Designing Studies and Evaluating Research Results: Type M and Type S Errors for Pearson Correlation Coefficient

Giulia Bertoldo [1], Claudio Zandonella Callegher [1], and Gianmarco Altoè [1,*]

[1] Department of Developmental Psychology and Socialisation, University of Padova, Padova, Italy

Submitted to *Meta-Psychology*. Click here to follow the fully transparent editorial process of this submission. Participate in open peer review by commenting through hypothes.is directly on this preprint.

Author Note

ORCiD

Giulia Bertoldo: 0000-0002-6960-3980

Claudio Zandonella Callegher:0000-0001-7721-6318

Gianmarco Altoè: 0000-0003-1154-9528

We have no known conflict of interest to disclose.

*Correspondence:
Gianmarco Altoè, Department of Developmental Psychology and Socialization,
University of Padova, Via Venezia 8, 35131 Padova, Italy
gianmarco.altoe@unipd.it

Abstract

Keywords— Correlation coefficient; Type M Error, Type S error, Design analysis, Effect size

---

1 Introduction

Psychological science is increasingly committed to scrutinize its published findings by promoting large-scale replication efforts (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014, 2018; Open Science Collaboration, 2015). In replication studies, the protocol of a previously published study is repeated as closely as possible with a new sample. Interestingly, many replication studies found smaller effects than originals, suggesting a sort of decline effect from original studies to replications (Camerer et al., 2018; Open Science Collaboration, 2015; Schooler, 2014). Among the many factors that could explain such decline, one relates to a feature of study design: statistical power (Protzko & Schooler, 2017). In particular, it is plausible for original studies to have lower statistical power than their replications. When underpowered studies are analyzed using thresholds, such as statistical significance levels, effects passing such thresholds have to exaggerate the true effect size (Button et al., 2013; Gelman, Skardhamar, & Aaltonen, 2017; Ioannidis, 2008; Ioannidis, Pereira, & Horwitz, 2013; Lane & Dunlap, 1978). Indeed, as will be extensively shown below, in underpowered studies only large effects correspond to values that can reject the null hypothesis and be statistically significant. As a consequence, if the original study was underpowerd and found an exaggerated estimate of the effect, the replication results will likely show a decline. The concept of statistical power finds its natural developement in the Neyman-Pearson framework of statistical inference and this is the framework that we adopt in this contribution. Further discussion on the Neyman and Perason approach and a comparison with Null Hypothesis Significance Testing (NHST) is available in Altoè et al. (2020) and Gigerenzer, Krauss, and Vitouch (2004).

Effect size exaggeration is a corollary consequence of low statistical power that might not be evident at first. Indeed, by definition, statistical power focuses on statistical significance, not effect size estimation (i.e., statistical power is the probability to find a statistically significant result if an effect of a certain size actually exists). However, effect size estimation and statistical significance are closely related. This point can be highlighted estimating the Type M (magnitude) and Type S (sign) errors characterizing a study design (Gelman & Carlin, 2014). Given a study design (i.e., sample size, statistical test directionality, and α level), Type M error indicates the factor by which a statistically significant effect would be, on average, exaggerated. Type S error indicates the probability to find a statistically significant effect in the opposite

direction to the one considered plausible. It is evident that both errors are defined starting from a reasoned guess on the plausible magnitude and direction of the effect under study, which is called *plausible effect size* (Gelman & Carlin, 2014). The analysis that researchers perform to consider Type M and Type S errors in their research practice is called *design analysis*, given the special focus posed into considering the design of a study (Altoè et al., 2020; Gelman & Carlin, 2014).

Why do these errors matter? The exaggeration of effect sizes, in the right or in the wrong direction, has important consequences on a theoretical and applied level. On a theoretical level, studies' designs with high Type M and Type S errors can foster distorted expectations on the effect under study, triggering a vicious cycle for the planning of future studies. This point is relevant also for the design of replication studies, which could turn out to be underpowered if they do not take into account possible inflations of the original effect (Button et al., 2013). When studies are used to inform policymaking and real-world interventions, implications can go beyond the academic research community and can impact society at large. In these settings, we could assist to a "hype and disappointment cycle" (Gelman, 2019), where true effects turn out to be much less impressive than expected. This can produce undesirable consequences on people's lives, a consideration that invites researchers to assume responsibility in effectively communicating the risks related to effects quantification.

Type M (magnitude) and Type S (sign) errors are not widely known in the psychological research community but their consideration during the research process has the potential to improve current research practices, for example, by increasing the awarness that design choices have on possible studies' results. In a previous work we illustrated Type M and Type S errors using Cohens' d as a measure of effect size (Altoè et al., 2020). In this contribution we aim to to further increase familiarity with Type M and Type S errors, considering another common effect size measures in psychology: Pearson correlation coefficient, ρ. In the following paragraphs we discuss what Type M and Type S errors are and how to perform a design analysis using the ad-hoc R functions that we developed.

2 Type M and Type S errors

Pearson correlation coefficient is a standardized effect size measure indicating the strength and the direction of the relationship between two continuous variables (Cohen, 1988; Ellis, 2010). Even though the correlation coefficient is widely known, we briefly go over its main features using an example. Imagine that we were interested to measure the relationship between anxiety and depression in a population and we plan a study with n participants, where, for each participant, we measure the level of anxiety (i.e., variable X) and the level of depression (i.e., variable Y). At the end of the study we will have n pairs of values X and Y. The correlation coefficient helps us answer the questions: how strong is the relationship between anxiety

and depression in this population? Is the relationship positive or negative? Correlation ranges from -1 to +1, indicating respectively two extreme scenarios of perfect negative relationship and perfect positive relationship. Since the correlation coefficient is a dimensionless number, it is a signal to noise ratio where the signal is given by the covariance between the two variables ($cov(x, y)$) and the noise is expressed by the product between the standard deviations of the two variables ($S_x S_y$; see Formula1). Following the conventional standards, in this contribution we will use the symbol ρ to indicate the correlation in the population and the symbol r to indicate the value measured in a sample.

$$r = \frac{cov(x, y)}{S_x S_y}. \tag{1}$$

Magnitude and sign are two important features characterizing Pearson correlation coefficient and effect size measures in general. And, when estimating effect sizes, errors could be committed exactly regarding these two aspects. Gelman and Carlin (2014) introduced two indexes to quantify these risks:

- Type M error, where M stands for magnitude, is also called Exaggeration Ratio - the factor by which a statistically significant effect is on average exaggerated.
- Type S error, sign - the probability to find a statistically significant result in the opposite direction to the plausible one.

How are these errors computed? In the next paragraphs we approach this question preferring an intuitive perspective. For a formal definition of these errors, we refer the reader to Altoè et al. (2020); Gelman and Carlin (2014); Lu, Qiu, and Deng (2018). Take as an example the previous fictitious study on the relationship between anxiety and depression and imagine we decide to sample 50 individuals (sample size, n = 50) and to set the α level to 5% and to perform a two-tailed test. On the basis of theoretical considerations, we expect the plausibly true correlation in the population to be quite strong and positive which we formalize as $\rho = .50$. In order to evaluate the Type M and Type S errors in this research design, imagine repeating the same study many times with new samples drawn from the same population and, for each study, register the observed correlation (r) and the corresponding p-value.

The first step to compute Type M error is to select only the observed correlation coefficients that are statistically significant in absolute value (for the moment, we do not care about the sign) and to calculate their mean. Type M error is given by the ratio between this mean (i.e., mean of statistically significant correlation coefficients in absolute value) and the plausible effect hypothesized at the beginning, which in this example is $\rho = .50$. Thus, given a study design, Type M error tells us what is the average overestimation of an effect that is statistically significant.

Type S error is computed as the proportion of statistically significant results that have the opposite sign compared to the plausible effect size. In the present example we hypothesized a positive relationship, specifically $\rho = .50$. Then, Type S error is the ratio between the number of times we observed a negative

statistically significant result and the total number of statistically significant results. In other words, Type S error indicates the probability to obtain a statistically significant result in the opposite direction to the one hypothesized.

The central and possibly most difficult point in this process is reasoning on what could be the plausible magnitude and direction of the effect of interest. This critical process, which is central also in traditional a priori power analysis, is an opportunity for researchers to aggregate, formalize and incorporate prior information on the phenomenon under investigation (Gelman & Carlin, 2014). What is plausible can be determined on theoretical grounds, using expert knowledge elicitations techniques and consulting literature reviews and meta-analysis, always taking into account the presence of effect sizes inflation in the published literature (Todo:) (Anderson, 2019). The plausible effect size approximates the true effect, which is never known but can be thought as "that which would be observed in a hypothetical infinitely large sample" (Gelman & Carlin, 2014, p. 642). For a more exhaustive description on plausible effect size, we refer the interested reader to Altoè et al. (2020); Gelman and Carlin (2014).

3 Design Analysis

Researchers can consider Type M and Type S errors in their practice by performing a *design analysis* (Altoè et al., 2020; Gelman & Carlin, 2014). Ideally, a design analysis should be performed when designing a study. In this phase it is specifically called *prospective design analysis* and it can be used as a sample size planning strategy where statistical power is considered together with Type M and Type S errors. However, design analysis can also be beneficial to evaluate the inferential risks in studies that have already been conducted and where the study design is known. In these cases, Type M and Type S errors can support results interpretation by communicating the inferential risks in that research design. When design analysis happens at this later stage, it takes the name of *retrospective design analysis*.

In the following sections we illustrate how to perform prospective and retrospective design analysis using some examples. We developed two R functions to perform design analysis for Pearson correlation, which are available at the page (Todo:) ... . The function to perform a prospective design analysis is `pro_r()`. It requires as input the plausible effect size (`rho`), the statistical power (`power`) the directionality of the test (`alternative`) which can be set as: "`two.sided`", "`less`" or "`greater`" . Type I error rate (`sig_level`) is set as default at 5% and can be changed by the user. The `pro_r()` function returns the necessary sample size to achieve the desired statistical power, Type M error rate, the Type S error probability, and the critical value(s) above which a statistically significant result can be found. The function to perform retrospective design analysis is `retro_r()`. It requires as input the plausible effect size, the sample size used in the study, and the directionality of the test that was performed. Also in this case, Type I error

rate is set as default at 5% and can be changed by the user. The function `retro_r()` returns the Type M error rate, the Type S error probability, and the critical value(s). (Todo:) For further details on (how to use)computational aspects regarding the R functions, we refer to Appendix A.

4 Case Study

To familiarize the reader with Type M and Type S errors, we start our discussion with a retrospective design analysis of a published study. However, the ideal temporal sequence in the research process would be to perform a prospective design analysis in the planning stage of a research project. This is the time when the design is being laid out and useful improvements can be made to obtain more robust results. In this contribution, the order of presentation aims first, to provide an understanding of how to interpret Type M and Type S errors, and then discuss how they could be minimized. The following case study was chosen for illustrative purposes only and, by no means our objective is to judge the study beyond illustrating an application of how to calculate Type M and Type S errors on a published study.

We consider the study published in *Science* by Eisenberger, Lieberman, and Williams (2003) entitled: "Does Rejection Hurt? An fMRI Study of Social Exclusion". The research question originated from the observation that the Anterior Cingulate Cortex (ACC) is a region of the brain known to be involved in the experience of physical pain. Could pain from social stimuli, such as social exclusion, share similar neural underpinnings? To test this hypothesis, 13 participants were recruited and each one had to play a virtual game with other two players while undergoing functional Magnetic Resonance Imaging (fMRI). The other two players were fictitious, and participants were actually playing against a computer program. Players had to toss a virtual ball among each other in three conditions: social inclusion, explicit social exclusion and implicit social exclusion. In the social inclusion condition the participant regularly received the ball. In the explicit social exclusion condition the participant was told that, due to technical problems, he was not going to play that round. In the implicit social exclusion condition, the participant experienced being intentionally left out from the game by the other two players. At the end of the experiment, each participant completed a self-report measure regarding their perceived distress when they were intentionally left out by the other players. Considering only the implicit social exclusion condition, a correlation coefficient was estimated between the measure of distress and neural activity in the Anterior Cingulate Cortex. As suggested by the large and statistically significant correlation coefficient between perceived distress and activity in the ACC, (Todo:) $r = .88$, $p < .005$ (Eisenberger et al., 2003, p. 291), authors concluded that social and physical pain seem to share similar neural underpinnings.

Before proceeding to the retrospective design analysis, we refer the interested reader to some background history regarding this study. This was one of the many studies included in the famous paper

"Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" (Vul, Harris, Winkielman, & Pashler, 2009) which raised important issues regarding the analysis of neuroscientific data. In particular, this paper noted that the magnitude of correlation coefficients between fMRI measures and behavioral measures were beyond what could be considered plausible. We refer the interested reader also to the commentary by Yarkoni (2009), who noted that the implausibly high correlations in fMRI studies could be largely explained by the low statistical power of experiments.

A retrospective design analysis should start with thorough reasoning on the plausible size and direction of the effect under study. In order to produce valid inferences, a lot of attention should be devoted to this point by integrating external information. For the sake of this example, we turn to the considerations made by Vul and Pashler (2017) who (Todo:) considered correlations between personality measures and neural activity to be likely around $\rho = .25$. A correlation of $\rho = .50$ was deemed plausible but optimistic and a correlation of $\rho = .75$ was considered theoretically plausible but unrealistic.

5 Retrospective Design Analysis

To perform a retrospective design analysis on the case study, we need information on the research design and the plausible effect size. Based on the previous considerations, we set the plausible effect size to be $\rho = .25$. Information on the sample size was not available in the original study (Eisenberger et al., 2003) and was retrieved from Vul et al. (2009) to be $n = 13$. The α level and the directionality of the test were not reported in the original study, so for the purpose of this example we will consider $\alpha = .05$ and a two-tailed test. Given this study design, what are the inferential risks in terms of effect size estimation?

We can use the R function `retro_r()`, which inputs and outputs are displayed in Figure 1. In this study, the statistical power is .13, that is to say, there is a 13% probability to reject the null hypothesis, if an effect of at least $\rho = |.25|$ exists. Consider this point together with the results obtained in the experiment: r = .88, p < .005 (Eisenberger et al., 2003, p. 291). It is clear that, even though the probability to reject the null hypothesis is low (power of 13%), this event could happen. And when it does happen, it is tempting to believe that results are (Todo:) impressive (Gelman & Loken, 2014). However, this design comes with serious inferential risks for the estimation of effect sizes, which could be grasped by presenting Type M and Type S errors. A glance at their value communicates that it is not impossible to find a statistically significant result, but when it does happen, the effect sizes could be largely overestimated - Type M = 2.58 - and maybe even in the wrong direction - Type S = .03. The Type M error rate of (Todo:) around 2.60 indicates that a statistically significant correlation is on average about two and a half times the plausible value. In other words, statistically significant results emerging in such a research design will on average overestimate the plausible correlation coefficient by 160%. The Type S error of .03 suggests that there is a

three percent probability to find a statistically significant result in the opposite direction, in this example, a negative relationship.

```
> retro_r(rho = .25, n = 13, alternative = "two.sided", sig_level = .05, seed = 2020)

        Design Analysis

Hypothesized effect: rho = 0.25

Study characteristics:
   n    alternative   sig_level
   13   two.sided     0.05

Inferential risks:
   power   typeM   typeS
   0.127   2.583   0.028

Critical value: r =  ±0.553
```

Figure 1: Input and Output of the function `retro_r()` for retrospective design analysis. Case study: Eisenberger et al. (2003). The plausible correlation coefficient is $\rho = .25$, the sample size is 13, and the statistical test is two-tailed.

In this research design, the critical values above which a statistically significant result is declared correspond to $\rho = \pm.55$ (Figure 1). These values are highlighted in Figure 2 as the vertical lines in the sampling distribution of correlation coefficients under the null hypothesis. Notice that the plausible effect size lies in the region of acceptance of the null hypothesis. Therefore, it is impossible to simultaneously find a statistically significant result and estimate an effect close to the plausible one ($\rho = .25$). The figure represents the so called Winner's curse: "the 'lucky' scientist who makes a discovery is cursed by finding an inflated estimate of that effect" (Button et al., 2013).

6 Prospective Design Analysis

Ideally, Type M and Type S errors should be considered in the design phase of a study during the decision-making process regarding the experimental protocol. At this stage, prospective design analysis can be used as a sample size planning strategy which aims to minimize Type M and Type S errors in the upcoming study.

Imagine that we were part of the research team in the previous case study exploring the relationship between activity in the Anterior Cerebral Cortex and perceived distress. When drafting the research protocol, we face the inevitable discussion on how many participants we are going to recruit. This choice depends on available resources, type of study design, constraints of various nature and, importantly, the plausible magnitude and direction of the phenomenon that we are going to study. As previously mention, deciding on a plausible effect size is a fundamental step which requires great effort and should not be (Todo:) a game where different numbers are tried until the desired sample size is reached. Instead,

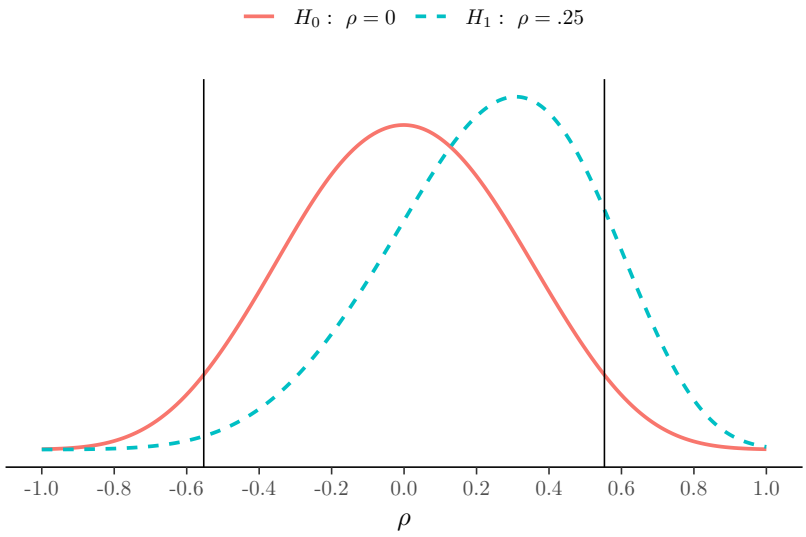

Figure 2: Winner's course. H_0 = Null Hypothesis, H_1 = Alternative Hypothesis. When sample size, directionality of the test and Type error probability are set, also the smallest effect size above which is possible to find a statistically significant result is set. In this case, the plausible effect size, $\rho = .25$, lies in the region where it is not possible to reject H_0. Thus, it is impossible to simultaneously find a statistically significant result and an effect close to the plausible one. In other words, a statistically significant effect must exaggerate the plausible effect size.

proposing a plausible effect size is where the expert knowledge of the researcher can be formalized and can greatly contribute to the informativeness of the study that is being planned. For the sake of this examples, we adopt the previous consideration and we suppose that common agreement is reached on a plausible correlation coefficient to be around $\rho = .25$. Finally, we would like to leave open the possibility to explore whether the relationship goes in the opposite direction to the one hypothesized, so we decide to perform a two-tailed test.

We can implement the prospective design analysis using the function `pro_r()` which inputs and outputs are displayed in Figure 3. About 125 participants are necessary to have 80 % probability to detect an effect of at least $\rho = \pm.25$, if it actually exists. With this sample size, the Type S error is minimized and approximates zero. In this study design, the Type M error is 1.12 indicating that statistically significant results are on average exaggerated by 12%. It is possible to notice that the critical values are $\rho = \pm.18$, further highlighting that our plausible effect size is actually included among those values that lead to the acceptance of the alternative hypothesis.

In a design analysis, it is advisable to investigate how the inferential risks would change according to different scenarios in terms of statistical power and plausible effect size. Changes in both these factors impact Type M and Type S errors. For example, maintaining the plausible correlation of $\rho = .25$, if we

```
> pro_r(rho = .25, power = .8, alternative = "two.sided", sig_level = .05, seed = 2020)

        Design Analysis

Hypothesized effect: rho = 0.25

Study characteristics:
   n     alternative   sig_level
   125   two.sided     0.05

Inferential risks:
   power   typeM   typeS
   0.806   1.111   0

Critical value: r =  ±0.176
```

Figure 3: Input and Output of the function `pro_r()` for prospective design analysis. Plausible correlation coefficient is $\rho = .25$, statistical power is 80% and the statistical test is two-tailed.

decrease statistical power from .80 to .60 only 76 participants are required (see Table 1). However, this is associated with an increased Type M error rate from 1.12 to 1.28. That is to say, with 76 subjects the plausible effect size will be on average overestimated by 28%. Alternatively, imagine that we would like to maintain a statistical power of 80%, what happens if the plausible effect size is slightly larger or smaller? The necessary sample size would spike to 344 for a $\rho = .15$ and decrease to 60 for $\rho = .35$. In both scenarios, the Type M error remains about 1.12, which reflects the more general point that for 80% power, Type M error is around 1.10. In all these scenario, Type S error is close to zero, hence not worrisome.

Table 1: Prospective design analysis in different scenarios of plausible effect size and statistical power.

ρ	Power	Sample Size	Type M	Type S	Critical r value
0.25	0.6	76	1.280	0	±0.226
0.15	0.8	344	1.116	0	±0.106
0.35	0.8	60	1.115	0	±0.254

Note: In all cases, alternative = `"two.sided"` and sig_level = .05.

For completeness, Figure 4 summarizes the relationship between statistical power, Type M and Type S errors as a function of sample size in three scenarios of plausible correlation coefficients. We display the three values that Vul and Pashler (2017) considered for correlations between fMRI measures and behavioral measures with different degrees of plausibility. An effect of $\rho = .75$ was deemed theoretically plausible but unrealistic, $\rho = .50$ was more plausible but optimistic, and $\rho = .25$ was more likely. The curves illustrate a general point: Type M and Type S error increase with smaller sample sizes, smaller plausible effect sizes and lower statistical power. Also, the figure shows that statistical power, Type M and Type S errors are related to each other: as power increases, Type M and Type S errors decrease.

At first, it might seem that Type M and Type S errors are redundant with the information provided

by statistical power. Even though they are related, we believe that Type M and Type S errors bring added value during the design phase of a research protocol because they facilitate a connection between how a study is planned and how results will actually be evaluated. That is to say, final results will comprise of a test statistics with an associated p-value and effect size measure. If the interest is maximizing the accuracy with which effects will be estimated, then Type M and Type S errors directly communicate the consequences of design choices on effect size estimation.

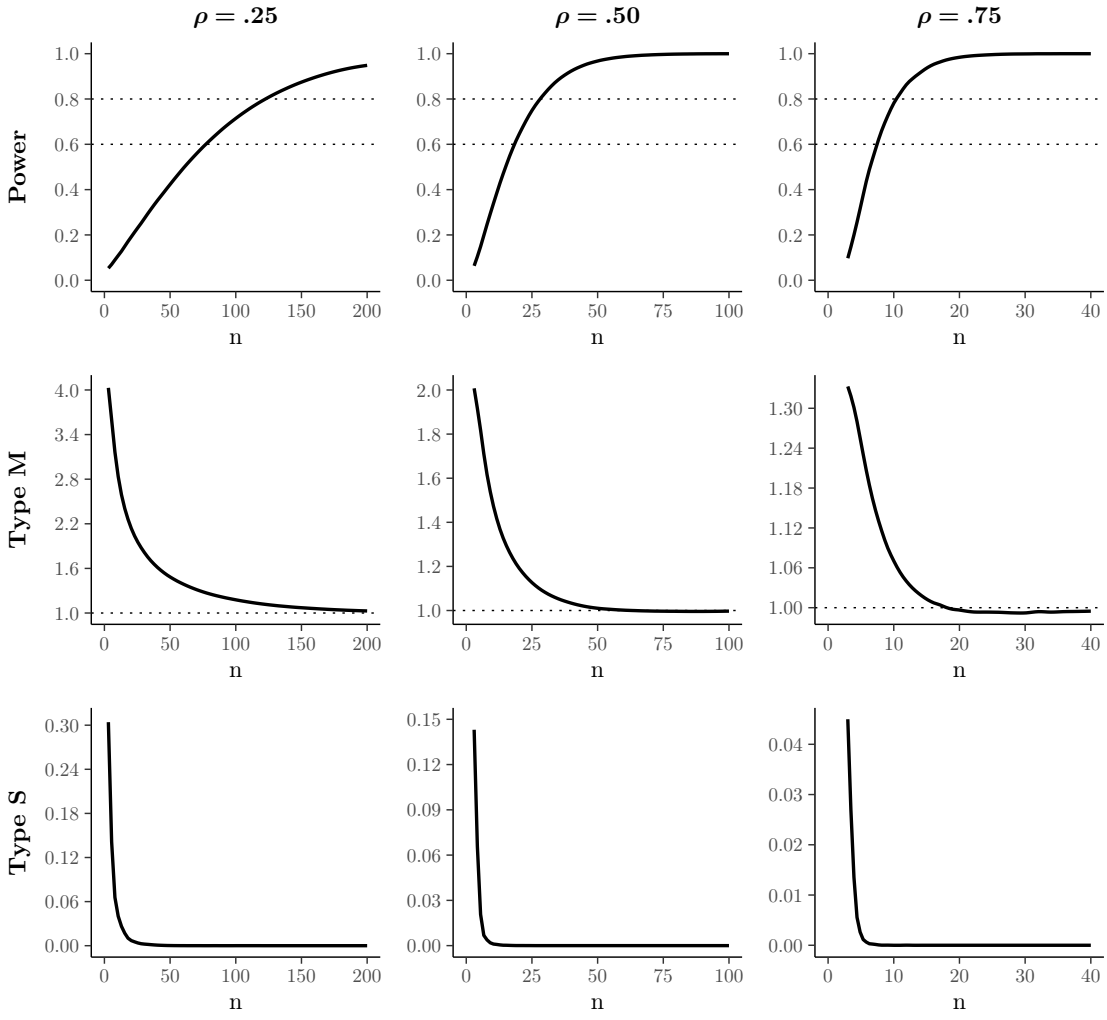

Figure 4: How Type M, Type S and Statistical power vary as a function of sample size in three different scenarios of plausible effect size (ρ = .25, ρ = .50, ρ = .75). Note that, for the sake of interpretability, we decided to use different scales for both the x-axis and y-axis in the three scenarios of plausible effect size.

7 Varying α levels and Hypotheses Directionality

So far, we did not discuss two other important decisions that researchers have to take when designing a study: statistical significance threshold or α level, and directionality of the statistical test, one-tailed or two-tailed. In this section we illustrate how different choices regarding these aspects impact Type M and Type S errors.

A lot has been written regarding the automatic adoption of a conventional α level of 5% (e.g., Gigerenzer et al., 2004; Lakens et al., 2018). This practice is increasingly discouraged, and researchers are invited to think about the best trade-off between α level and statistical power, considering the aim of the study and available resources. The α level impacts Type M and Type S errors as much as it impacts statistical power. Everything else equal, Type M error increases with decreasing α level (i.e., negative relationship), whereas Type S error decreases with decreasing α level (i.e., positive relationship). To further illustrate the relation between Type M error and α level, let us take as an example the previous case study with a sample of 13 participants, plausible effect size $\rho = .25$ and two-tailed test. Table 2 shows that by lowering the α level from 10% to .10%, the critical values move from (Todo:) $\rho = \pm.48$ to $\rho = \pm.80$. This suggests that, with these new higher thresholds, the exaggeration of effects will be even more pronounced because effects have to be even larger to pass such higher critical values. Instead, the relationship between Type S error and α level can be clarified thinking that by lowering the statistical significance threshold, we are being more conservative to falsely reject the null hypothesis in general which implies that we are also being more conservative to falsely rejecting the null hypothesis in the wrong direction.

Table 2: How changes in α level impact Power, Type M error, Type S error and critical values.

α-level	Power	Type M	Type S	Critical r value
0.100	0.2	2.369	0.040	±0.476
0.050	0.1	2.583	0.028	±0.553
0.010	0.0	2.977	0.011	±0.684
0.005	0.0	3.088	0.014	±0.726
0.001	0.0	3.340	0.000	±0.801

Note: In all cases, n = 13 and alternative = `"two.sided"`.

Another important choice in study design is the directionality of the test (i.e., one-tailed or two-tailed). Design analysis invites reasoning on the plausible effect size and hypothesizing the direction of the effect, not only its magnitude. So why should a researcher perform nondirectional statistical tests when there is a hypothesized direction? Performing a two-tailed test leaves open the possibility to find an unexpected result in the opposite direction (Cohen, 1988), a possibility which may be of special interest for preliminary exploratory studies. However, in more advanced stages of a research program (i.e., confirmatory study), directional hypotheses benefit from higher statistical power and lower Type M error rates (Figure 5). As an

example, let us consider the differences between a two-tailed test and a one-tailed test in the previous case study. We can perform a new prospective design analysis (Figure 6) with plausible correlation of $\rho = .25$, 80% statistical power, but this time setting the argument `alternative` in the R function to `"greater"`. A comparison of the two prospective design analyses, (Todo:) Figure 3 and Figure 6, suggests that the same Type M error rate of about 10% is guaranteed with 94 participants, instead of 125 subjects. Note that, Type S error is not possible in directional statistical tests. Indeed, all the statistically significant results are obtainable only in the hypothesized direction, not the opposite one.

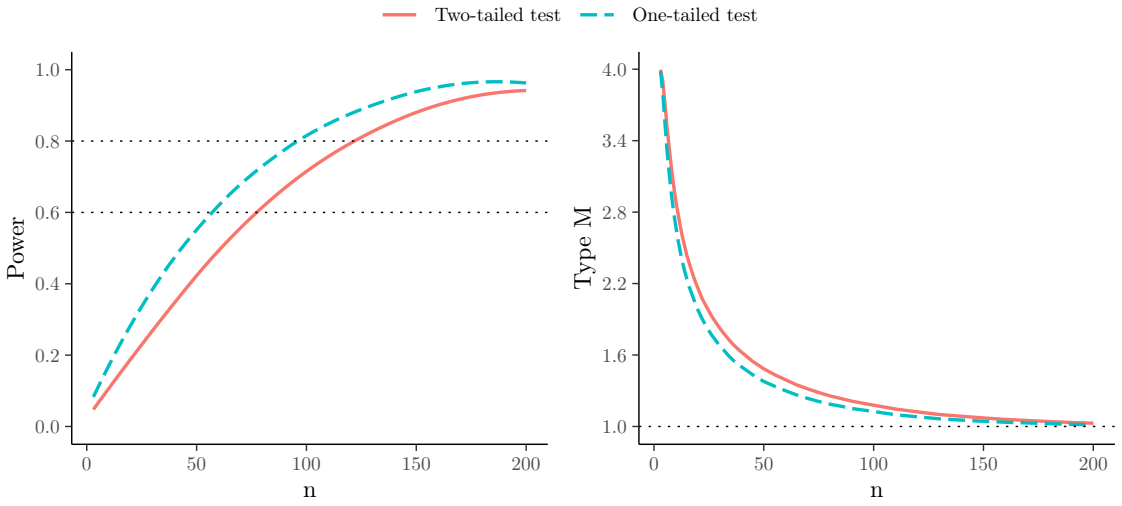

Figure 5: Comparison of Type M error rate and Power level between one tailed and two tailed test with $\rho = .25$, $\alpha = .05$. n = sample size.

```
> pro_r(rho = .25, power = .8, alternative = "greater", sig_level = .05, seed = 2020)

        Design Analysis

Hypothesized effect: rho = 0.25

Study characteristics:
   n    alternative   sig_level
   94   greater       0.05

Inferential risks:
   power   typeM   typeS
   0.793   1.14    0

Critical value: r =  0.171
```

Figure 6: Input and Output of the function `pro_r()` for prospective design analysis. Plausible correlation coefficient is $\rho = .25$, statistical power is 80% and the statistical test is one-tailed.

Valid conclusions require decisions on test directionality and α level to be taken a priori, not while data are being analyzed (Cohen, 1988). These decisions can take place during a prospective design analysis,

which aligns with the increasing interest in psychological science to transparently communicate and justify design choices through studies' preregistration in public repositories (e.g., Open Science Framework; Aspredicted.com). Preregistration of studies' protocol is particularly valuable for researchers endorsing an error statistics philosophy of science, where the evaluation of research results takes into account the severity with which claims are tested (Lakens, 2019; Mayo, 2018). Severity depends on the degree with which a research protocol tries to falsify a claim. For example, a one-tailed statistical test provides greater severity than a two-tailed statistical test. As noted by Lakens (2019), preregistration is important to openly share a priori decisions, such as test-directionality, providing valuable information for researchers interested in evaluating the severity of research claims.

8 Discussion and Conclusion

In the scientific community, it is quite widespread the idea that the literature is affected by a problem with effect size exaggeration. This issue is usually explained in terms of studies' low statistical power combined with the use of thresholds of statistical significance (Button et al., 2013; Ioannidis, 2008; Ioannidis et al., 2013; Lane & Dunlap, 1978; Yarkoni, 2009; Young, Ioannidis, & Al-Ubaydli, 2008). Statistically significant results can be obtained even in underpowered studies and it is precisely in these cases that we should worry the most about issues of overestimation. Type M and Type S errors quantify and highlight the inferential risks directly in terms of effect size estimation, which are implied by the concept of statistical power, but might not be recognizable outright. So far, only a handful of papers explicitly mentioned Type M and Type S errors (Altoè et al., 2020; Gelman, 2018; Gelman & Carlin, 2013, 2014; Gelman et al., 2017; Gelman & Tuerlinckx, 2000; Lu et al., 2018; Vasishth, Mertzen, Jäger, & Gelman, 2018). With the broader goal of facilitating their consideration in psychological science, in the present contribution we illustrated how Type M and Type S errors are considered in a design analysis using one of the most common effect size measures in psychology, Pearson correlation coefficient.

Peculiar to design analysis is the focus on the implications of design choices on effect sizes estimation rather than statistical significance only. We illustrated how Type M and Type S errors can be prevented with a *prospective design analysis*. In the planning stage of a research project, design analysis has the potential to increase researchers' awareness of the consequences that their sample size choices have on uncertainty about final estimates of the effects. This favors reasoning in similar terms to those in which results will be evaluated, that is to say, effect size estimation. But understanding the inferential risks in a study design is also beneficial once results are obtained. We presented *retrospective design analysis* on a published study, and the same process can be useful for studies in general, especially those ending without the necessary sample size to maximize statistical power and minimize Type M and Type S errors. In all cases, presenting their values, effectively communicates the uncertainty of the results. In particular,

Type M and Type S errors put a red flag when results are statistically significant, but the effect size could be largely overestimated and in the wrong direction. Finally, both prospective and retrospective design analysis favors cumulative science encouraging the incorporation of expert knowledge in the definition of the plausible effect sizes.

To make design analysis accessible to the research community, we provide the R functions to perform prospective design analysis and retrospective design analysis for Pearson correlation coefficient (Todo:) (link) together with a short guide on how to use the R functions and a summary of the examples presented in this contribution (Appendix B).

(Todo:) Aggiunta revisione Altoè

Choices regarding studies' design impact effect size estimation. Type M (magnitude) error and Type S (sign) error directly quantify these inferential risks. Their consideration in a prospective design analysis increases awareness of what are the consequences of sample size choice reasoning in similar terms to those used in results evaluation. Instead, retrospective design analysis provides further guidance on interpreting research results. More broadly, design analysis reminds researchers that statistical inference should start before data collection and does not end when results are obtained.

(Todo:) rivedere numerazione paragrafi e sottoparagrafi

References

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis. *Frontiers in Psychology*, *10*. doi: https://doi.org/10.3389/fpsyg.2019.02893

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., & Munafò, M. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... 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, T.-H., Huber, J., Johannesson, M., ... 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. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. doi: https://doi.org/10.1016/j.jesp.2015.10.012

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292. doi: https://doi.org/10.1126/science.1089134

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes*. Cambridge University Press.

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507. doi: 10.2307/2331838

Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, *44*(1), 16–23. doi: https://doi.org/10.1177/0146167217729162

Gelman, A. (2019). *Embracing Variation and Accepting Uncertainty: Implications for Science and Metascience* [Video]. (Metascience 2019 Symposium: The Emerging Field of Research on the scientific Process. September 5th–8th, Stanford University. https://www.metascience2019.org/presentations/andrew-gelman/)

Gelman, A., & Carlin, J. (2013). *Retrospective design analysis using external information* [Unpublished].

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. 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., Skardhamar, T., & Aaltonen, M. (2017). Type M Error Might Explain Weisburd's Paradox. *Journal of Quantitative Criminology*. doi: https://doi.org/10.1007/s10940-017-9374-5

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373–390. doi: https://doi.org/10.1007/s001800000040

Gigerenzer, G., Krauss, S., & 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 (pp. 392–409). 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. doi: https://doi.org/10.4135/9781412986311.n21

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated:. *Epidemiology*, *19*(5), 640–648. doi: https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Pereira, T. V., & Horwitz, R. I. (2013). Emergence of Large Treatment Effects From Small Trials—Reply. *JAMA*, *309*(8), 768–769. doi: https://doi.org/10.1001/jama.2012.208831

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. doi: https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., ... Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. doi: https://doi.org/10.1177/2515245918810225

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis* (Preprint). PsyArXiv. doi: 10.31234/osf.io/jbh4w

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., ... Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. doi: 10.1038/s41562-018-0311-x

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 107–112. doi: https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

Lu, J., Qiu, Y., & Deng, A. (2018). A note on Type S/M errors in hypothesis testing. *British Journal of Mathematical and Statistical Psychology*. doi: https://doi.org/10.1111/bmsp.12132

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (First ed.). Cambridge University Press. doi: 10.1017/9781107286184

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. doi: https://doi.org/10.1126/science.aac4716

Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. In *Psychological science under scrutiny: Recent challenges and proposed solutions.* (pp. 85–107). Wiley-Blackwell. doi: https://doi.org/10.1002/9781119095910.ch6

Schooler, J. (2014). Turning the Lens of Science on Itself: Verbal Overshadowing, Replication, and Metascience. *Perspectives on Psychological Science*, *9*(5), 579–584. doi: https://doi.org/10.1177/1745691614547878

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. doi: https://doi.org/10.1016/j.jml.2018.07.004

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. doi: https://doi.org/10.1111/j.1745-6924.2009.01125.x

Vul, E., & Pashler, H. (2017). Suspiciously high correlations in brain imaging research. In *Psychological science under scrutiny* (pp. 196–220). John Wiley & Sons, Ltd. doi: https://doi.org/10.1002/9781119095910.ch11

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 294–298. doi: https://doi.org/10.1111/j.1745-6924.2009.01127.x

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLOS Medicine*, *5*(10), 1–5. doi: https://doi.org/10.1371/journal.pmed.0050201

9 Session Information

```
sessionInfo(package = NULL)

## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] docstring_1.0.0  MASS_7.3-51.6    gtable_0.3.0     gridExtra_2.3    kableExtra_1.1.0
##  [6] forcats_0.5.0    stringr_1.4.0    dplyr_0.8.5      purrr_0.3.4      readr_1.3.1
## [11] tidyr_1.0.2      tibble_3.0.1     ggplot2_3.3.0    tidyverse_1.3.0  knitr_1.28
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3         lubridate_1.7.8   lattice_0.20-38   assertthat_0.2.1
##  [5] digest_0.6.25      R6_2.4.1          cellranger_1.1.0  backports_1.1.5
##  [9] reprex_0.3.0       evaluate_0.14     httr_1.4.1        pillar_1.4.3
## [13] rlang_0.4.5        readxl_1.3.1      rstudioapi_0.11   Matrix_1.2-17
## [17] tikzDevice_0.12.3 rmarkdown_2.1     splines_3.6.1     labeling_0.3
## [21] webshot_0.5.2      munsell_0.5.0     tinytex_0.22      broom_0.5.6
## [25] compiler_3.6.1     modelr_0.1.6      xfun_0.12         pkgconfig_2.0.3
## [29] mgcv_1.8-29        htmltools_0.4.0   tidyselect_1.0.0  roxygen2_7.1.0
## [33] codetools_0.2-16   fansi_0.4.1       viridisLite_0.3.0 crayon_1.3.4
## [37] dbplyr_1.4.3       withr_2.1.2       nlme_3.1-141      jsonlite_1.6.1
## [41] lifecycle_0.2.0    DBI_1.1.0         formatR_1.7       magrittr_1.5
## [45] scales_1.1.0       cli_2.0.2         stringi_1.4.6     farver_2.0.3
## [49] fs_1.4.1           xml2_1.2.2        ellipsis_0.3.0    generics_0.0.2
## [53] vctrs_0.2.4        tools_3.6.1       glue_1.3.1        hms_0.5.3
## [57] colorspace_1.4-1   filehash_2.4-2    rvest_0.3.5       haven_2.2.0
```

Appendix A: Pearson Correlation and Design Analysis

To conduct a design analysis, it is necessary to know the sampling distribution of the effect of interest not only under the Null-Hypothesis (H_0), but also under the alternative Hypothesis (H_1). The sampling distribution is the distribution of effects we would observe if n observations were sampled over and over again from a population with a given effect. This allows us to evaluate the statistical power and inferential risks of the study considered.

Regarding Pearson's correlation between two normally distributed variables, the sampling distribution is bounded between -1 and 1 and its shape depends on the values of ρ and n, repectively the population correlation value and the sample size. The sampling distribution is approximately Normal if $\rho = 0$. Whereas, for positive or negative values of ρ, it is negatively skewed or positively skewed, respectively. Skewness is greater for higher absolute values of ρ but decreases when larger sample sizes are considered. In Figure A.1, correlation sampling distributions are presented for increasing values of ρ and fixed sample size ($n = 30$).

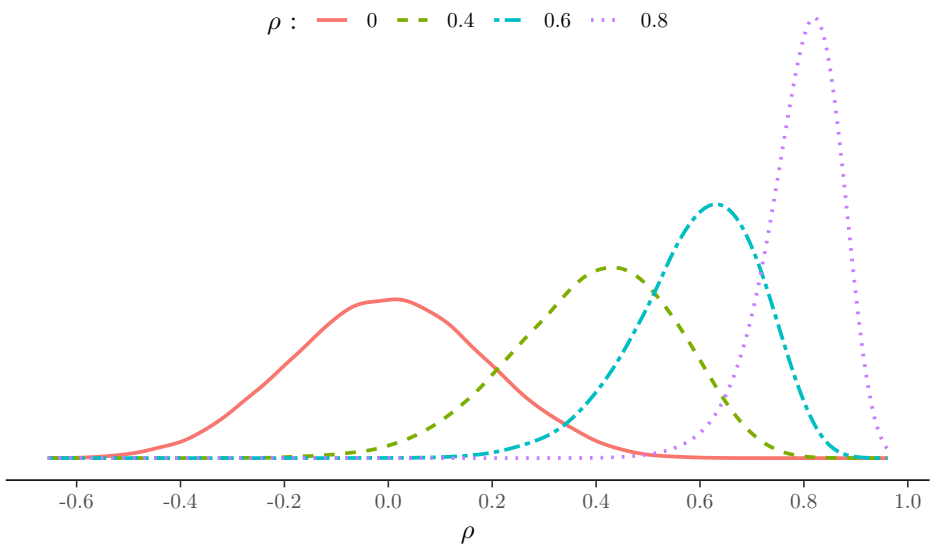

Figure A.1: Pearson Correlation coseefficient sampling distributions for increasing values of ρ and fixed sample size ($n = 30$)

In the following paragraphs, we consider the consequence of Pearson's correlation sampling distribution on statistical inference and the behaviour of Type M and Type S errors as a function of statistical power.

A.1 Statistical inference

To test an hypothesis or to derive confidence intervals, it is necessary that the sampling distribution of the effect of interest follows a known distribution. In the case of $H_0 : \rho = 0$, the sample correlation is approximately Normally distributed with Standard Error: $\text{SE}(r) = \sqrt{(1-r^2)/(n-2)}$. Thus, statistical inference is performed considering the t-statistic:

$$t = \frac{r}{\text{SE}(r)} = r\sqrt{\frac{n-2}{1-r^2}}, \tag{A.1}$$

that follows a t-distribution with $df = n - 2$.

However, in the case of $\rho \neq 0$, assuming normality would lead to unreliable results. As we have previously seen, for large values of ρ and small sample sizes the sampling distribution is skewed. To overcome this issue, the Fisher transformation was introduced (Fisher, 1915):

$$F(r) = \frac{1}{2}\, ln\frac{1+r}{1-r} = \text{arctanh}(r). \tag{A.2}$$

Applying this transformation, the resulting sample distribution is approximately Normal with mean $= F(\rho)$ and SE $= \frac{1}{\sqrt{n-3}}$. Thus, statistical inference is performed considering the Z-scores.

Alternatively, other methods can be used to obtain reliable results, as for example Monte Carlo simulation. Monte Carlo simulation is based on random sampling to approximate the quantities of interest. In the case of correlation, n observations are iteratively simulated from a bivariate normal distribution with a given ρ, and the observed correlation is considered. As the number of iterations increases, the distribution of simulated correlation values approximates the actual correlation sampling distribution and it can be used to compute the quantities of interest.

Although Monte Carlo methods are more computationally demanding than analytic solutions, this approach allows us to obtain reliable results in a wider range of conditions even when no closed-form solutions are available. For these reasons, the functions `pro_r()` and `retro_r()`, presented in this paper, are based on Monte Carlo simulation to compute power, Type M, and Type S error values. This guarantees a more general framework where other future applications can be easily integrated into the functions.

A.2 Type M and Type S errors

Design Analysis was first introduced by Gelman and Carlin (2014) assuming that the sampling distribution of the effect estimate follows a t-distribution. This is the case, for example, of Cohen's d effect size. Cohen's d is used to measure the mean difference between two groups on a continuous outcome. The behaviour of Type M and Type S errors as a function of statistical power in the case of Cohen's d is presented in Figure A.2.

For different values of hypothetical population effect size (Todo:) ($d = .2, .5, .7, .9$), we can observe

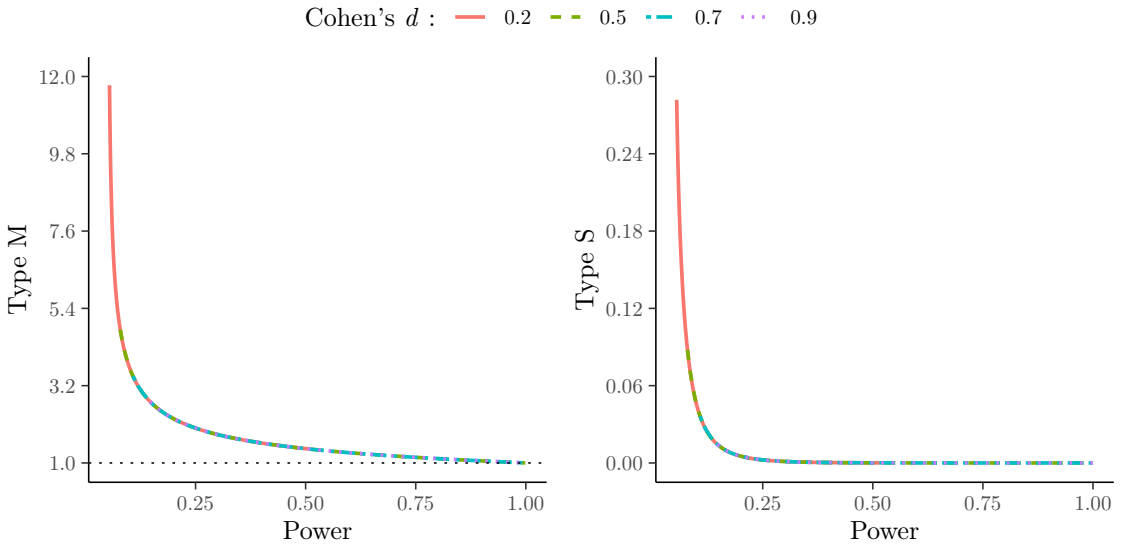

Figure A.2: The behaviour of Type M and Type S errors as a function of statistical power in the case of Cohen's *d*. Note that the four lines are overlapping.

that, for high levels of power, Type S and Type M errors are low. Conversely, the Type S and Type M errors are high for low values of power. In particular, the relation between power and inferential errors is not influenced by the value of d (i.e., the four lines are overlapping). Limit cases are obtained for power = 1 and 0.05 (note that the lowest value of power is given by the alpha value chosen as the statistical significance threshold). In the former case, Type S error is 0 and Type M error is 1. In the latter case, Type S error is 0.5 and the Type M error value goes to infinity.

In the case of Pearson's Correlation, we noted above that the sampling distribution is skewed for large values of ρ and small sample sizes. Moreover, the support is bounded between -1 and 1. Thus, the relations between power, Type M, and Type S error are influenced by the value of hypothetical population effect size (see Figure **??**).

We can observe how, for different values of correlation (Todo:) (ρ = .2, .5, .7, .9), Type M error increases at different rates when the power decrease, whereas Type S error follows a consistent pattern (note that differences are due to numerical approximation). We can intuitively explain this behaviour considering that, for low levels of power, the sampling distribution includes a wider range of correlation values. However, correlation values can not exceed the value 1 and therefore the distribution becomes progressively more skewed. This does not influence the proportion of statistically significant sampled correlations with the incorrect sign (Type S error), but it affects the mean absolute value of statistically significant sampled correlations (used to compute Type M error). In particular, sampling distributions for greater values of ρ becomes skewed more rapidly and thus Type-M error increases at a lower rate.

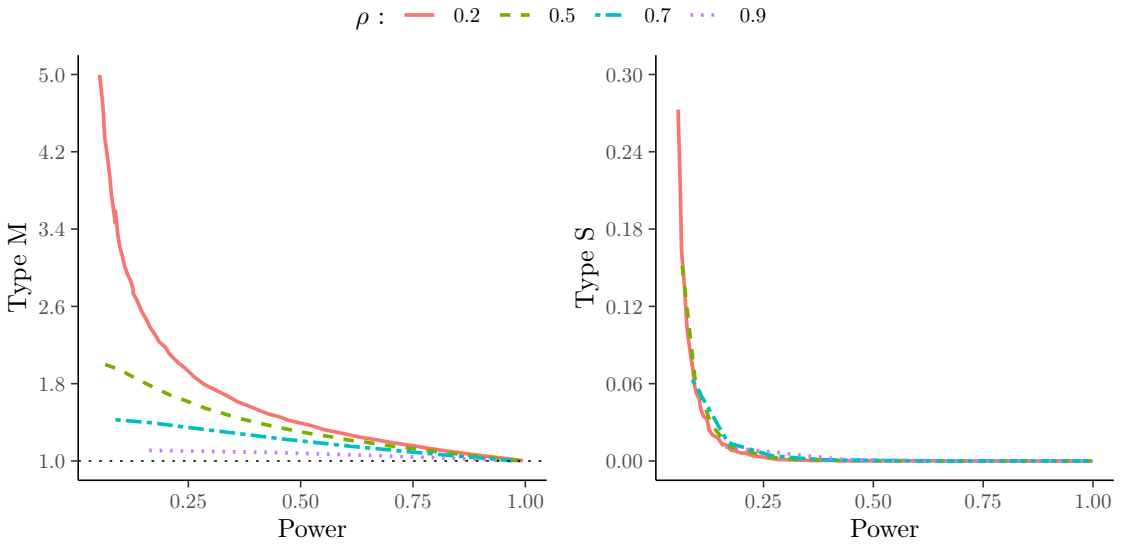

Figure A.3: The behaviour of Type M and Type S errors as a function of statistical power in the case of Peardson's correlation ρ.

Finally, since the correlation values are bounded, Type-M error for a given value of ρ can increase only to a maximum value given by $\frac{1}{\rho}$. For example, for $\rho = .5$ the maximum Type-M error is 2 as $.5 \times 2 = 1$ (i.e., the maximum correlation value).

In this appendix, we discussed for completeness the implications of conducting a Design Analysis in the case of Pearson's correlation effect size. We considered extreme scenarios that are unlikely to happen in real research settings. Nevertheless, we thought this was important for evaluating the statistical behaviour and properties of Type M and Type S error in the case of Pearson's correlation as well as helping researchers to deeply understand Design Analysis.

Appendix B: Pearson Correlation and Design Analysis