that the difference between proportions in a simulation is under the lower

critical value or over the upper critical value (in other words, every time that results are significant), and the total number of simulations.

- The type *M* error is calculated finding the mean, in absolute value, of the difference between independent proportions when they are significant divided by the plausible proportions. In other words, the significant effect size found in simulations divided by the plausible effect size set by the user.
 - The type *S* error is calculated by dividing the number of negative significant effect sizes by the number of significant results, and this is a proportion too.
6. The last step is just the output of the functions, specifying the plausible effect size set by the user, the power, and both errors.

The prospective function is a bit more complex, an internal function of `propssp()` is directly the retrospective function. Let's see how the prospective function works:

1. As in the `retrop()`, the first step is to verify if the first proportion is greater than the other, if not, the same output as the retrospective function is given to the user.
2. Then the function verifies if the maximum sample sizes is big enough to proceed to the calculations. This step is done by setting both sample sizes to the max and doing a retrospective design analysis with the proportion set by the user in the prospective function. If the maximum sample size is not big enough to obtain the desired power, the output tells the user to modify the range.
3. This step is where, if the maximum sample size is enough, calculations to find out the sample size based on the desired power and the give plausible effect size are made through the retrospective function. A technique called bisection is used; the median of the sample size range is taken and used for a retrospective design analysis, if the power with that sample size is not near enough (tolerance) to the desired power, the process is repeated bisecting

the upper or lower part of the first bisection based on, respectively, if the power was lower or larger than the desired one. This process is repeated until the function converges to a sample size that gives the desired power. For example, with a range between 10 and 30 (being 30 enough for the desired power), the bisection function finds 20 as the median and calculates the power applying, as said, a retrospective design analysis. If the calculated power with $n = 20$ is lower than the desired power, the bisection function selects the median between 20 and 30, if the power with $n = 20$ is higher than the desired power, the bisection function selects the median between 10 and 20. This process, as already mentioned, is repeated until a sample size that gives the desired power ($\pm$ tolerance) is found.

4. A counter makes sure that the process has an end: in some cases the bisection function is not able to find a sample size that gives the desired power $\pm$ tolerance, in these cases the counter makes sure that the process is repeated maximum 20 times (as default setting, but it can be changed at any time by the user). If this happens, the output will tell the user to change the iterations or the tolerance, but the user can choose to change the desired power too.
5. If the previous steps don't give any output, the function gives the final output with the estimated power, the sample size per group, the type M error and the type S error.

Since we are simulating binomial data, data are discrete and not all possible proportions can be studied or be observed. This means that it can happen that the function does not get to a definitive result (Step 3 or Step 4), and this happens with extremely small sample sizes (i.e.: $n \leq 15$ for each group). In these cases, as already mentioned, you can directly change the desired power if you prefer to have a different level of power but with the same precision, or you can modify tolerance, to get less (or more) precise power estimates around the desired power. This last option usually, but not always, leads to more iterations to find a sample size that gives you the desired power with the desired level of precision.

Now that we explained step by step how our new functions work, we can apply them to a real case study to show that they can be used exactly as the *PRDA* package and to simply give an example.

3.3. Case study

The example we are going to use is taken from a real case study (see Meuer and Woessner, 2018). This study is the second phase of an electronic monitoring (EM) study conducted to find out, first, whether EM is a helpful tool to avoid imprisonment and to prepare for release and rehabilitate (see Schwedler and Woessner, 2017), and second (Meuer and Woessner, 2018), a recidivism study to find out whether EM reduces recidivism after its use is terminated.

The study was conducted with two experimental groups and two control groups with volunteer subjects directly from some German prisons. The experimental groups had two different areas of investigation:

- The first group regarded home detention for up to six months before prison release. In this case, the subjects had to have a permanent residence, they had to stay at home except for some permitted activities such as work, therapy and to meet their probation officers.
- The second group, in the form of day parole, had an early work release and were supervised in halfway houses.

In both cases the subjects had to have a job for at least 20 hours a week, they had the EM on their ankle, and thanks to the GPS-operated monitoring device their location was supervised and limited with inclusion and exclusion zones. The control groups are subjects that served their sentence in prison until release.

Before continuing, we should clarify some relevant aspects regarding the considered study and the use of design analysis relative to it:

1. The purpose of the analysis applied to this study is merely illustrative.
2. The data considered are not collected by applying a true experimental design (the subjects are not randomized between conditions) and considerations

relating to the difficulty in selecting high sampling numbers should be taken into account given the particularities of this study.

3. We will only consider the two dichotomous variables indicated in the study (i.e.: experimental vs. control group and recidivists vs. not recidivists) without considering other potentially confounding variables.
4. We will consider only one of the two cases investigated by the researchers, namely that of the early work release.

Despite these limitations, the application of design analysis can provide important insights relatively to the interpretation of the obtained results (retrospective design analysis) and to future large-scale study planning (prospective design analysis). Anyways, getting back to the study, in *Table 3.1* you can see the results they found.

Table 3.1: Recidivism statistics collected by Meuer and Woessner (2018) for the early work release case.

	Control group	Experimental group
Sample size (<i>N</i>)	30	24
Recidivists	15	7
Recidivists (%)	50	29.2

Notes: Adapted from Meuer and Woessner (2018). The results were not found to be significant.

As it can be noted, the use of EM seems to have a positive effect (i.e.: recidivism is lower in the experimental condition). From the inferential point of view, their results are not statistically significant ($p = .12$). Even though design analysis is usually used to evaluate researches with significant results, while in this case the results are not significant, it can also be used to evaluate researches with

non-significant results and can still give important insights. Let's see what the contribute of design analysis can be with non-significant results too.

We will conduct a design analysis on their results assuming three different plausible effect sizes. The effect sizes we are going to assume plausible are a 15%, a 10% and a 5% difference between the control group and the experimental group. These effect sizes serve only as an example and are not based on extensive research on the object of study. Anyways, as we already mentioned, in the case of proportions power changes based on where the difference is situated on the proportion scale; this means that in the case of proportions it necessary not only to find a plausible difference between proportions, but a “plausible position on the proportion scale” too. We will assume that this “plausible position” is around the 40% area. Therefore, we will add and subtract half of the effect size from the starting point. In both retrospective and prospective design analyses we will use a .05 significance level and several levels of desired power.

3.3.2. Retrospective design analysis

In the case of retrospective design analysis, we are directly evaluating other researchers work knowing their results, in this case Meuer and Woessner's (2018). The sample sizes reported in *Table 3.1* are used to retrospectively evaluate the research using the plausible effect sizes we assumed, **NOT** the effect sizes they found in their research. Doing this will give you an idea on the power their research has and the type *M* and *S* error to better understand how misleading (or precise) their results can be. Since in *Figure 3.1* we already showed how power changes based on the positions on the proportion scale, thus sample size, type *M* error and type *S* error varies too; we will now use only the starting point mentioned a few lines above (40%), but the starting point can be modified if doing a sensitivity analysis to test several scenarios.

As you can see in *Table 3.2*, in all cases power is extremely low and type *M* errors are very high. Type *S* errors are too high too, in fact in the 10% plausible effect size it is too high to be considered acceptable, while in the 15% plausible

effect size it can be considered barely sufficient since it means that you would find an effect size in the wrong direction 1 out of 200 times; a Type *S* error more near to 0 would be far more recommended to consider results more solid.

Table 3.2: Retrospective design analyses of the early work release case.

Effect size		
5%	10%	15%
Power: 6.5%	Power: 11.4%	Power: 19.9%
Type <i>M</i> : 6.33	Type <i>M</i> : 3.25	Type <i>M</i> : 2.23
Type <i>S</i> : 15.5%	Type <i>S</i> : 3%	Type <i>S</i> : 0.5%

Notes: Code available in Appendix F.

In these cases, again, if the effect size were based on extensive external research and on expert knowledge elicitation techniques, we would say that the study's results are not solid at all, not credible nor reliable since the effect size has to be overestimated a lot to be significant, and there are high probabilities that the result are in the wrong direction too.

To better give the idea of the low power and the overestimation of the effect, for the retrospective analyses we will use the same histograms used explaining step by step the retrospective function (*Figure 3.2* and *Figure 3.3*) but adapted with their relative data simulation. In some cases (i.e.: the one with smaller plausible effect sizes) the differences are barely visible between the simulations under H_0 and under H_1 .

The 10% and 15% plausible effect size cases in the early work release are the most “powerful” cases (*Figure 3.5* and *Figure 3.4* respectively). In these two cases you can clearly see the difference between the simulations under H_0 and under H_1 , but in all three cases you can clearly see that the simulations under H_1 are not

above the critical values (that in all cases is around $|0.26|$). This critical value indicates that, to find a significant effect, you should find a 26% difference between proportions with such sample sizes, but the plausible effect size is much lower, so it would not be possible to find anything significant if not overestimating the effect size.

This demonstrates that there is no way to find significant results with such small sample sizes for such plausible effect sizes. Retrospective design analysis' purpose is to help researchers evaluating theirs or others researches, now let's see what prospective design analysis' purpose is.

Figure 3.4: 15% plausible effect size in the early working release case (1 million simulations). *Above:* the curve under H_0 with the critical values ($|0.267|$; $\alpha = .05$; red lines). *Below:* the curve under H_1 with the critical values. Code available in Appendix F.

Figure 3.5: 10% plausible effect size in the early working release case (1 million simulations). *Above:* the curve under H_0 with the critical values ($|0.267|$; $\alpha = .05$; red lines). *Below:* the curve under H_1 with the critical values. Code available in Appendix F.

Figure 3.6: 5% plausible effect size in the early working release case (1 million simulations). *Above:* the curve under H_0 with the critical values ($|0.267|$; $\alpha = .05$; red lines). *Below:* the curve under H_1 with the critical values. Code available in Appendix F.

3.3.3. Prospective design analysis

In the case of prospective design analysis, we have to imagine not to know the data or any other specific information about Meuer and Woessner's (2018) results. Here we will act as if we were planning the study and the only information we have is all the previous research from which we can estimate plausible effect sizes and "plausible positions on the proportion scale" and experts' opinion.

Conducting the three prospective design analyses you instantly realise that in some cases the function asks the user to set a larger range for sample size. This means that for some of these plausible effect sizes you need really big samples, and consider that we are using "only" an 80% power. Trying with a 95% power would mean getting even bigger sample size, so for the sake of comparing sample sizes, we tried with a 60% and a 70% power too.

In *Table 3.3* we can clearly see that sample sizes are, in some cases, quite large, except for lower power such as 60% or 70% for the bigger effect sizes. Obviously, the best option is the 80% power, or even the 95% power. In the case of early work release, being near to the middle of the proportion scale, the sample size required for the same amount of power and the same plausible effect size will be bigger than in cases that are more near to the extremes of the proportion scale.

If our plausible effect sizes were not assumptions only in order to make an example but were based on extensive research and on expert knowledge elicitation techniques, we could say that in Meuer and Woessner's (2018) research no prospective power calculations were made to select the sample size, otherwise they would have realised that their research would not have had a lot of meaning because of the imprecision of the measure of the effect size.

Prospective design analysis' purpose is to help researchers to plan better studies before they collect the data, not to evaluate other researches, but using it also in this way can help to better understand what other researches did and could have done while planning their studies, consequently helping to plan a better study. With the data we calculated we would have chosen a much larger sample size, and

these is a great reason to start worrying more about planning studies in the correct way. Using design analysis is a good and “easy” way to start doing it in a reasonable way rather than a mechanical way, since it leads to use reason thanks to the need of a plausible effect size. This does not mean that it requires reason, so it can be misused as much as all the other statistical methods, but our hope is that leading to use reason is enough to have better planning in future studies.

Table 3.3: Early work release prospective design analyses results with 60%, 70%, 80%, and 95% power and 5%, 10%, and 15% plausible effect size.

	Effect size		
	5% (37.5% - 42.5%)	10% (35% - 45%)	15% (32.5% - 47.5%)
60% power	N: 939	N: 232	N: 101
	Type M: 1.29	Type M: 1.29	Type M: 1.3
	Type S: 0	Type S: 0	Type S: 0
70% power	N: 1188	N: 286	N: 124
	Type M: 1.2	Type M: 1.2	Type M: 1.2
	Type S: 0	Type S: 0	Type S: 0
80% power	N: 1500	N: 370	N: 159
	Type M: 1.12	Type M: 1.12	Type M: 1.13
	Type S: 0	Type S: 0	Type S: 0
95% power	N: 2500	N: 631	N: 262
	Type M: 1.03	Type M: 1.03	Type M: 1.03
	Type S: 0	Type S: 0	Type S: 0

Notes: *N* = sample size. Sample size is referred to a single group, so total sample size equals sample size x2. Code available in Appendix G.

Fourth Chapter

Discussion and conclusions

*“On two occasions I have been asked, — "Pray,
Mr. Babbage, if you put into the machine wrong figures,
will the right answers come out?" [...] I am not able rightly to apprehend the kind of confusion
of ideas that could provoke such a question.”*
(Charles Babbage, 1864)

4.1. General discussion

With this work we tried to create greater awareness to the problem of underpowered studies and the misuse, voluntary or not, of statistical methods to find statistically significant effects. As seen in the first chapter, the reasons of the replicability crisis in psychology, and in science in general, can be multiple: publication practices, *QRPs*, replication misunderstanding and low power. Low power is a long-standing problem in psychology that nowadays has not been solved yet, even though in some journals power calculations are becoming a requirement for studies to be published. Here we concentrated the most on the low power aspect, suggesting design analysis as a possible method not only to help researchers to evaluate and plan studies, but also as a way to reason more about what they are doing and trying to find out,

becoming more conscious about the consequences of underpowered studies and on the effects and sample sizes they are working with.

Design analysis, with its inferential errors (type *M* and type *S*) and the requirement of a plausible effect size, leads researchers to reason more and study more thoroughly the topic they are addressing. Hopefully this is enough to persuade scientists to plan and conduct better research, even though it is surely not. In fact, with the current publication practices, some changes need to be done directly by journals, and design analysis cannot have a direct role in this.

While design analysis can help evaluate already completed researches and interpret results, it is really important the role it can have in planning new studies. Indeed, planning better studies is essential to stop the vicious cycle of planning studies based on other badly planned studies with exaggerated effect sizes and small samples. Samples calculated on exaggerated effects results in other underpowered studies making it look like as if the power is sufficient or even high, especially in the case of post-hoc power calculations, as mentioned in the second chapter. This is the reason why in design analysis a plausible effect size calculated from external data is used rather than the effect size calculated from the data at hand.

The purpose of research should not be statistical significance but finding out and understanding phenomena. Scientists need to get back to study phenomena rather than trying to publish at any cost, and journals should encourage and motivate this kind of behaviour. Surely it is not an easy thing to do, but it is what must be done. We are trying to give our little contribution by providing scientists new methods, approaches, and tools to actively take action in our attempt...one of many attempts made in psychological and, in general, in scientific literature.

Design analysis, as much as all the other possible solutions, is not effortless at all; to address such a problem in psychology, and in science, a lot of efforts must be made, and design analysis is only a small part of it. Thankfully nowadays we can use technology to do most of the calculations and help us improve in the complexity of statistical methods and tools used for these purposes; but this is at the same time a downside, since being able to do a huge number of analyses with small efforts leads to do all kind of analyses to find a significant effect. Technology is only a tool

that does exactly what we tell it to do, it does it million times faster and without any mistake (unless the mistake is in our inputs or in programming codes), but it does not make magics and the effort on our side must be done to have reliable and solid results: *garbage in, garbage out*.

In conclusion, being able to conduct design analysis on the difference between independent proportions too is a little step forward in giving the opportunity to scientists who want to undertake our own path. Design analysis, and transparent communication of the calculated inferential risks, are not the only and unique solutions, since pre-registration, transparency on the data and on the methods, the correct use of statistics and validity of the measurements are part of it too; but it is one of the ways we can make scientists more aware and reach our goal if enough of them undertake, in a way or another, the same purpose. Offering new and more tools to scientists makes it easier to “promote” this metascientific reasoning and knowledge that is much needed in the scientific world nowadays; in this work we referred in particular to psychology, but it is applicable to science in general.

4.2. Limits

The first limit we must mention is that our function has only didactic purposes, under the IT aspect it can be optimised. In fact, there are some repeated lines and processes to make each internal function autonomous that otherwise could be removed. The second and quite obvious limit of our work is that there is not yet a function to conduct design analysis on a single proportion rather than a difference between independent proportions.

As said several times in the third chapter, we simulated binomial data, because of the discrete nature of binomial data not all proportions and values can be studied or observed; this can lead to our function to not be able to calculate all the results or even to give an error as output. Both cases occur only with extremely small sample sizes (i.e.: $n \leq 15$ for each group).

Another limitation concerns the function for the prospective design analysis, in fact in this function there is only the option to have equal sample size for the two groups, at the moment there is not the option to set a fixed number for one group and calculate the necessary sample size of the other to reach the desired power level, but this will be surely part of a future update of our functions.

In this work we used quite simplified examples that might not represent the complexity of the cases in which most scientists would find themselves while planning and conducting research. We used such simple example to give the idea of how design analysis works, clarifying how to estimate plausible effect size and how to read and interpret type M and type S errors, how it can help scientists and how to use some of the functions to conduct design analysis. Furthermore, we considered only the case of plausible effect size as point value and did not consider the case of plausible effect size as intervals of plausible values with different kind of distributions within the interval (see Altoè et al., 2020). This deepening has been left out for its complexity not only in the explanation, but in the application too.

4.3. Further developments

Further developments are directly linked to some of the limits. In fact, as said, the first future development we can think about is enhance the function so that in the prospective design analysis is possible to set a group to a fixed sample size and calculate the sample size of the other group to reach the desired level of power. This enhancement would help a lot in cases where it is difficult to have a lot of participants in one group because of the characteristics needed to be part of it or when the subjects are difficult to reach.

Along with this, the development of a function able to calculate design analysis on a single proportion would be very useful to the scientific community. Even though we did not address the subject yet, Cohen in 1988 explaining the h index already covered the issue.

Furthermore, once the functions we offered in this work will be optimised and further developed and expanded, they will become part of the *PRDA* package

presented at the end of second chapter. At the moment is still independent from the package, and will be available on the OSF repository, along with this work, after graduation discussion. With the integration of our function in the *PRDA* package would be worthwhile to be able to integrate the possibility to estimate the plausible effect size as an interval with several kind of distributions within it.

Some interesting developments of this work could be some functions or interactive apps that considers only the plausible proportion difference without the “plausible position” on the proportion scale, giving as output the power, the type M , the type S , and the sample size as a function of the “plausible position” on the proportion scale as a graph or as the user interfaces the app. This can already be done by our function but not without manual intervention by the user (as in *Figure 3.1* and Appendix B). Intervention that sometimes can be somewhat time-consuming, especially if you have to consider not only power (as in *Figure 3.1*), but sample size, and both the inferential errors too.

Other developments that would be interesting to see are the implementation of functions for design analysis for even more indexes and for more complex cases, so that more scientists are able to conduct design analysis in several conditions, and apps that run on the web to use design analysis for those who are not able to use *R* or other statistical software based directly on coding.

References

- Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagni, A., Finos, L. and Pastore, M. (2020) Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Front. Psychol.* 10:2893. DOI: <https://doi.org/10.3389/fpsyg.2019.02893>
- Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., Cubelli, R. (2017) Questionable research practices among italian research psychologists. *PLoS ONE*, 12(3), e0172792. DOI: <https://doi.org/10.1371/journal.pone.0172792>
- Anderson, S. F., Kelley, K. and Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, 28 (11), 1547-1562. DOI: <https://doi.org/10.1177/0956797617723724>
- Anderson, S. F. (2019). Best (but oft forgotten) practices: Sample size planning for powerful studies. *The American Journal of Clinical Nutrition*, 110(2), 280–295. DOI: <https://doi.org/10.1093/ajcn/nqz058>
- Anvari, F. and Lakens, D. (2018) The replicability crisis and public trust in psychological science, *Comprehensive Results in Social Psychology*, 3:3, 266-286, DOI: <https://doi.org/10.1080/23743603.2019.1684822>
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27, 108–119. DOI: <http://dx.doi.org/10.1002/per.1919>
- Babbage, C., (1864). Passages from the life of a philosopher. Chapter 5 “Difference Engine No. 1”, pp. 67. Retrieved from: http://djm.cc/library/Passages_Life_of_a_Philosopher_Babbage_edited.pdf
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454. DOI: <https://doi.org/10.1038/533452a>
- Baker, M. and Dolgin, E. (2017) Cancer reproducibility project releases first results. *Nature*, 541(7637), 269–270. DOI: <https://doi.org/10.1038/541269a>

- Bertoldo, G. (2019, November 6). Dealing with the replication crisis in psychological science: The contribution of Type *M* and Type *S* errors. DOI: <https://doi.org/10.31237/osf.io/w63h7>
- Bertoldo, G., Callegher, C. Z., Altoè, G. (2021) Designing Studies and Evaluating Research Results: Type *M* and Type *S* Errors for Pearson Correlation Coefficient. Manuscript submitted for publication. Preprint DOI: <https://doi.org/10.31234/osf.io/q9f86>
- Broomell, S. B., and Kane, P. B. (2017). Public perception and communication of scientific uncertainty. *Journal of Experimental Psychology: General*, 146, 286–304. DOI: <https://doi.org/10.1037/xge0000260>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365–376. DOI: <https://doi.org/10.1038/nrn3475>
- Callegher, C. Z., Bertoldo, G., Toffalini, E., Vesely, A., Andreella, A., Pastore, M., Altoè, G. (2021). *PRDA*: An R package for Prospective and Retrospective Design Analysis. *Journal of Open Source Software*, 6(58), 2810. DOI: <https://doi.org/10.21105/joss.02810>
- Callegher, C. Z., Pastore, M., Andreella, A., Vesely, A., Toffalini, E., Bertoldo, G., Altoè, G. (2020). Conduct a Prospective or Retrospective Design Analysis. Retrieved from <https://cran.r-project.org/web/packages/PRDA/PRDA.pdf>
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Alemberg, M., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isakkson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016) Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436. DOI: <https://doi.org/10.1126/science.aaf0918>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, H. T., Huber, J., Johannesson, M., Kirchler, M., Nave Nosek, B., Pfeiffer, T., Altmejd, A., Buttrick, A., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L.,

- Imai, T., Isaksson, S., Manfredi, D., Rose, J. E., Wagenmakers, E. -J. and Wu, H. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644. DOI: <https://doi.org/10.1038/s41562-018-0399-z>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review *Journal of Abnormal & Social Psychology*, 65(3), 145-153. DOI: <https://doi.org/10.1037/h0045186>
- Cohen, J. (1970). Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement*, 30, 811-831. DOI: <https://doi.org/10.1177/001316447003000404>
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dziak, J. J., Dierker, L. C., Abar, B. (2018) The interpretation of statistical power after the data have been gathered. *Curr Psychol.* 1-8. DOI: <https://doi.org/10.1007/s12144-018-0018-1>
- Ebersole C. R., Axt J. R., Nosek B. A. (2016) Scientists' Reputations Are Based on Getting It Right, Not Being Right. *PLoS Biol*, 14(5): e1002460. DOI: <https://doi.org/10.1371/journal.pbio.1002460>
- Fiedler, K., Schwarz, N., (2016) Questionable research practices revisited. *Social Psychological and Personality Science* 7(1), 45-52. DOI: <https://doi.org/10.1177/1948550615612150>
- Flake, J. K., and Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. DOI: <https://doi.org/10.31234/osf.io/hs7wm>
- Fletcher, S. C. (2021) The role of replication in psychological science. *European Journal for Philosophy of Science*, 11(23). DOI: <https://doi.org/10.1007/s13194-020-00329-2>
- Flora, D. B. (2020). Thinking about effect sizes: From the replication crisis to a cumulative psychological science. *Canadian Psychology/Psychologie Canadienne*, 61(4), 318-330. DOI: <http://dx.doi.org/10.1037/cap0000218>

- Freedman, L. P., Cockburn, I. M., Simcoe, T. S. (2015) The Economics of reproducibility in preclinical research. *PLoS Biology*, 13(6): e1002165. DOI: <https://doi.org/10.1371/journal.pbio.1002165>
- Franco, A., Malhotra, N. and Simonovits, G. (2014) Publication bias in the social sciences: unlocking the file drawer. *Science*, 345, 1502–1505. DOI: <https://doi.org/10.1126/science.1255484>
- Gelman, A. (2013a). Too good to be true. *Slate*, 24 Jul. <https://slate.com/technology/2013/07/statistics-and-psychology-multiple-comparisons-give-spurious-results.html>
- [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. (pp. 1–17). Department of Statistics, Columbia University. <https://doi.org/dx.doi.org/10.1037/a0037714>
- Gelman, A. and Carlin, J. (2014). Beyond Power Calculations: Assessing Type *S* (Sign) and Type *M* (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651. DOI: <https://doi.org/10.1177/1745691614551642>
- Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American scientist*, 102(6), 460–466. DOI: <https://doi.org/10.1511/2014.111.460>
- Gelman, A. (2019). Don’t calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, 269, e9. DOI: <https://doi.org/10.1097/SLA.0000000000002908>
- Gigerenzer, G., Krauss, S., and Vitouch, O. (2004). “The null ritual: what you always wanted to know about significance testing but were afraid to ask”, in *The Sage Handbook of Quantitative Methodology for the Social Sciences*, ed D. Kaplan (Thousand Oaks, CA: Sage), 391–408. DOI: <https://doi.org/10.4135/9781412986311.n21>
- Goodman, S.N. (1992). A comment on replication, *p-values* and evidence. *Stat Med*. 11, 875-879. DOI: <https://doi.org/10.1002/sim.4780110705>
- Goodman, S. N., Berlin, J. A. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting

- results. *Ann. Intern. Med.* 121, 200-206. DOI: <https://doi.org/10.7326/0003-4819-121-3-199408010-00008>
- Hirshleifer, D., Teoh, S. H. (2003) Herd behaviour and cascading in capital markets: A review and synthesis. *Eur. Financ. Manage.* 9, 25-66. DOI: <https://doi.org/10.1111/1468-036X.00207>
- Hendriks, F., Kienhues, D. and Bromme, R. (2020) Replication crisis = trust crisis? The effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Understanding of Science* 29, 270–288. DOI: <https://doi.org/10.1177/0963662520902383>
- Hoenig, J. N., and Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55, 19-24. DOI: <https://doi.org/10.1198/000313001300339897>
- Holmes, A., Illowsky, B., and Dean, S. (November 29th, 2017). Introductory Business Statistics, Chapter 10.4 Comparing Two Independent Population Proportions, OpenStax. Retrieved from: <https://openstax.org/books/introductory-business-statistics/pages/10-4-comparing-two-independent-population-proportions>
- Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLoS Med* 2(8): e124. DOI: <https://doi.org/10.1371/journal.pmed.0020124>
- John, L. K., Loewenstein, G., Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. DOI: <https://doi.org/10.1177/0956797611430953>
- Kelley, K., Rausch, J. R. (2006) Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. *Psychol Methods*, 11:363–385. DOI: <https://doi.org/10.1037/1082-989X.11.4.363>
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17, 137–152. DOI: <http://dx.doi.org/10.1037/a0028086>

- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196–217.
https://doi.org/10.1207/s15327957pspr0203_4
- Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3), 221–230. DOI: https://doi.org/10.24602/sjpr.62.3_221
- Lakens, D., Scheel, A. M. & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. DOI: <https://doi.org/10.1177/2515245918770963>
- Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, 45, 626–629. DOI: <http://dx.doi.org/10.1016/j.jpsychires.2010.10.008>
- Lilienfeld, S. O., and Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, 61(4), 281–288. DOI: <https://doi.org/10.1037/cap0000236>
- Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585. <https://doi.org/10.1126/science.aal3618>
- Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, 8, 750–751. DOI: <http://doi.org/10.1037/h0063675>
- Makel, M., Plucker, J., Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science*, 7, 537–542. DOI: <https://doi.org/10.1177/1745691612460688>
- Masicampo, E.J., Lalande, D.R. (2012) A peculiar prevalence of *p* values just below .05. *Q J Exp Psychol*, 65:2271–9. DOI: <https://doi.org/10.1080/17470218.2012.711335>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9, 147–163. DOI: <https://doi.org/10.1037/1082-989X.9.2.147>
- Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of*

- Psychology*, 59, 537–563. DOI:
<http://dx.doi.org/10.1146/annurev.psych.59.103006.093735>
- Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70(6), 487–498. <https://doi.org/10.1037/a0039400>
- Mede, N. G., Schafer, M. S., Ziegler, R., Weisskopf, M. (2020). The “replication crisis” in the public eye: Germans’ awareness and perceptions of the (ir)reproducibility of scientific research. *Public Understanding of Science*, 30(1), 91–102. <https://doi.org/10.1177/0963662520954370>
- Meuer, K. and Woessner, G. (2018). Does electronic monitoring as a means of release preparation reduce subsequent recidivism? A randomized controlled trial in Germany. *European Journal of Criminology*, 1-22. DOI:
<https://doi.org/10.1177/1477370818809663>
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1 (1), 0021, DOI:
<https://doi.org/10.1038/s41562-016-0021>
- Nosek, B. A., Spies, J. R. and Motyl, M. (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7, 615–631. DOI:
<https://doi.org/10.1177/1745691612459058>
- Nuijten, M. B., Bakker, M., Maassen, E., Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences*, 41, e143. DOI: <https://doi.org/10.1017/s0140525x18000791>
- Oakley, J. (February 17th, 2021). MAS113 Part 2: Data Science, Chapter 8 Hypothesis testing: comparing two proportions. Retrieved from:
<http://www.jeremy-oakley.staff.shef.ac.uk/mas113/notes/hypothesis-testing-comparing-two-proportions.html>
- O’Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, 73, 69–81. DOI:
<https://doi.org/10.1080/00031305.2018.1518265>

- Open Science Collaboration, *Science* 349, aac4716 (2015). DOI: <https://doi.org/10.1126/science.aac4716>
- Perugini, M. (2014). La crisi internazionale di credibilità della psicologia come un'opportunità di crescita. *Giornale italiano di psicologia*, 41, 23-46. DOI: <https://doi.org/10.1421/77194>
- Perugini, M., Gallucci, M., and Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9, 319– 332. DOI: <https://doi.org/10.1177/1745691614528519>
- Poole, C. (1987) Beyond the confidence interval. *Am J Public Health*. 77, 195-199. DOI: <https://doi.org/10.2105/ajph.77.2.195>
- Richard, F. D., Bond, C. F., Jr, & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7, 331–363. DOI: <https://doi.org/10.1037/1089-2680.7.4.331>
- Sedlmeier, P. and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2), 309-316. DOI: <https://doi.org/10.1037/10109-032>
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, *Psychological Science*, 22, 1359–1366. DOI: <https://doi.org/10.1177/0956797611417632>
- Schwedler, A., and Woessner, G. (2017). Identifying the rehabilitative potential of electronically monitored release preparation. A randomized controlled study in Germany. *International Journal of Offender Therapy and Comparative Criminology*, 61(8): 839–856. DOI: <https://doi.org/10.1177/0306624X15612060>
- Szucs, D., Ioannidis, J. P. A. (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol*, 15(3): e2000797. DOI: <https://doi.org/10.1371/journal.pbio.2000797>

- Taylor, D. J., and Muller, K. E. (1996). Bias in linear model power and sample size calculations due to estimating noncentrality. *Communications in Statistics: Theory and Methods*, 25, 1–14. DOI: <https://doi.org/10.1080/03610929608831787>
- Tversky, A. and Kahneman, D. (1971) Belief in the law of small numbers. *Psychol. Bull.* 75, 105–110. DOI: <https://doi.org/10.1037/h0031322>
- Westfall, J. (June 22nd, 2015). Don't fight the power (analysis). Retrieved from: <http://jakewestfall.org/blog/index.php/2015/06/16/dont-fight-the-power-analysis/>
- Wingen, T., Berkessel, J., and English, B. (2020). No replication, no trust? How Low Replicability Influences Trust in Psychology. *Social Psychological and Personality Science*, 11(4), 454-463. DOI: <https://doi.org/10.1177/1948550619877412>
- Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, 5, 1418–1422. DOI: <https://doi.org/10.1371/journal.pmed.0050201>
- Zwaan, R. A., Etz, A., Lucas, R. E., Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41, e120. DOI: <https://doi.org/10.1017/S0140525X17001972>

Appendices

Appendix A: Code for prospective and retrospective design analysis applied to Wingen et al. (2020).

```
install.packages("PRDA")
library(PRDA)

# Study 1

set.seed(122021) # set seed to make results reproducible
prospective(effect_size = .2, power = .95, test_method = "pearson",
             alternative = "greater")
set.seed(122021)
retrospective(effect_size = .2, sample_n1 = 271,
              test_method = "pearson", alternative = "greater")
set.seed(122021)
retrospective(effect_size = .15, sample_n1 = 271,
              test_method = "pearson", alternative = "greater")

# Study 5

set.seed(122021)
prospective(.2, .80, test_method = "two_sample",
            alternative = "greater")
set.seed(122021)
prospective(.2, .95, test_method = "two_sample",
            alternative = "greater")
set.seed(122021)
retrospective(.36, 98, 90, test_method = "two_sample",
              alternative = "greater")
set.seed(122021)
retrospective(.48, 98, 90, test_method = "two_sample",
              alternative = "greater")
set.seed(122021)
retrospective(.2, 98, 90, test_method = "two_sample",
              alternative = "greater")
set.seed(122021)
retrospective(.2, 98, 90, test_method = "two_sample",
```

```
alternative = "greater")
```


Appendix B: Code for the plot about the power of a difference of .05 between independent proportions with several sample sizes (Figure 3.1).

```
a<-seq(0.05,1,.05)
b<-seq(0,.95,.05)
num0<-c(800,800)
vet.pow0<-rep(NA,length(a))
num1<-c(450,450)
vet.pow1<-rep(NA,length(a))
num2<-c(250,250)
vet.pow2<-rep(NA,length(a))
num3<-c(125,125)
vet.pow3<-rep(NA,length(a))
num4<-c(40,40)
vet.pow4<-rep(NA,length(a))
set.seed(122021)
for (i in 1:length(a))
  vet.pow0[i]<-retrop(c(a[i],b[i]),n=num0)$pow
for (i in 1:length(a))
  vet.pow1[i]<-retrop(c(a[i],b[i]),n=num1)$pow
for (i in 1:length(a))
  vet.pow2[i]<-retrop(c(a[i],b[i]),n=num2)$pow
for (i in 1:length(a))
  vet.pow3[i]<-retrop(c(a[i],b[i]),n=num3)$pow
for (i in 1:length(a))
  vet.pow4[i]<-retrop(c(a[i],b[i]),n=num4)$pow
plot(a,vet.pow0,xlab="Greater proportion",ylab="Power", type="b",
     col = "green", xlim = c(0,1), ylim = c(0,1.1))
points(a, vet.pow1, col="greenyellow", pch="|")
lines(a, vet.pow1, col="greenyellow",lty=1)
points(a, vet.pow2, col="yellow3", pch="+")
lines(a, vet.pow2, col="yellow3",lty=2)
points(a, vet.pow3, col="orange",pch="°")
lines(a, vet.pow3, col="orange", lty=3)
points(a, vet.pow4, col="red",pch="*")
lines(a, vet.pow4, col="red", lty=4)
legend(.364,1.1,
      legend=c("n=800","n=450","n=250","n=125", "n=40"),
      col=c("green","greenyellow","yellow3","orange","red"),
```

```
title = expression('Sample size (n=n'[1]*'n'[2]*')'),  
pch=c("o","|","+","°","*"), lty=c(1,2,3,4), ncol=1)
```

Appendix C: Code for converting proportions to Cohen's h and vice versa.

```
# Conversion From proportions to Cohen's h
hprop <- function(props = c(prop1,prop2)){
  prop1 <- props[1] ; prop2 <- props[2]
  if (prop1 <= prop2)
    stop("First proportion must be greater than second proportion")
  phi1 <- 2*asin(sqrt(props[1]))
  phi2 <- 2*asin(sqrt(props[2]))
  h <- (phi1 - phi2)
  ris <- list(phi1 = phi1, phi2 = phi2, h = h)
  return(ris)
}

# Conversion from Cohen's h to proportions
diffprop <- function(phi = c(phi1,phi2)){
  phi1 <- phi[1] ; phi2 <- phi[2]
  if (phi1 <= phi2)
    stop("First phi must be greater than second phi")
  h <- (phi1 - phi2)
  prop1 <- (-cos(phi1)+1)/2
  prop2 <- (-cos(phi2)+1)/2
  diffp <- prop1 - prop2
  ris <- list(h = h, prop1 = prop1 , prop2 = prop2, diffp = diffp)
  return(ris)
}
```

Appendix D: Code of the functions for retrospective and prospective design analysis for the difference between independent proportions.

```
# Internal function to always round 0.5 up
# (see: https://stackoverflow.com/questions/12688717/round-up-from-5)
round2 <- function(x, digits = 0) {
  posneg <- sign(x)
  z <- abs(x) * 10^digits
  z <- z + 0.5
  z <- trunc(z)
  z <- z / 10^digits
  z * posneg
}

#####

# Internal function to compare proportions based on normal approximation
zprops<-function(x,n,alpha=.05,B=1e6){
  prop1<-x[1]/n[1] ; prop2<-x[2]/n[2]
  if (prop1<=prop2){
    stop("First proportion must be greater than second proportion")
  }
  diff.prop<-prop1-prop2
  props.0<- (x[1]+x[2]) / (n[1]+n[2])
  p1_0=rbinom(n=B,n[1],p=props.0)/n[1]
  p2_0=rbinom(n=B,n[2],p=props.0)/n[2]
  dsim0<-p1_0-p2_0
  vc.low<-as.numeric(quantile(dsim0,alpha/2))
  vc.up<-as.numeric(quantile(dsim0,1-alpha/2))
  p.value= ( sum(dsim0<=(-diff.prop)) + sum (dsim0>=diff.prop) ) /B
  ris=list(diff.prop=diff.prop,p.value=round2(p.value,4),
           alpha=alpha,vc.low=vc.low,vc.up=vc.up)
  return(ris)
}

#####
```

```

# Retrospective function
retrop<-function(props=c(prop1,prop2),n=c(n1,n2),alpha=.05,B=1e6){
  vet.pow=rep(NA,B)
  x<-rep(NA,2)
  prop1<-props[1] ; prop2<-props[2]
  x[1]<-round2(n[1]*prop1) ; x[2]<-round2(n[2]*prop2)
  if (prop1<=prop2){
    stop("First proportion must be greater than second proportion")
  }
  diff.prop.plau<-prop1-prop2
  p1_1=rbinom(n=B,n[1],p=prop1)/n[1]
  p2_1=rbinom(n=B,n[2],p=prop2)/n[2]
  diff_1=p1_1-p2_1
  lb<-zprops(x=x,n=n,alpha=alpha,B=B)$vc.low
  ub<-zprops(x=x,n=n,alpha=alpha,B=B)$vc.up
  for (i in 1:B) vet.pow[i]= (diff_1[i]<=lb | diff_1[i]>=ub)
  pow <-sum(vet.pow)/B
  typeS <- sum((diff_1<=lb)) / sum(vet.pow)
  typeM <- mean( abs( diff_1[diff_1<=lb | diff_1>=ub]))/diff.prop.plau
  ris=list(diff.prop.plau=round2(diff.prop.plau,2),power=round2(pow,3),
          typeM=round2(typeM,3),typeS=round(typeS,3))
  return(ris)
}

#####

# Prospective function
prosp<-function(props=c(p1,p2),alpha=.05,power=.80,B=1e6,
                rangen=c(15,1000),tol = .01,iter=20,verbose=FALSE)
{
  if (props[1]<=props[2]){
    stop("First proportion must be greater than second proportion")
  }
  j=0
  diffp=props[1]-props[2]
  n1<-n2<-rangen[2]
  vals<-retrop(props=c(props[1],props[2]),n=c(n1,n2),alpha=alpha,B=B)
  if (vals$power<power)

```

```

{ cat(paste0("Actual power = ", vals$power, " with n = ",
            rangen[2], " (per group), " ), "\n")
  cat(paste0("    try to set minimum of rangen to ", rangen[2],
            " and increase maximum of rangen.", "\n"))
  return(vals)
}
n_seq <- seq( rangen[1], rangen[2], by = 1 )
(n_target <- round2(median(n_seq)))
find_power=FALSE
while( (!find_power) ) {
  if (verbose!=FALSE) {cat("Estimating power with n per group
=",n_target, "\n")}
  n1<-n2<-n_target
  est_P <- vals<-
retrop(props=c(props[1],props[2]),n=c(n_target,n_target),
        alpha=alpha,B=B)
  est_power <- est_P$power
  if (verbose!=FALSE) {cat("Estimated power est_power
=",est_power, "\n")}
  if ( (est_power<=(power+tol)) & (est_power>(power-tol)) ) {
    find_power <- TRUE
  } else {
    if (length(n_seq)==1) {
      print(n_seq)
      stop(" ")
    }

    if ( est_power > (power-tol) ) {
      (n_seq <- seq( min(n_seq), n_target, by = 1))
      (n_target <- round2(median(n_seq)))
    } else {
      (n_seq <- seq( n_target, max(n_seq), by = 1))
      (n_target <- round2(median(n_seq)))
    }
  }
  j=j+1
  if (j==iter)
  { cat(paste0("The algorithm did not achieve convergence with iter =
",
              j, "\n"))
    cat(paste0("Try to increase tol and/or iter" , "\n"))
    return(NA) }
}

```

```
}  
  ris <- list(power=round2(power,2),est_power=round2(est_power,2),  
n_group= n_target,  
              typeM=round2(est_P$typeM,3), typeS=round2(est_P$typeS,3))  
  return(ris)  
}
```

Appendix E: *R* functions for design analysis with independent proportions.

The *R* functions presented in the paper to perform design analyses on the difference between independent proportions are:

- `retrop()`
- `prosp()`

These functions are described in detail in the following section. Readers can use these functions to conduct design analyses on their own data. These functions can also be used to further extend the *R* code to different or more complex cases that were not part of the aims of this work. These functions have a didactic purpose only, and can be optimised under an IT point of view.

R functions

retrop()

The function

```
retrop ( props = c ( prop1, prop2 ), n = c ( n1, n2 ), alpha = .05,  
        B = 1e6)
```

The function `retrop()` runs retrospective design analyses on difference between independent proportion with a fixed type I error (`alpha`). The user must specify:

- Both plausible proportions (`prop1` and `prop2`).
- Sample size relatively to each proportion (`n1` and `n2`).

Function arguments

- `props` = the plausible proportions.
- `n` = sample size.
- `alpha` = type I error, the default setting is .05.
- `B` = number of replications of the simulations.

Function return values

- `diff.prop.plau` = the plausible difference between proportions we have set.
- `power` = the calculated power.
- `typeM` = type *M* error.
- `typeS` = type *S* error.

Examples

```
> retrop(c(.25,.15), c(67,55), alpha = .01)  
$diff.prop.plau
```

```
[1] 0.1
```

```
$power
```

```
[1] 0.118
```

```
$typeM
```

```
[1] 2.194
```

```
$typeS
```

```
[1] 0
```

propssp()

The function

```
propssp ( props = c ( p1, p2 ), alpha = .05, power = .80, B = 1e6,  
          rangen = c ( 15 , 1000 ), tol = .01, iter = 20, verbose = FALSE )
```

The function `propssp()` runs prospective design analyses on difference between independent proportion with a fixed type I error (`alpha`). The user must specify:

- Both plausible proportions (`p1` and `p2`)

Function arguments

- `props` = the plausible proportions.
- `alpha` = type I error, the default setting is `.05`.
- `power` = the desired power.
- `B` = number of replications of the simulations.
- `rangen` = a vector of two values indicating the range of `n` within which the function calculates power, the default setting is from 15 to 1000.
- `tol` = numerical tolerance used to calculate power; the default setting is `.01`.
- `iter` = number of times the function repeats calculation to get to the desired power before asking to change `iter` or tolerance (power can be changed too); the default setting is 20.
- `verbose` = if `TRUE` the function gives feedback on the calculation it is doing on sample size and estimated power; the default setting is `FALSE`.

Function return values

- `power` = the set desired power.
- `est_power` = the calculated power.
- `n_group` = the sample size for each group.

- `typeM` = type M error.
- `typeS` = type S error.

Examples

```
> propssp(c(.125,.175), power = .95, rangen = c(1000,2000))
$power
[1] 0.95

$est_power
[1] 0.94

$n_group
[1] 1250

$typeM
[1] 1.035

$typeS
[1] 0
```

```
> propssp(c(.20,.10), alpha = .1, tol = .05, verbose = TRUE)
Estimating power with n per group = 508
Estimated power est_power = 0.998
Estimating power with n per group = 262
Estimated power est_power = 0.945
Estimating power with n per group = 139
Estimated power est_power = 0.769
$power
[1] 0.8

$est_power
[1] 0.77

$n_group
[1] 139

$typeM
```

```
[1] 1.169
```

```
$types
```

```
[1] 0
```

Appendix F: code for retrospective design analyses and difference between H_0 and H_1 histograms on the case study.

```
alpha = .05
B = 1e6

#15% plausible effect size

set.seed(122021)
retrowork15 <- retrop(c(.40+.15/2, .40-.15/2), c(30,24))
nw <- c(30,24)
prop1w15 <- .40+.15/2
prop2w15 <- .40-.15/2
xw15 <- rep(NA,2)
xw15[1]<-round2(nw[1]*prop1w15) ; xw15[2]<-round2(nw[2]*prop2w15)
props.0w15 <- (xw15[1]+xw15[2])/(nw[1]+nw[2])
p1w15_0 = rbinom(n=B,nw[1],p=props.0w15)/nw[1]
p2w15_0 = rbinom(n=B,nw[2],p=props.0w15)/nw[2]
dsim0w15 <- p1w15_0-p2w15_0
vc.loww15 <- as.numeric(quantile(dsim0w15,alpha/2))
vc.upw15 <- as.numeric(quantile(dsim0w15,1-alpha/2))
p1w15_1=rbinom(n=B,nw[1],p=prop1w15)/nw[1]
p2w15_1=rbinom(n=B,nw[2],p=prop2w15)/nw[2]
diff_1w15=p1w15_1-p2w15_1
hist(dsim0w15, breaks = 30, main = "",
      xlab = expression('Difference between proportions under H'[0]),
      col = "grey87", xlim = c(-0.7,0.7))
abline(v=c(vc.loww15, vc.upw15), col= "red")
hist(diff_1w15, main = "", breaks = 30,
      xlab = expression('Difference between proportions under H'[1]),
      col = "grey87", xlim = c(-0.7,0.7))
abline(v=c(vc.loww15, vc.upw15), col= "red")

#10% plausible effect size

set.seed(122021)
retrowork10 <- retrop(c(.40+.10/2, .40-.10/2), c(30,24))
prop1w10 <- .40+.10/2
prop2w10 <- .40-.10/2
xw10 <- rep(NA,2)
```

```

xw10[1]<-round2(nw[1]*prop1w10) ; xw10[2]<-round2(nw[2]*prop2w10)
props.0w10 <- (xw10[1]+xw10[2])/(nw[1]+nw[2])
p1w10_0 = rbinom(n=B,nw[1],p=props.0w10)/nw[1]
p2w10_0 = rbinom(n=B,nw[2],p=props.0w10)/nw[2]
dsim0w10 <- p1w10_0-p2w10_0
vc.loww10 <- as.numeric(quantile(dsim0w10,alpha/2))
vc.upw10 <- as.numeric(quantile(dsim0w10,1-alpha/2))
p1w10_1=rbinom(n=B,nw[1],p=prop1w10)/nw[1]
p2w10_1=rbinom(n=B,nw[2],p=prop2w10)/nw[2]
diff_1w10=p1w10_1-p2w10_1
hist(dsim0w10, breaks = 30, main = "",
      xlab = expression('Difference between proportions under H'[0]),
      col = "grey87", xlim = c(-0.7,0.7))
abline(v=c(vc.loww10, vc.upw10), col= "red")
hist(diff_1w10, main = "", breaks = 30,
      xlab = expression('Difference between proportions under H'[1]),
      col = "grey87", xlim = c(-0.7,0.7))
abline(v=c(vc.loww10, vc.upw10), col= "red")

#5% plausible effect size

set.seed(122021)
retrowork05 <- retrop(c(.40+.05/2,.40-.05/2), c(30,24))
prop1w05 <- .40+.05/2
prop2w05 <- .40-.05/2
xw05 <- rep(NA,2)
xw05[1]<-round2(nw[1]*prop1w05) ; xw05[2]<-round2(nw[2]*prop2w05)
props.0w05 <- (xw05[1]+xw05[2])/(nw[1]+nw[2])
p1w05_0 = rbinom(n=B,nw[1],p=props.0w05)/nw[1]
p2w05_0 = rbinom(n=B,nw[2],p=props.0w05)/nw[2]
dsim0w05 <- p1w05_0-p2w05_0
vc.loww05 <- as.numeric(quantile(dsim0w05,alpha/2))
vc.upw05 <- as.numeric(quantile(dsim0w05,1-alpha/2))
p1w05_1=rbinom(n=B,nw[1],p=prop1w05)/nw[1]
p2w05_1=rbinom(n=B,nw[2],p=prop2w05)/nw[2]
diff_1w05=p1w05_1-p2w05_1
hist(dsim0w05, breaks = 30, main = "",
      xlab = expression('Difference between proportions under H'[0]),
      col = "grey87", xlim = c(-0.7,0.7))
abline(v=c(vc.loww05, vc.upw05), col= "red")
hist(diff_1w05, main = "", breaks = 30,
      xlab = expression('Difference between proportions under H'[1]),

```

```
col = "grey87", xlim = c(-0.7,0.7))  
abline(v=c(vc.loww05, vc.upw05), col= "red")
```


Appendix G: Code for prospective design analyses on the case study.

```
set.seed(122021)

#Early work release (95% power)

prowork1595 <- propssp(c(.40+.15/2,.40-.15/2), power = .95)
prowork1095 <- propssp(c(.40+.10/2,.40-.10/2), power = .95)
prowork0595 <- propssp(c(.40+.05/2,.40-.05/2), power = .95,
                      rangen = c(2000,3000))

#Early work release (80% power)

prowork1580 <- propssp(c(.40+.15/2,.40-.15/2))
prowork1080 <- propssp(c(.40+.10/2,.40-.10/2))
prowork0580 <- propssp(c(.40+.05/2,.40-.05/2),
                      rangen = c(1000,2000))

#Early work release (70% power)

prowork1570 <- propssp(c(.40+.15/2,.40-.15/2), power = .7)
prowork1070 <- propssp(c(.40+.10/2,.40-.10/2), power = .7)
prowork0570 <- propssp(c(.40+.05/2,.40-.05/2), power = .7,
                      rangen = c(1000,2000))

#Early work release (60% power)

prowork1560 <- propssp(c(.40+.15/2,.40-.15/2), power = .6)
prowork1060 <- propssp(c(.40+.10/2,.40-.10/2), power = .6)
prowork0560 <- propssp(c(.40+.05/2,.40-.05/2), power = .6)
```