Simulate to stimulate statistical thinking

simulating p-values and their robustness to irregularities in the data

Jean-Jacques Orban de Xivry

May 28, 2019

Introduction

Statistical significance and the concept of p-values are at the heart of a philosophical debate (Wasserstein and Lazar 2016). Should we abandon the use of p-values (Szucs and Ioannidis 2017), should we use a more severe significance threshold (Benjamin et al. 2018), an adapted one (Daniel Lakens et al. 2018) or remove it entirely (Amrhein and Greenland 2018)? These questions are often hotly debated in scientific papers and social media. Yet, the problem with p-values is that people do not understand them properly or do not understand the circumstances of when to use a p-value (Colquhoun and London 2014; Gagnier and Morgenstern 2017; Greenland et al. 2016; Lyu et al. 2020). A p-value can be defined as follows: "*the P value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility*" (Greenland et al. 2016) or "*p-value is the probability that the data would be equal to or more extreme than what would be expected by chance.*". Yet, such definition does not provide a lot of insights for most people on how p-values behave in different circumstances. Here, we use our ability to generate random samples of different populations as a tool to get insights into p-values (Colquhoun and London 2014; Martins 2018; O'Hara 2019; Tintle et al. 2015). These simulations of randomly correlated or uncorrelated variables are done with correlations as the statistical test of interest. These simulations highlight the many problems that have been discussed in meta-science (the science of science): publication bias (Nissen et al. 2016), p-hacking (Head et al. 2015), correction for multiple comparisons (Makin and Orban de Xivry 2019), the dance of the p-values (Cumming 2011). A correct understanding of the p-value can provide leverage to the exploration of other related concepts such as power and effect sizes and how they are related to p-values.

Indeed, p-values on their own have very little (societal) value if they are not coupled to the effect size of an effect (Hubbard and Lindsay 2008). Indeed, while we sometimes want to know whether a relationship exist between two variables (e.g. air pollution and respiratory disease), we often also want to know the importance of this link (Calin-Jageman and Cumming 2019). Does hair pollution explains 0.1 or 25% of respiratory disease incidence. The ability to estimate the importance of such relationship is referred to as the (Gardner and Altman 1986). Estimation is important not for distinguishing between two theories but to assess the magnitude of the relationship (Sullivan and Feinn 2012). The estimation problem is made complicated by publication bias (Joober et al. 2012; Mlinarić, Horvat, and Smolčić 2017; Song, Hooper, and Loke 2013). That is, people rarely published a non-significant effect but readily publish significant effects. The importance of the estimation problems and how it is affected by publication bias will be addressed by using simulations to understand how publication bias threatens the estimation of effect size magnitude.

This publication bias is also linked to the presentation of results that are too good to be true (Francis 2013). Simulation of artificial data can also provide us with information about the type of results that one could expect (Daniël Lakens and Etz 2017). That is, if two variables are correlated and we run three times the same experiment, how many significant p-values do you expect? Note that tackling this question will tell what the expected patterns of p-values are in the presence of an effect.

The simulations and corresponding p-values reported in this paper are based on statistical test of correlation between two variables. Correlation was chosen as an example of statistical test because it is very frequently used in scientific papers but also suffers from a few pitfalls (Aggarwal and Ranganathan 2016). The correlation coefficient represents the strength of the association between two variables (Lee Rodgers and Nicewander 1988; Taylor 1990). While the correlation coefficient has been deemed to be robust against violations of the assumptions (Havlicek and Peterson 1976), we now know that it is at least extremely sensitive to outliers (Abdullah 1990; Pernet, Wilcox, and Rousselet 2013; Zimmerman 1994; Rousselet and Pernet 2012). That is, as we will see below, a slight irregularity in the data is sufficient to obtain significant but spurious correlations.

Therefore, the goal of this paper is to show that simulation of artificial data can be a valid tool to understand the behavior of p-values (O'Hara 2019; Tintle et al. 2015), to get insights into the estimation problem (Pernet, Wilcox, and Rousselet 2013), to compare different analysis techniques [Dorothy V. M. Bishop (2023)](Carter et al. 2019) and to find the best techniques to account for irregularities in the data (Rousselet and Pernet 2012) in order to avoid common statistical mistakes (Makin and Orban de Xivry 2019) or to fool ourselves (D. Bishop 2020). This paper is partially inspired by the blog posts of Prof. Dorothy Bishop: Blog 1, Blog 2, Blog3, Blog 4, Presentation

Learning about statistical concepts from simulated correlations.

Let's imagine the following experiment. Researchers from Klow in Syldavia[1] are willing to measure height and working memory capacity in a population of young healthy participants only in the city of Klow. Because of their limited budget, they are able to measure these parameters in a population of 15 participants. For each individual participant i, they have two observations: (x_i, y_i) where x_i is the height of the participant and y_i is his/her working memory capacity. Then, the researchers decided to test the correlation between these two variables for their sample population. There is no reason whatsoever to expect a correlation between height and working memory capacity. That is, we would expect this correlation to be equal to zero. We can simulate this experiment numerically by generating 15 random numbers corresponding to the simulated height of each participant and 15 different random numbers corresponding to the simulated working memory capacity of each participant. For simulations of each 15 random numbers, a mean of 0 and a standard deviation of 1 were chosen for simplicity but all the arguments below hold if one picks another mean and/or another standard deviation. By doing this, we have generated 15 pairs of height and working memory capacity (the measured parameters). To follow the idea of our Syldavian researchers, we can then compute the correlation between height and working memory capacity of our randomly generated parameters across the population of 15 individuals.

Simulating a single correlation between two randomly generated samples

To generate two independent variables (with zero correlations), we will implement the pseudo-algorithm explained above in function of the tools that we have:

1. Pick randomly 15 numbers from a Gaussian distribution (mean = 0 and standard deviation =1) and assign one of these numbers to the height of each of the participants (x_1 to x_{15})
2. Pick randomly 15 numbers from a Gaussian distribution (mean = 0 and standard deviation =1) and assign one of these numbers to the working memory capacity of each of the participants (y_1 to y_{15})
3. Compute the correlation (and the associated p-value) between height and working memory capacity (between x and y)

[1] inspired from the comics Tintin and the King Ottokar's Sceptre by Hergé.

By doing so, we obtain the following matrices:

Table 1: randomly generated parameters for the individuals of the population

	height (xi)	working memory capacity (yi)
Subj # 1	-1.4805676	-0.7034643
Subj # 2	1.5771695	1.1888792
Subj # 3	-0.9567445	0.3405123
Subj # 4	-0.9200052	0.5069682
Subj # 5	-1.9976421	-0.2933051
Subj # 6	-0.2722960	0.2236414
Subj # 7	-0.3153487	2.0072015
Subj # 8	-0.6282552	1.0119791
Subj # 9	-0.1064639	-0.3024592
Subj # 10	0.4280148	-1.0252448
Subj # 11	-0.7777196	-0.2673848
Subj # 12	-1.2938823	-0.1991057
Subj # 13	-0.7795665	0.1311226
Subj # 14	0.0119518	0.1457999
Subj # 15	-0.1524162	0.3620647

These values are unitless. Some of them are positive other negative values. Now that we have generated data for height and working memory capacity for each individual, we can test the correlation between these two parameters. We will first use the Pearson's correlation, which is used in most scientific studies. Off course, one should not expect any correlation between height and working memory capacity as these are sampled from random independent distributions. The two randomly generated parameters have a correlation coefficient r= 0.31 (CI: [-0.24, 0.71], p-value = 0.25). Interestingly, despite the two variables should be independent (height and working memory were generated randomly and independently), their correlation coefficient is not exactly zero. Yet, the value of the correlation coefficient is low (weak effect size) and the associated p-value informs us that there is no evidence that the correlation is different than zero (we cannot reject the null hypothesis). In other words, we cannot conclude from this data that the variables of height and working memory capacity are related in this population. However, we have to be careful not to interpret the absence of significance ($p>0.05$) as the absence of evidence for rejecting the null hypothesis is not evidence for the absence of an effect (Altman and Bland 1995). Therefore, we cannot conclude that these are not correlated.

Now, this is the data from a single city, Klow. Let's now simulate what would happen if the same correlation was computed for 10000 different cities across Syldavia (Sprodj, Niedzdrow, etc.) with enough inhabitants. In this case, we should repeat the process 10000 times to know whether the result obtained for Klow and illustrated on Figure 1 is a special case or not. That is, for every simulated city (N=10000), we will generate random numbers for the two parameters (height and working memory capacity). We will then obtain the size (correlation coefficient) and significance (p-value) of the correlation between these parameters for every one of the 10000 towns.

Simulating multiple correlations between two randomly generated parameters

Being able to perform these simulations in any programming language has certainly a large added value as we will show below. Therefore, we encourage the interested reader to use the code below to perform these simulations in R. For the readers who do not have the time or the willingness to do so, they can use the web application developed for this paper (See ShinyApp section below).

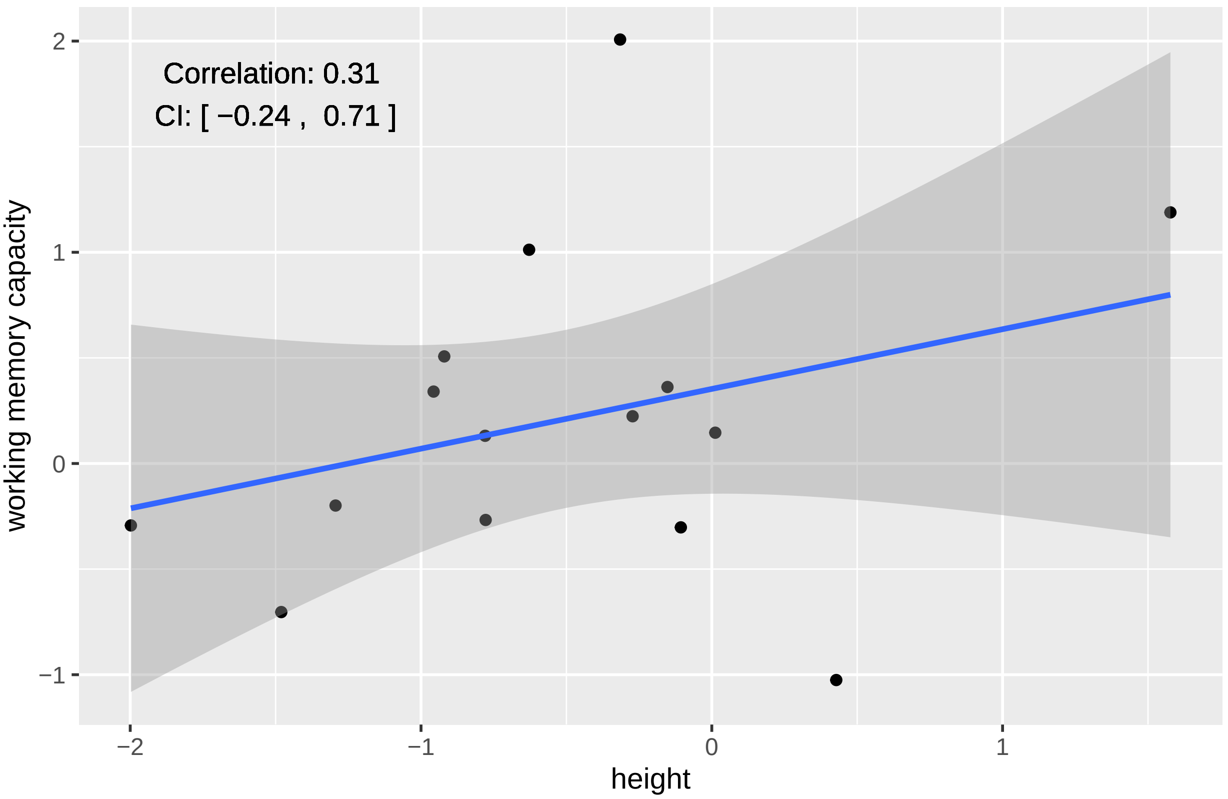


Figure 1: Figure 1: Absence of correlation from two randomly generated samples from two uncorrelated variables

Simulating data that in R

Simmulating randomly generated parameters for a given number of participants can be done in R with the *mvrnorm* function from the $MASS$ package (Venables and Ripley 2002).

To do so, we will use the ability of the *mvrnorm* function to simultaneously generate random values for the x and y uncorrelated variables. This can be done on the basis of the covariance matrix Sigma:

$$Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where the off-diagonal element of the covariance matrix indicate the theoretical value of the correlation between the variables x and y (Here, r=0). Using this covariance matrix, we will use the *mvnorm* function to build two uncorrelated variables at the same time.

```
# covariance matrix as defined above. Diagonal elements are equal to 1. Off-diagonal elements are zero.
Sigma <- matrix(c(1,0,0,1),2,2)
# using the mvrnorm function to produce a matrix whose first column provides us with the values of heig
D <- mvrnorm(n = PopSize, rep(0, 2), Sigma, empirical = FALSE) #PopSize is defined above. PopSize=15
x <- D[,1]
y <- D[,2]
```

We will then repeat this process 10000 times to obtain a reliable distribution of the p-values. This is repeated 10000 times (Nsim= 10000) in the code below:

```
set.seed(123)# to make sure everybody gets the same results
# Nsim=10000 # number of simulations --> defined above
# PopSize = 15 #number os elements in each sample --> defined above

# covariance matrix
Sigma <- matrix(c(1,0,0,1),2,2)
# pre-allocating space for the dataframe
res <- data.frame(p=rep(NA,Nsim),R=rep(NA,Nsim),Sig=rep(NA,Nsim))
# Repeating the produciton of random datasets Nsim times
for (i in c(1:Nsim)){
  # drawing 15 elements for each random variable (x and y) with a covariance matrix equal to Sigma
  # the first column of D corresponds to the samples of the x parameter (e.g. the height) and the secon
  D <- mvrnorm(n = PopSize, rep(0, 2), Sigma, empirical = FALSE)
  # correlation between the randomly generated x and y
  R<-rcorr(D[,1],D[,2])
  # storing the correlation value for future use
  res[i,"R"] <- R$r[1,2]
  # storing the p-value for future use
  res[i,"p"]<-R$P[1,2]
  # tag p-values smaller than the type II error threshold (here 0.05)
  res[i,"Sig"]<-res[i,"p"]<0.05 #
}
```

In the code above the R-dataframe *res* stores the results of the correlation between x (first column of D - $D[,1]$) and y (second column of D = $D[,2]$) for each iteration of the simulations. The value of the correlation coefficient ($R\$r[1,2]$) and the associated p-value ($R\$p[1,2]$) are stored in this dataframe. One can use this dataframe to generate the histogram illustrating the distribution of p-values (see Figure 2 below).

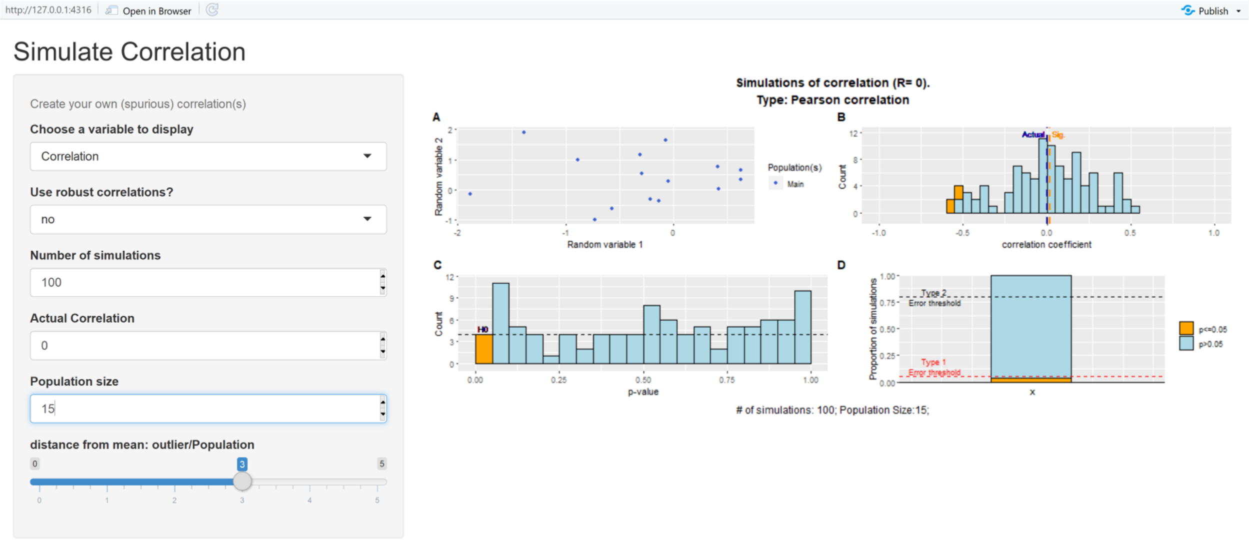
```
library(ggplot2)
# plotting the distribution of p-values
ggplot(res, aes(x=p, fill=Sig)) +
  geom_histogram(aes(x=p,fill=Sig),binwidth=BinSize,boundary = 0.05,color="black")+
  scale_fill_manual(values = Fcolor)
```

Using the ShinyApp

For the readers that are not (yet) familiar with the R language, one can simply follow the following steps to be able to interactively follow this tutorial:

1. download and install R and Rstudio (both needed) (e.g. https://courses.edx.org/courses/UTAustinX/UT.7.01x/3T2014/56c5437b88fa43cf828bff5371c6a924/)
2. install the shiny package (run the following from Rstudio console: `install.packages("shiny")`)
3. call the Shiny library (run the following from Rstudio console: `library(shiny)`)
4. run the following from an R console: runGitHub("SpuriousCorrelation","jjodx")

and you can reproduce all the manipulations explained below thanks to the program that is now open



Simulate Correlation

Create your own (spurious) correlation(s)

Choose a variable to display: Correlation

Use robust correlations?: no

Number of simulations: 100

Actual Correlation: 0

Population size: 15

distance from mean: outlier/Population: 3 (0–5)

Simulations of correlation (R= 0).
Type: Pearson correlation

# of simulations: 100; Population Size:15;

What is the expected p-value distribution if there are no correlations

Simulating such correlation repeatedly can allow one to look at the distribution of p-values when no correlation is present in the underlying variables. As described in many other papers (Benjamini and Hochberg 1995), this distribution is completely flat and 5% of the p-values are below the typical 5% significance threshold.

Interestingly, even in the absence of correlation between the parameters x and y, there is a subset of correlations associated with a p-value under the significance threshold (type I error). Those are highlighted in orange in Figure 2 and are called false-positives. That is, these correlations are significant despite the fact that the underlying independent variables x and y of 15 participants are, by design, not correlated. Given that the distribution of p-values is uniform and that p-values fall in the interval [0,1], there will always be 5% of the p-values under the threshold of 0.05, independently of the number of samples for the two variables.

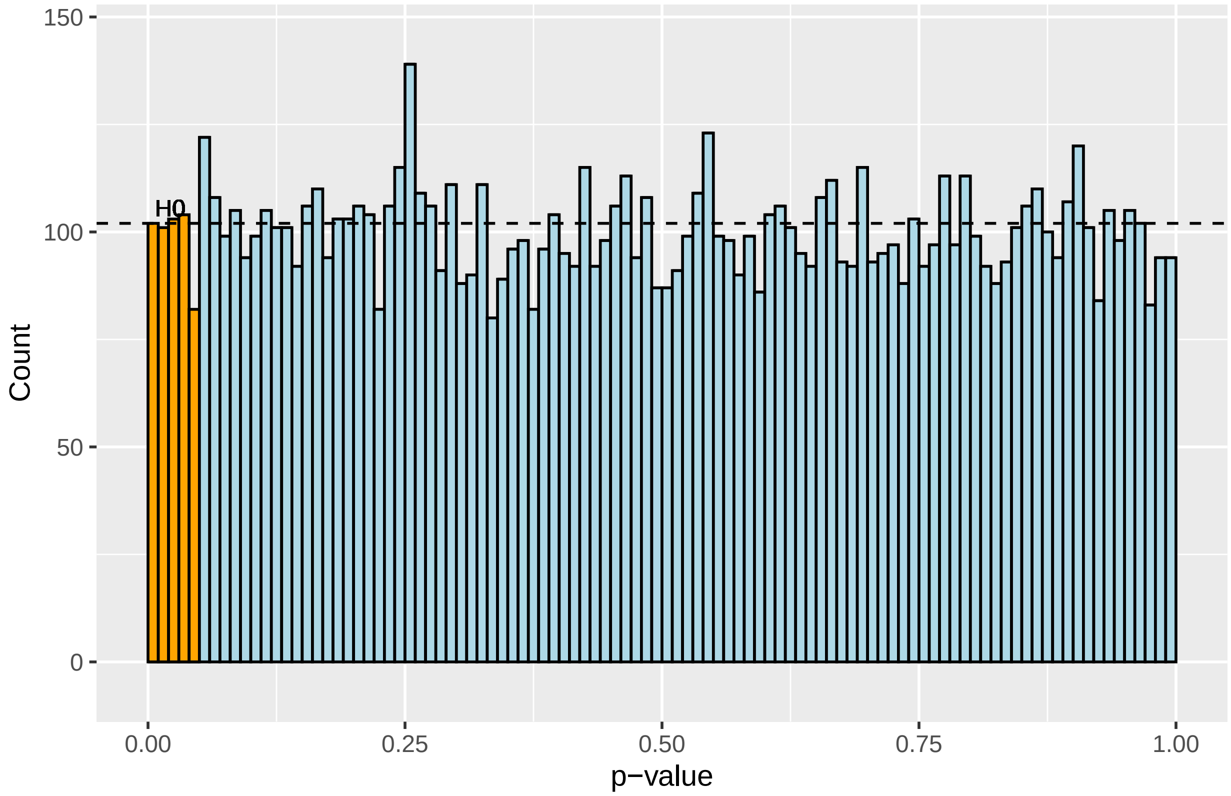


Figure 2: Figure 2: Distribution of p-values when no effect is present. Each simulated p-value corresponds to the correlation between samples taken from two uncorrelated variables

Assignment: *check that, when the two variables are uncorrelated, increasing the population size (value of PopSize in the code) does not influence the distribution of p-values.*

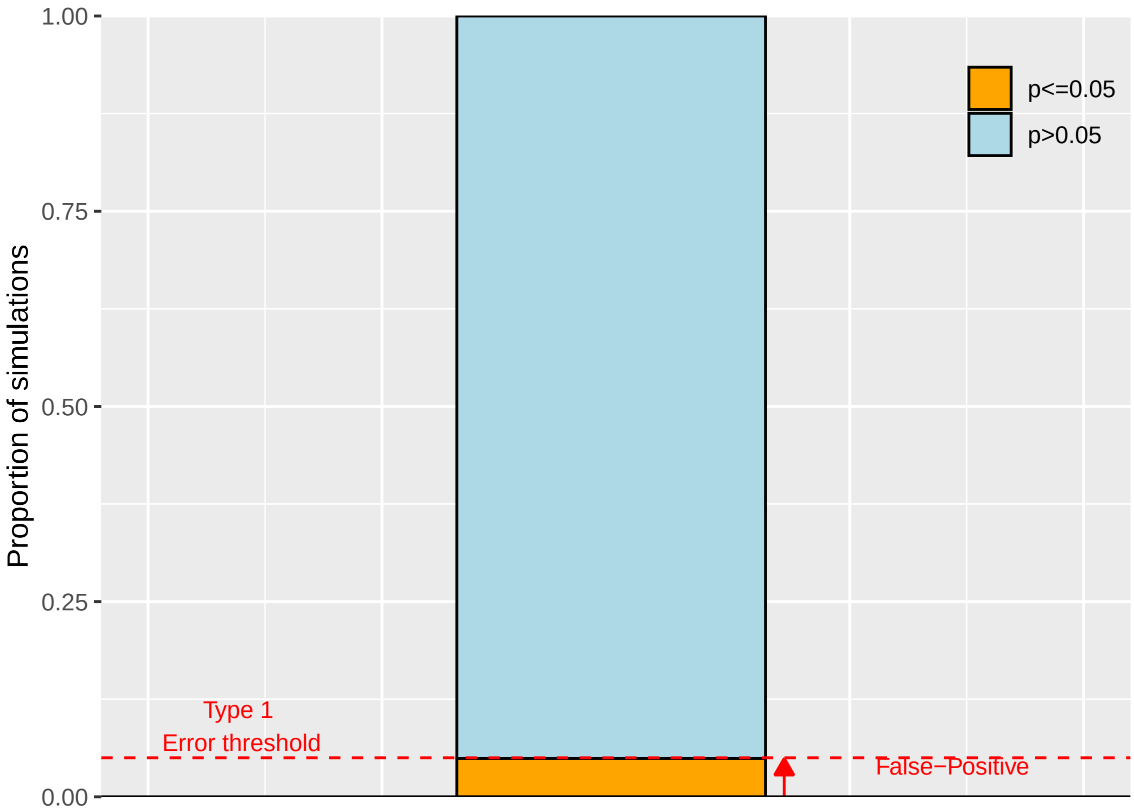


Figure 3: Figure 3: proportion of significant and non-significant p-values when the simulated R is 0 and the sample size is 15

In the absence of correlation between the underlying variables, the significant p-values are referred to as false-positive. We expect 5% of such correlations for a type I error threshold (aka the infamous significance threshold) of 0.05 (red horizontal dashed line in Figure 3).

Understanding the influence of sample size on the magnitude of the false-positive correlations

So far, we have looked at the distribution of the p-values associated with the correlation but not at the magnitude of the correlation coefficient. The distribution of the correlation coefficient will allow us to understand the problem with small sample size and why significant results always look convincing with small sample size. In Figure 4, one can see that the simulated correlations with 15 samples span a range between -0.83 and 0.88. However, only extreme correlations (in orange on Figure 4) are significant. That is, with small sample size (here 15 observations for each variable), the false-positive correlations are large (absolute value > 0.51) and look very convincing even though they are spurious.

This is a fallacy consistent with the law of small numbers (Tversky and Kahneman 1971): "*In evaluating replications, his or others', he has unreasonably high expectations about the replicability of significant results. He underestimates the breadth of confidence intervals.*" One could rephrase it as follows: *If it so big and significant, it can only be true.* It is therefore important to note that, when correlating two variables and looking at a high and significant correlation, it is impossible to say whether this is a true effect or a false-positive correlation. Replication of this correlation in an independent and hopefully larger sample can only

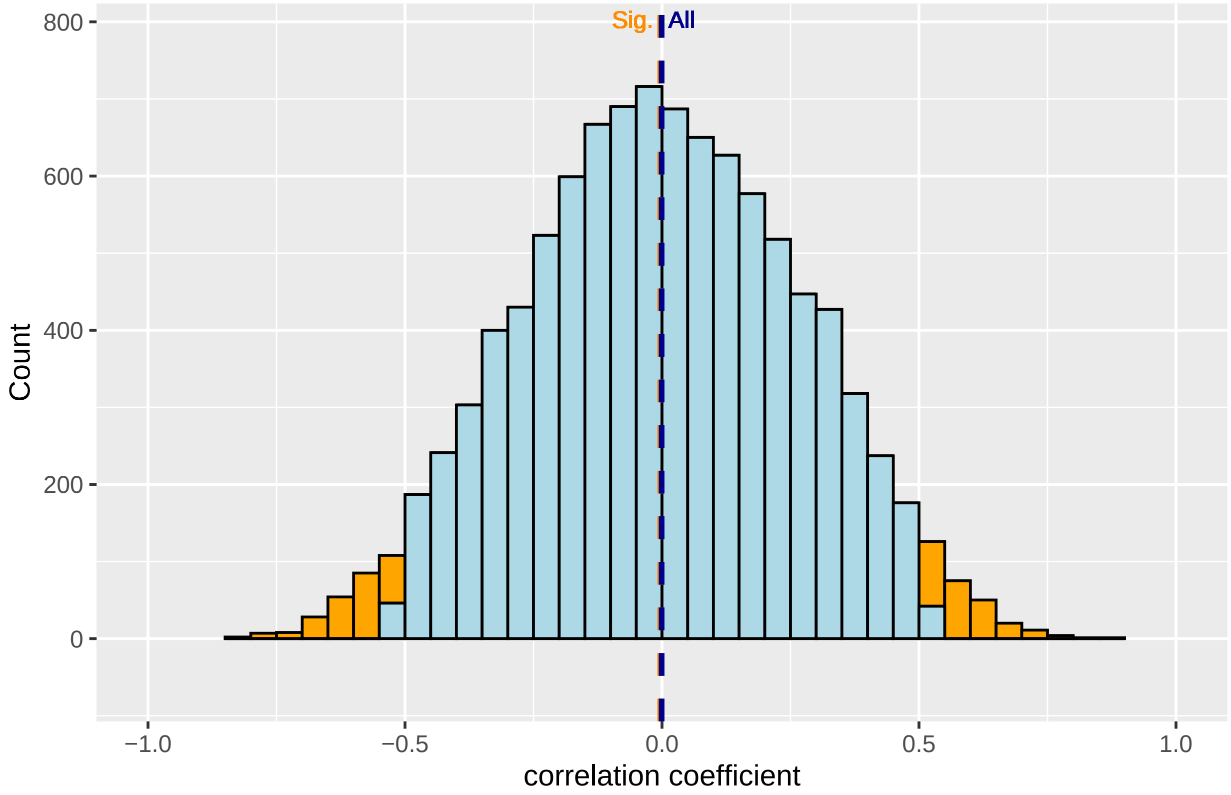


Figure 4: Figure 4: distribution of the correlations coefficient when r= 0 and N= 15

increase the probability that this correlation is true if it is again significant. In the other case, it increases the probability that the first correlation was a false-positive one. Yet, even two correlations are insufficient to draw a firm conclusion about the reliability of the effect but they at least give a better idea of the investigated correlation. In other words high-correlations on a limited number of data points should require skepticism for any scientists. I am well aware that it is not always easy to do so (see the beautiful correlation on N=6 in (Orban de Xivry, Missal, and Lefèvre 2009)).

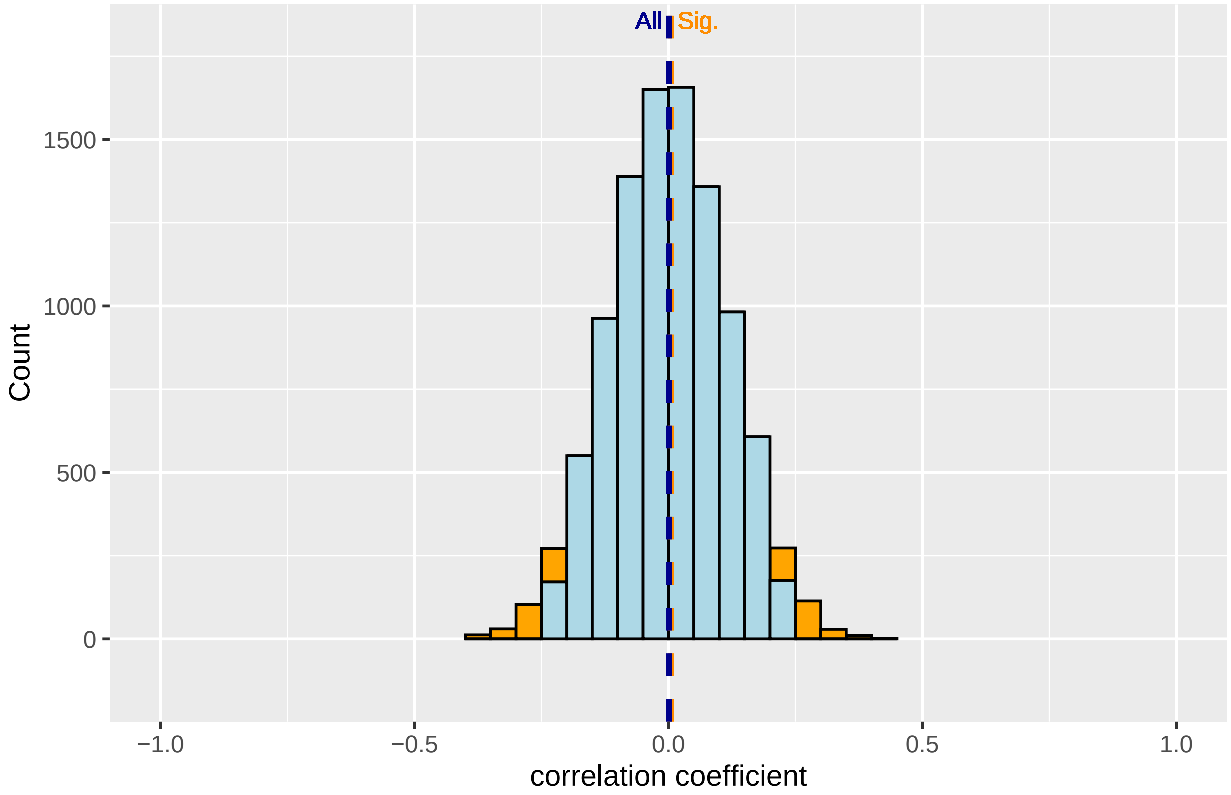


Figure 5: Figure 5: distribution of the correlations coefficient when R= 0 and the sample size is 75

I have insisted on the fact that the high false-positive correlations described above were linked to the small sample size used (N=15). Now, we will investigate the influence of sample size on the magnitude of these high false-positive correlations. To do so, we will repeat the simulations above but with a sample size of N=75 instead of N=15. As one can see on Figure 5, the value of the false-positive correlations has been dramatically reduced to absolute values larger than 0.23). In other words, if there is no effect, it is easy to obtain impressive but spurious correlations with low sample size while false-positive correlations are much less impressive with high sample size. Therefore, any scientist should be skeptical of exploratory correlations with low N.

> ***Assignment:*** *How big does the population size need to be to have all false positive correlations below a medium effect size (r=0.3)?*

Understanding the multiple comparison problem.

These simulations also show that if we keep generating random parameters, we are assured of getting at least one significant correlation. With 10000 simulations, there is almost 100% chance that at least one correlation will be significant, even if there is no correlation between the x and y variables. In our simulation, we observed one significant correlation after 14.5 tests (on average). That is, if one keeps testing the significance of correlation between different parameters, (s)he is assured of finding a significant correlation even if none actually exists. This is often referred to as the multiple comparison problem and has been famously used by scientists to show brain activity in a dead salmon (Bennett et al. 2011).

This illustrates that if someone keeps trying to test correlation between many possible variables, the probability of getting a significant correlation is 100% even if there should be no correlation.

> ***Assignment:*** *Repeatedly simulate a single correlation. How long did you have to wait to get at least one significant correlation? Now, simulate 50 different correlations. How often are none of the correlations significant? This demonstrates that it is not difficult to get a significant correlation even if the data is completely random.*

Simulating a single correlation r=0.5

It is also interesting to simulate cases where the correlation is not zero. While simulating correlations with r=0 taught us about false-positive correlations (type I error), the importance of sample size, the law of small numbers, and the multiple comparison problem, simulating correlation with $R \neq 0$ can teach us about power (and type II error), the precision of estimation and the consequences of publishing only significant studies.

To simulate two parameters with a given correlation, one can use the *mvnorm* function (Venables and Ripley 2002) but this time with a covariance matrix exhibiting the value of the simulated correlation on the off-diagonal elements. For instance, if we want to simulate two variables with a correlation of r=0.5, then the covariance matrix is:

$$Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

Table 2: randomly generated value for the individuals of the population

	height (xi)	weight
Subj # 1	-0.9304770	-1.6339413
Subj # 2	0.7714293	1.9603084
Subj # 3	-0.9988212	-0.6583089
Subj # 4	-1.0502320	-0.5432638
Subj # 5	-1.5833562	-1.8766614
Subj # 6	-0.3476360	-0.1239946
Subj # 7	-1.2767007	0.7305007
Subj # 8	-1.0500746	-0.0380954
Subj # 9	0.0590292	-0.2434301
Subj # 10	0.8832941	-0.1419507
Subj # 11	-0.5398325	-0.8072173
Subj # 12	-1.0209821	-1.2200878
Subj # 13	-0.7406857	-0.6095631
Subj # 14	-0.0625494	0.0832505
Subj # 15	-0.3130287	0.0490360

The generated values for the two parameters can be seen in the table above. One can see that individuals that appear to have a larger height also have a larger weight. This is even more apparent on Figure 6 where the relationship between height and weight is clearly visible.

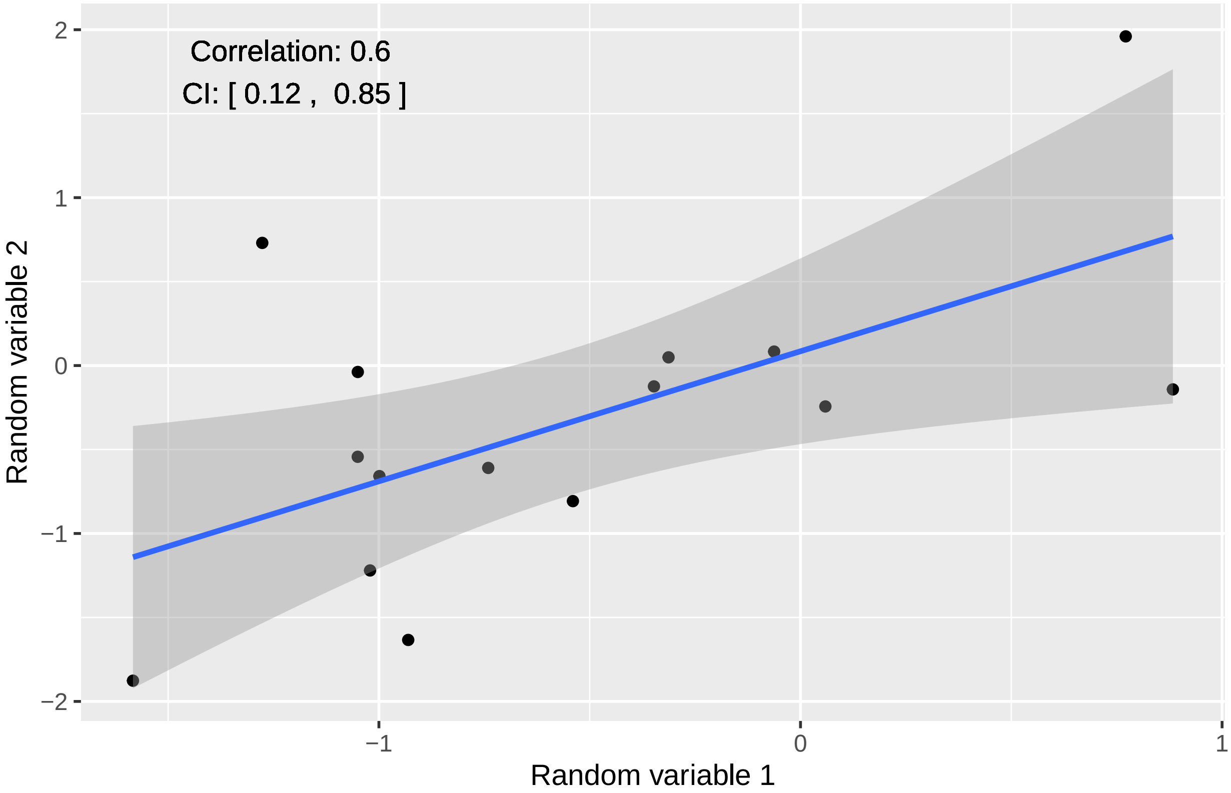


Figure 6: Figure 6: example of correlations when the simulated R is 0.5 and the sample size is 15

Together, these two parameters measured in 15 individuals are correlated with a correlation coefficient r= 0.6 (CI: [0.12, 0.85], p-value = 0.02). That is, the correlation between the two parameters (height and weight) is not exactly equal to the simulated value (r= 0.5). We will see later how we can recover the actual simulated correlation. This is reminiscent of when we generated uncorrelated parameters (Figure 1). Even though there was no correlation in the underlying variables (height and working memory capacity), the correlations that we found were not exactly zero but approximately zero.

What is the expected p-value distribution if there is a correlation r=0.5

Now, we will repeat this process a number of times in order to obtain the distribution of p-values when there is a correlation r=0.5, just as we did earlier when no correlation was present. When r=0, the distribution of p-values was completely flat (Figure 2). In contrast, when there is an underlying correlation between the two variables (r= 0.5), the distribution of p-values becomes skewed with more p-values under 0.05.

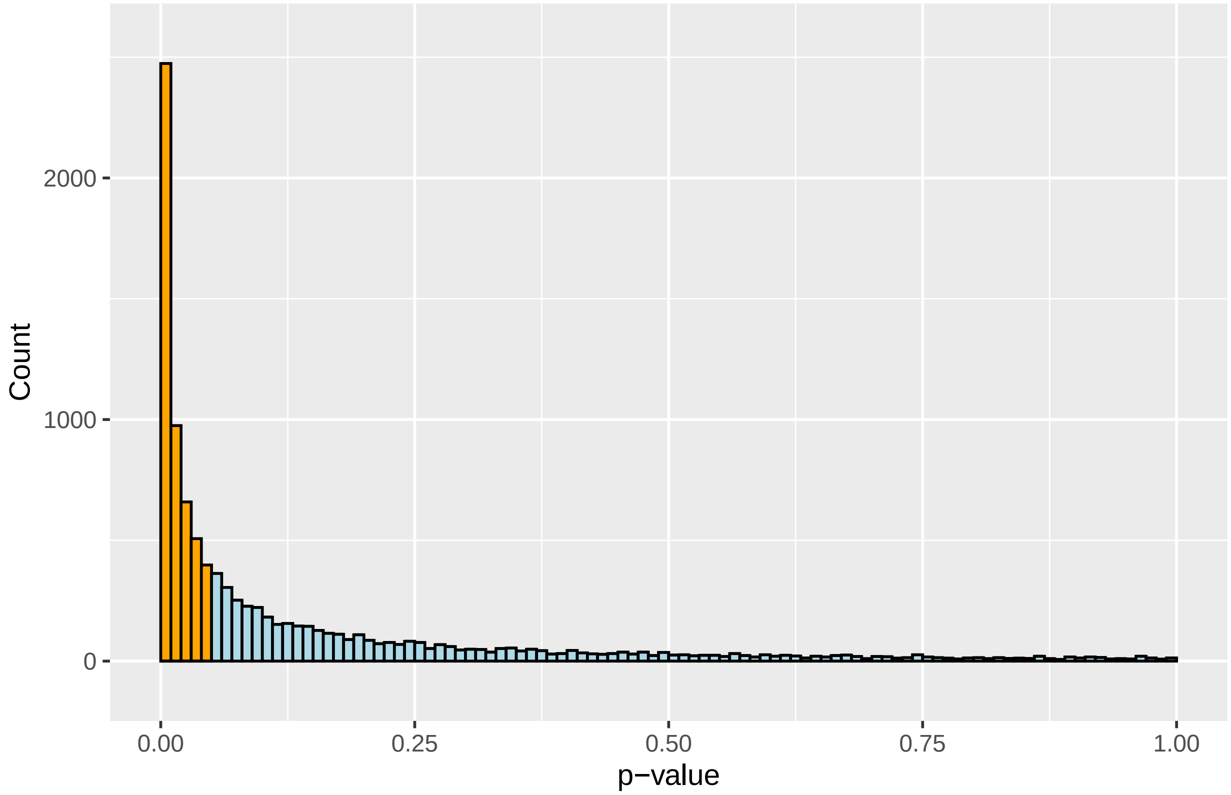


Figure 7: Figure 7: distribution of p-values when the simulated R is 0.5 and the sample size is 15

In this graph, it appears very clearly that p-values represents a continuous quantity and that there is nothing special about the 0.05 significance threshold. It also shows that in the presence of an effect, getting a p-value between 0.04 and 0.05 is not as likely as getting a p-value below 0.01. *This reinforces the statement that we should not dichotomize p-values in p<0.05 and p>0.05.* Figure 7 also shows that non-significant p-values cannot be considered as a proof that there are no correlation between the variables x and y. Indeed, despite the fact that the underlying variables are correlated (R = 0.5), a portion of the simulated correlations are not significant. The non-significant correlations on Figure 7 are false-negative. That is, these correlations are not significant even though the underlying variables x and y are correlated. False-negative correlations correspond to type II error, which is linked to the power of an experiment.

Assignment: *Explore the shape of the p-value distribution for correlation of different magnitudes and for different population sizes. Pay particular attention to the distribution of p-values below 0.05.*

Understanding power

The percentage of significant correlations is equal the power of an experiment designed to detect a correlation of r=0.5 with 15 observations. Indeed, the power of an experiment can be defined as the chance that an experiment designed to detect a given effect (here a correlation) will actually do so (100% minus power is also referred to as type II error or false negative). In our case, we performed as many experiments as we simulated correlations but only 50% of them were significant. That means that if we have 15 observations from two populations that are correlated with r=0.5, we have 50% of detecting a significant correlation (see Figure 8 for an illustration).

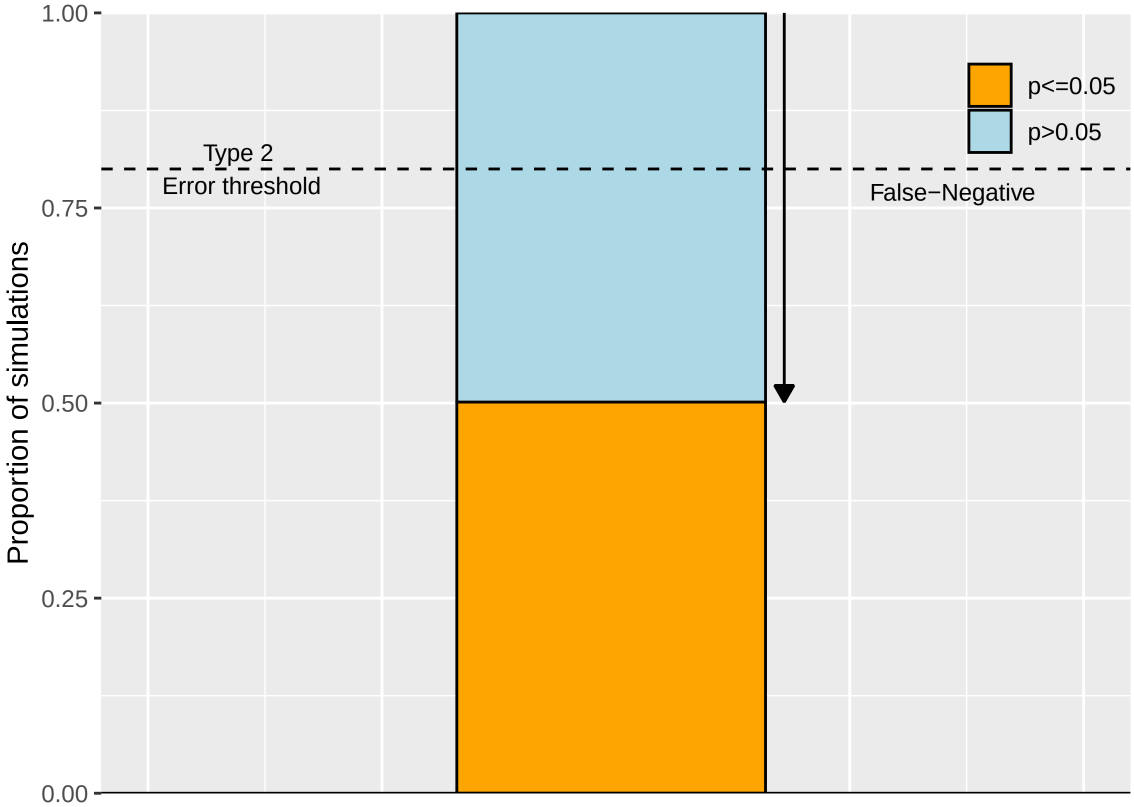


Figure 8: Figure 8: proportion of significant and non-significant p-values when the simulated R is 0.5 and the sample size is 15

As indicated on Figure 9, scientists often try to reach a power 80% (or 90%, or 95%) for an experiment (but they rarely do reach it (Button, Ioannidis, et al. 2013a)). Yet, this is here not the case. To detect a correlation r=0.5, one needs 28 observations. In this case, the percentage of significant correlations reaches the power requirement of 80%. The power of an experiment depends on the size of the effect (here, the strength of the correlation) and on the number of samples.

This manipulation demonstrates how simulations can yield insights about the required sample size for detecting a correlation of a given value (Here, r=0.5). The problem with power is that we rarely know in advance the effect size magnitude (e.g. the strength the correlation).

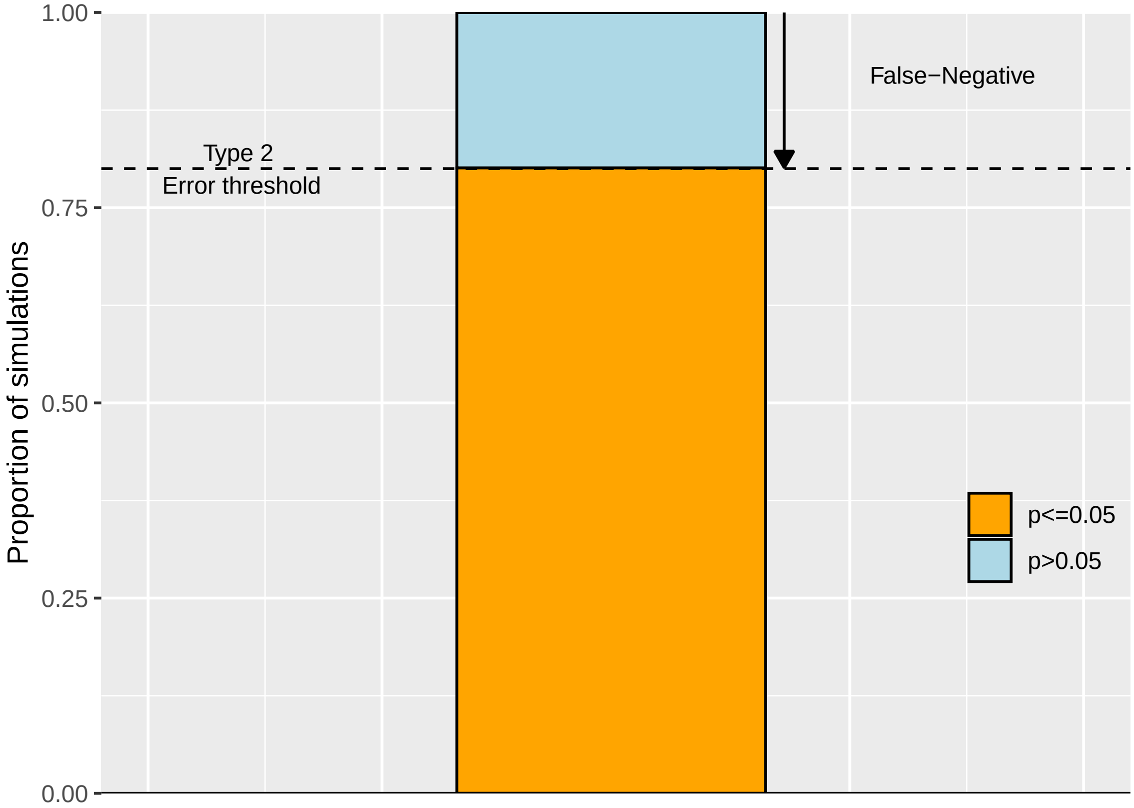


Figure 9: Figure 9: proportion of significant and non-significant p-values when the simulated R is 0.5 and the sample size is 28

Assignment: *How big needs the population size to be to reach a power of 80% for r=0.3 or r=0.1, a power of 90% for r=0.5?*

Effect of sample size on the accuracy of correlation estimation.

On Figure 10, we can see the distribution of correlation magnitude. Obviously, not all simulated correlations are exactly equal to 0.5 but are distributed around that value but are spread over a large interval. Indeed, the simulated correlations exhibited a large 95% confidence interval: between 0.01 and 0.81. One can look at the effect of sample size on the precision of the estimated correlation. It is obvious that increasing the sample size (N= 28) decreases the 95% confidence interval of simulated correlation ([0.17,0.74]). This has been described previously and illustrates that high power (here 80%) does not always correspond to high precision given the remaining variability of the simulated correlations.

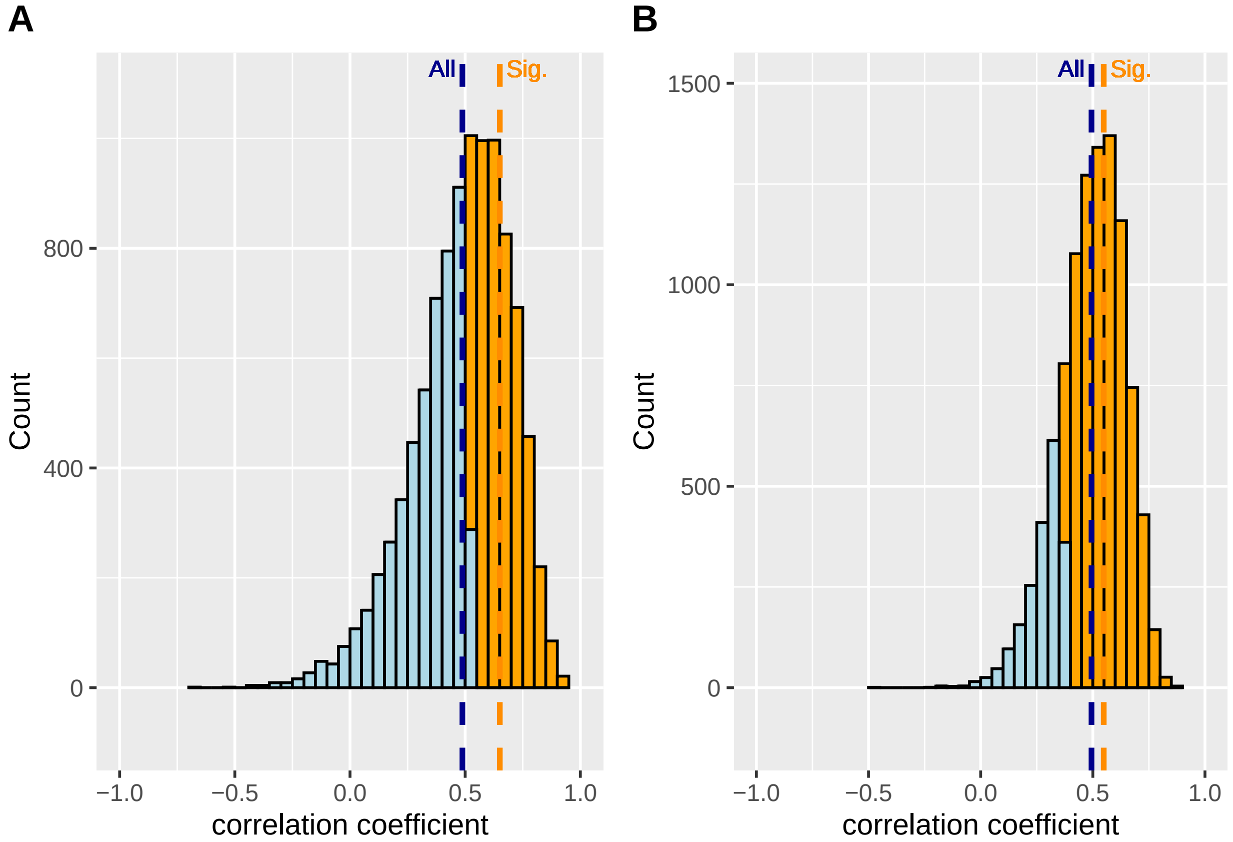


Figure 10: Figure 10: distribution of significant and non-significant correlation coefficients when the simulated R is 0.5 and the sample size is 15 (panel A) and 28 (panel B)

Understanding that publication bias inflates effect size

Figure 10 also indicates the importance of publishing non-significant findings. Indeed, if one looks at the average correlation for the significant correlations (orange dashed vertical line in right panel of Figure 10, with N= 28), this average correlation (mean r=0.55) overestimates the actual correlation value (r= 0.5). In contrast, the average correlation computed over the significant and non-significant correlations (dark blue dashed vertical line on the graphs above) is much closer to the actual simulated correlation (with N= 28, average correlation is 0.49). Furthermore, this overestimation of the actual correlation is even bigger with smaller sample sizes (with N= 15, average correlation from significant ones is r=0.65 while the average correlation from all studies is r= 0.49). That is, the mean correlation magnitude from significant studies only is always larger than the actual population correlation magnitude. With low power, significant effect sizes are even more inflated. Off course, many scientists are convinced that reporting non-significant experiments is less convincing. Polished stories due to selective reporting is a common communication device used to convince the reader (Corneille et al. 2023). Yet, this leads to results that are too good to be true (see section below).

> ***Assignment:*** *Explore the effect of population size and effect size (i.e. the correlation) on publication bias.*

Using confidence interval to assess the accuracy of the estimation

Confidence interval is one way to go beyond significance and to provide some information about the size of an effect (Calin-Jageman and Cumming 2019; Cumming 2011). The accuracy of the estimation of the correlation magnitude can be quantified by the associated confidence interval. The 95% confidence interval gives a range of plausible values for the correlation magnitude of the underlying population. This interval has a 95% chance of containing the true effect size (here, the true value fo the correlation). That is, 95% of the confidence interval contains the value r= 0.5. To check this definition, we computed the confidence interval for each correlation presented on panel B of Figure 9, and computed the proportion of simulated confidence intervals that included the true value of the correlation (r=0.5). Consistent with the definition, 95% of them did include the true effect size (see panel A of Figure 11).

While the confidence interval has a 95% chance of containing the true correlation magnitude, it does not have a 95% chance of containing the next simulated correlation magnitude. In other words, the probability that the correlation of a second experiment will fall into the confidence interval of the correlation of a first experiment is not 95%. Imagine that you perform a first experiment and obtained a confidence interval [0.12, 0.85] (as presented in Figure 6). What is then the chance that, if you try to replicate your first experiment with the same sample size, the correlation magnitude will fall into the confidence interval obtained by the first experiment? To simulate this case,we will test whether the correlation of the simulated correlation #n falls within the 95% confidence interval of correlation #n-1. The results show that there is ONLY a 84% probability that the magnitude of the next simulated correlation falls within the confidence interval of the previous one (see panel B of Figure 11).

We can also look at the effect of sample size by comparing the width of the confidence interval of the simulated correlations with N= 15 and N=28, respectively (Figure 12). The width of the confidence interval was computed as the difference between the upper and lower bound of the 95% confidence interval. For instance, the CI of the correlation coefficient reported in Figure 6 was equal to CI: [0.12, 0.85]. The width of this confidence interval is thus 0.73, which is pretty close to the median CI width when N=15.

Comparing the two panels of Figure 12 demonstrate that increasing the sample size increase the accuracy of the estimation of each individual correlation coefficient (confidence intervals become narrower) in addition to increasing the accuracy of the estimation over multiple experiments (spread of simulated correlations coefficient is smaller when sample size increases, compare both panels of Figure 10). This shows that increasing sample size has three major effects: 1) it increases power, 2) it improves accuracy of the effect size estimation at the population level and 3) it increases the accuracy of every individual correlation coefficient.

How can results be too good to be true?

It is clear from Figure 10 that, even in the presence of an effect, some correlations will turn out to be non-significant. Then, it is interesting to look at the probability of getting one or two non-significant results in a series of 5 experiments. To do so, we simulated 5 experiments 10000 times and then counted the proportion of significant correlations.

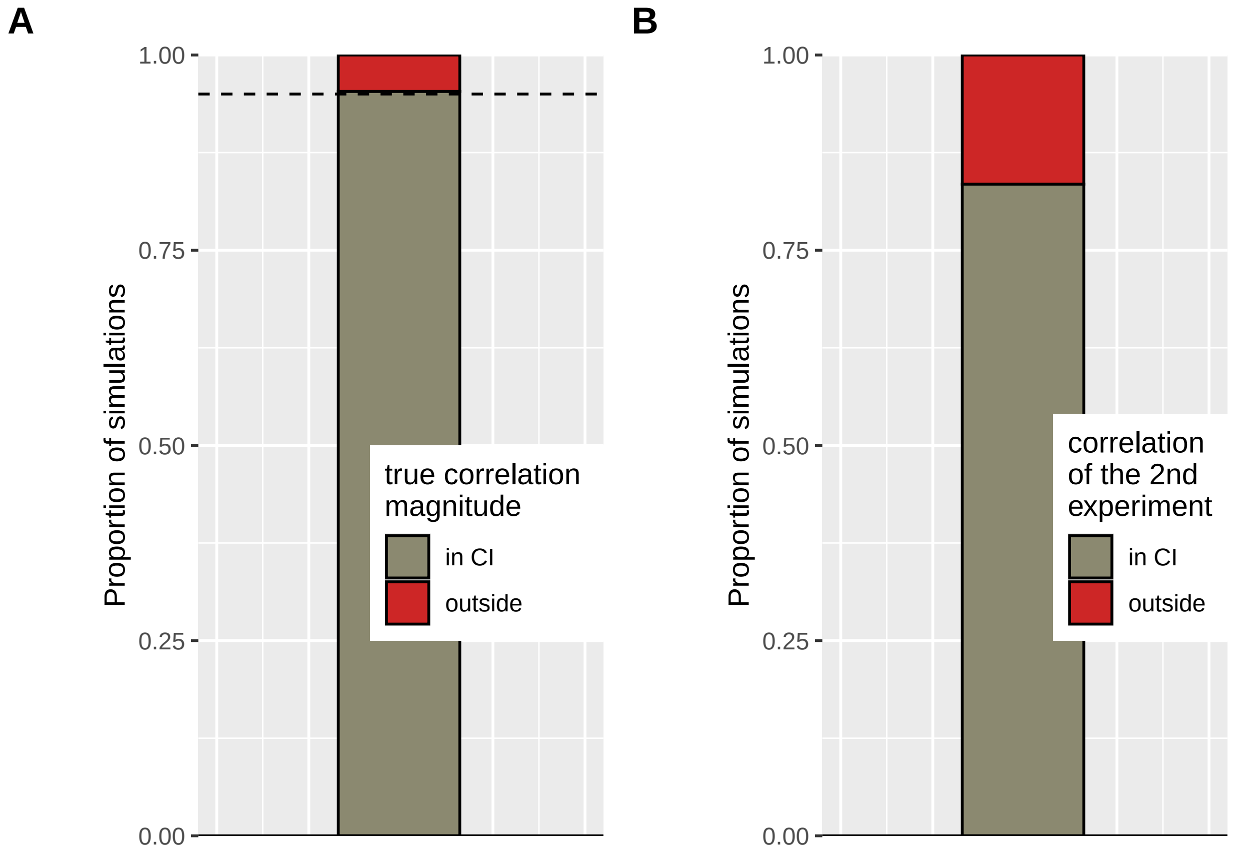


Figure 11: Figure 11: panel A: proportion of 95% confidence intervals including the actual effect size of r= 0.5 . Dashed black line is set at 95%. panel B: proportion of 95% confidence interval that contains the correlation coefficient of the next simulation. For both panels, the sample size used is 28 but this is true for any sample size

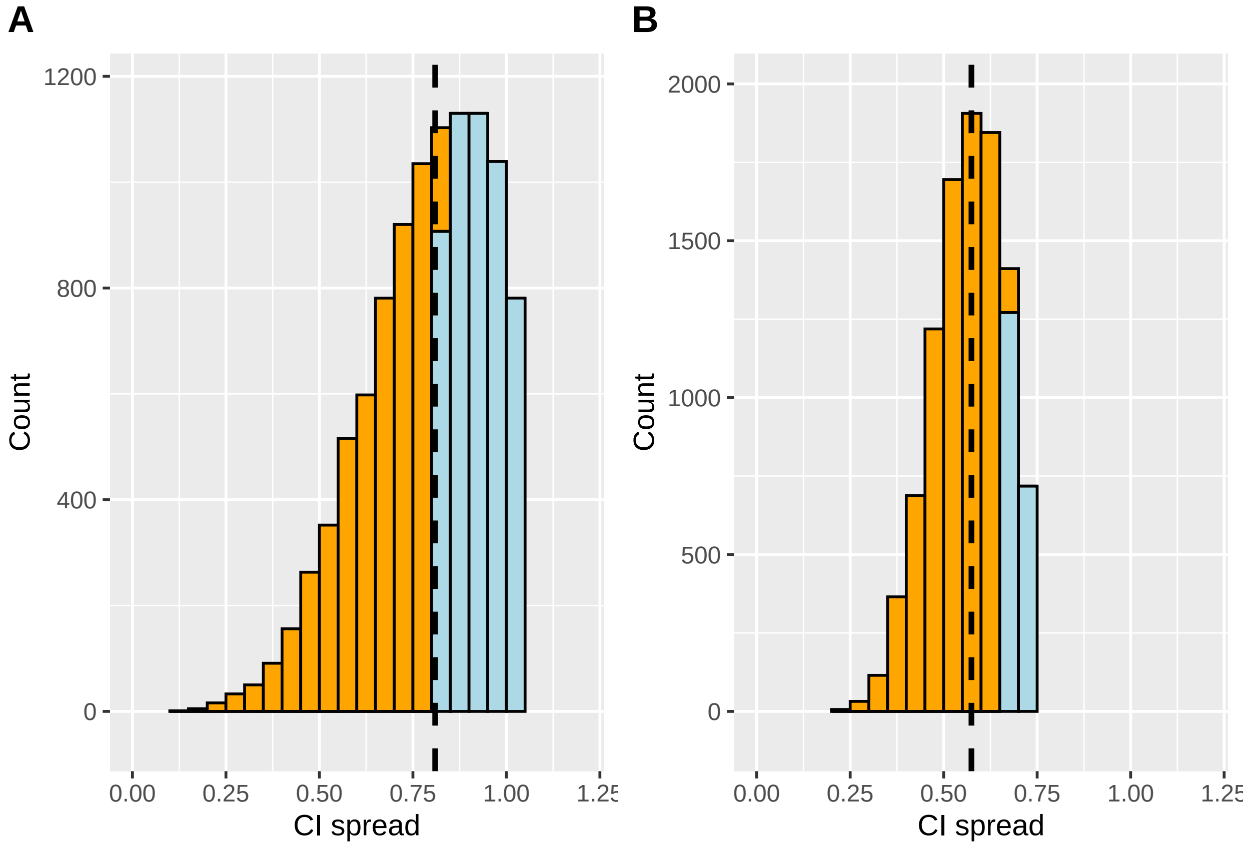


Figure 12: Figure 12: distribution of the width of confidence intervals around the correlations when the simulated R is 0.5 and the sample size is 15 (panel A) or 28 (panel B). Dashed vertical bar corresponds to the median CI width in both panels

To get a sense of the outcomes of this stimulation, one can use the Shiny App and repeatedly simulate 5 experiments (use the *generate again* button). In most of the cases, the 5 experiments do not yield 5 significant p-values. Actually, this only happens in 3.19% of the simulations.

> ***Assignment:*** *Select a number of simulated samples and see how often all the simulated correlations are significant. Test the effect of population size and correlation on this proportion.*

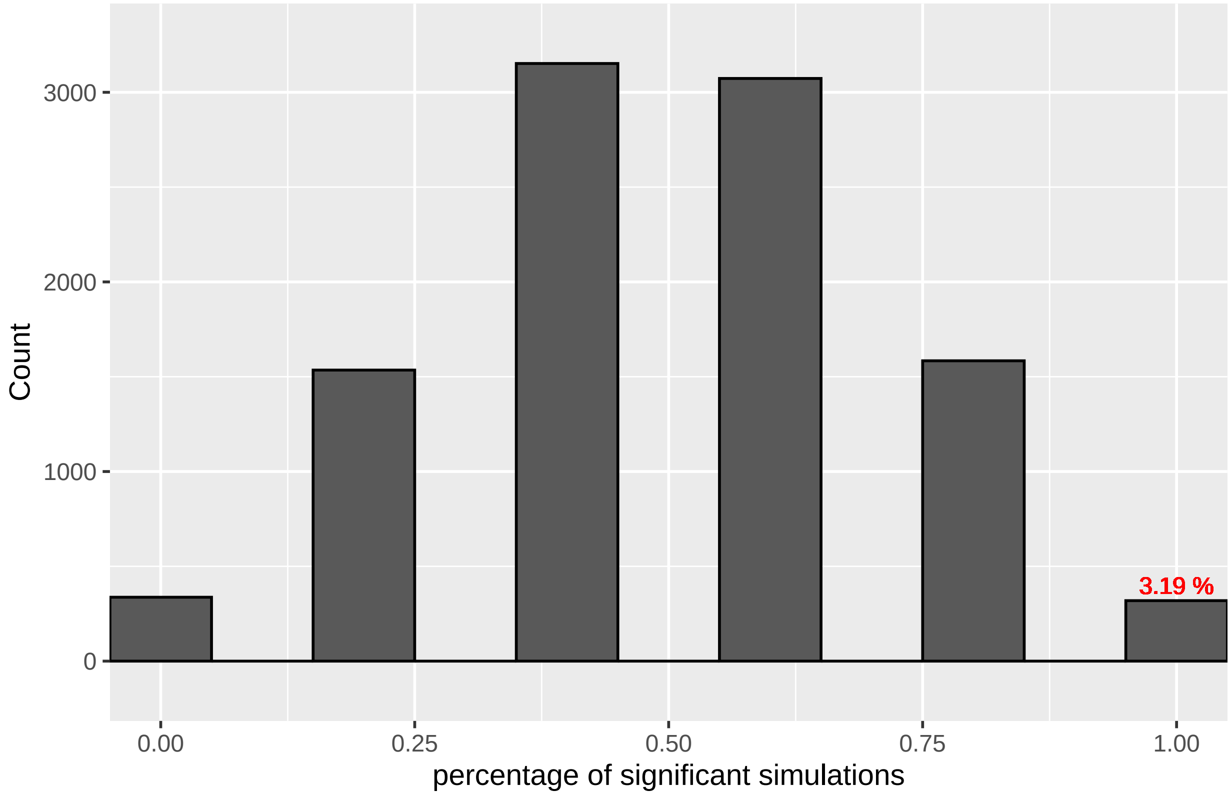


Figure 13: Figure 13: percentage of significant simulations among 5 simulations for R = 0.5 and sample size = 15

This result is really important because it means that scientists who are only presenting series of papers with only significant studies (often with small sample size, hence low power) are not telling you the entire truth (Francis 2013). Either they have plenty of studies in their file-drawer or they used questionable research practices to always obtain significant p-values. It is difficult to believe claims made from a series of only significant studies with small sample sizes (even across papers from the same laboratory). This cannot reflect the reality (Lane et al. 2016).

Note that the percentage of series of 5 simulations that are all significant increases with power and sample size (N=15: 3.19%; N=28: 32.63%). Yet, even with 80% power, these remain the minority (67.37% of the series of 5 simulations contain at least one non-significant p-value).

Furthermore, the probability of observing only significant p-values decreases with the number of experiments considered in one series. For instance, when considering a series of 10 experiments, the probability of observing 10 significant p-values is only 10.42%, despite having 80% power. Again, this reinforces the conclusion that every scientists should now and then report experiments that failed. Yet, we know this is rarely the case and these scientists present results that are likely better than the reality (Francis 2014).

Summary about p-values, power and effect size

In this first part, we started with the description of the p-value for a single correlation coefficient (Fisherian view). Then, we transitioned from one p-value to many p-values (N = 10000) in order to look at the distribution of p-values over multiple repetition of the same experiment (Neyman-Pearsonian view, (Goodman 1999)) and to discuss the concept of false-positives and false-negatives. Hopefully these many p-values brought the readers insights into how p-values behave and how they should be interpreted.

It is important to understand that the statistical problem here is well-defined. Indeed, in all simulations, we knew whether the two underlying variables were correlated or not. However, in science, the problem is often ill-defined. When a researcher performs a single experiment and get a p-value of 0.019, (s)he does not know whether this p-value corresponds to a true effect or to a false-positive correlation. This can only be determined at the end of several experiments albeit not all necessarily significant. As such, this suggests that we need a set of well-powered experiments to make a statement about the existence of an effect. Making statements about an effect is only possible if we get away from the law of small numbers (Tversky and Kahneman 1971), which appears to be difficult (D. V. M. Bishop, Thompson, and Parker 2022). In other words, we need several experiments (because they allow us to choose between two hypotheses, between the null effect or the existence of an effect) with large samples (because they provide more reliable estimates).

If we follow the same line of reasoning, one can investigate the distribution of p-values obtained from t-test aiming at comparing two independent groups. That is, instead of correlating the variables obtained from the random sampling, one can simply compare them with a t-test (the covariance matrix should have zeros on his off-diagonal elements). To simulate a difference between the two groups, one needs to change the mean of one of the groups.

> ***Assignment**: Let's test whether the attentive readers can solve the following problem (Gigerenzer 2018): A researcher designs an experiment with 50% power to detect an effect of a medium size and obtains a significant difference between means, p < .05. How likely is this researcher to replicate this effect in a second experiment with the same sample size (1) if there is an effect or (2) if there is no effect.*

> ***Assignment**: You read two papers, each reporting the results of three experiments. Which paper seems more plausible? A: p=0.024, p=0.034 and p=0.031 B: p=0.25, p=0.032 and p=0.0001 (source)*

> ***Assignment**: Which one is more likely to replicate (for the same effect size): a correlation of r=0.7 with N=10 or a correlation of r=0.3 with N=100?*

Using simulations to understand how irregularities in the data can affect p-values and to test possible solutions.

Looking at false-positive correlations or at power, as we have done above, is only correct if the studied p-values reflect an actual p-value. That is, this should not be artificially low because of the researchers' degrees-of-freedom or because the data do not conform to the assumptions underlying the statistical test used. The first point has been explored by others many times (Vrieze 2021; Gopalakrishna et al. 2022; Ravn and Sørensen 2021; Büttner et al. 2020). We will therefore focus the discussion on irregularities in the data and how irregularities in the data can yield spurious correlations. Indeed, correlations are very sensitive to violations of assumptions. Here, we will show that 1) in the absence of correlation (r=0), the presence of an outlier dramatically increases the false-positive rate (more than 5% of correlations are significant when r=0) (Abdullah 1990; Pernet, Wilcox, and Rousselet 2013; Zimmerman 1994; Rousselet and Pernet 2012); 2) in the absence of a correlation, data pooled between two different groups (e.g. young adults vs. old adults) with different means along the two variables will have a massive effect on the false-positive rate. In the presence of an effect, there is a large overestimation of the effect size in both cases (presence of an outlier or pooled data across two subgroups).

The outlier problem: simulation and solution

To generate an outlier, we will first generate the x and y parameters for the 15 individuals. We will then compute their mean and standard deviation, the data from the last individual will then be replaced by a point 3 standard deviation away from the mean (mean + 3*SD) for both x and y parameters. By doing so, we obtain a graphic with one point detached from the other points (red point in Figure 14).

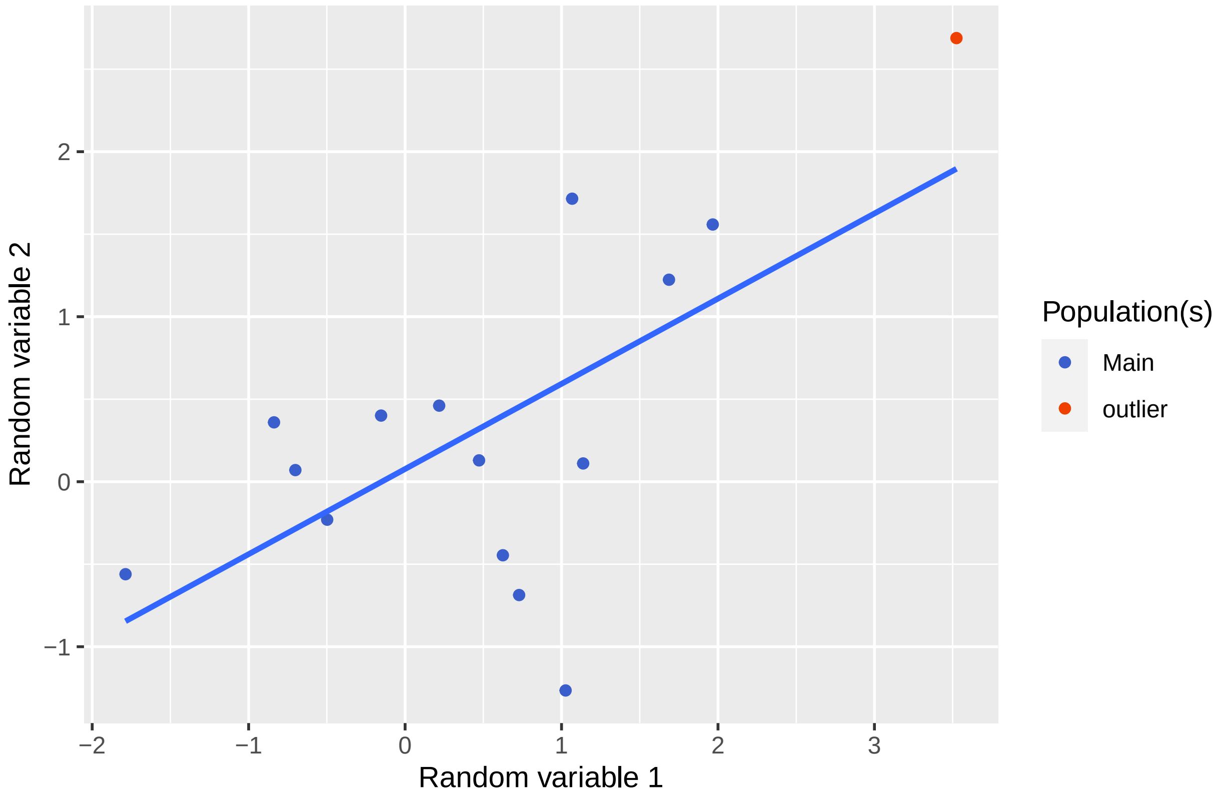


Figure 14: Figure 14: example of a correlation biased by an outlier (in red) when the simulated R is 0 , the sample size is 15 and the outlier is 3 standard deviation away from the mean

Despite the absence of correlation between the two underlying variables (r=0), we observe a strong correlation r=0.64 (CI: [0.2, 0.87], p=0.01). However, a single simulation does not tell us how an outlier impact the false-positive rate. Therefore, as we did above, we generated many such samples and tested whether the presence of an outlier increases the chance of getting a significant correlation when the two underlying variables are not correlated (r=0).

How sensitive are correlations to a single outlier?

As we did for highlighting the behavior of p-values, we repeatedly generated random numbers to obtain samples of the x and y variables. In addition, we replace the last point with an outlier as we did above. Importantly, given that these two variables are randomly generated, the expected correlation is r=0. Therefore, we expect a uniform distribution of p-values (as in Figure 2). In addition, only 5% of the simulated correlations should be significant (false-positive rate). Figure 15 illustrates that both of these expectations are falsified by the simulations. Therefore, a single outlier in a population of 15 points increase the type I error tremendously. That is, there are 25 % significant p-values (right panel of Figure 15)!

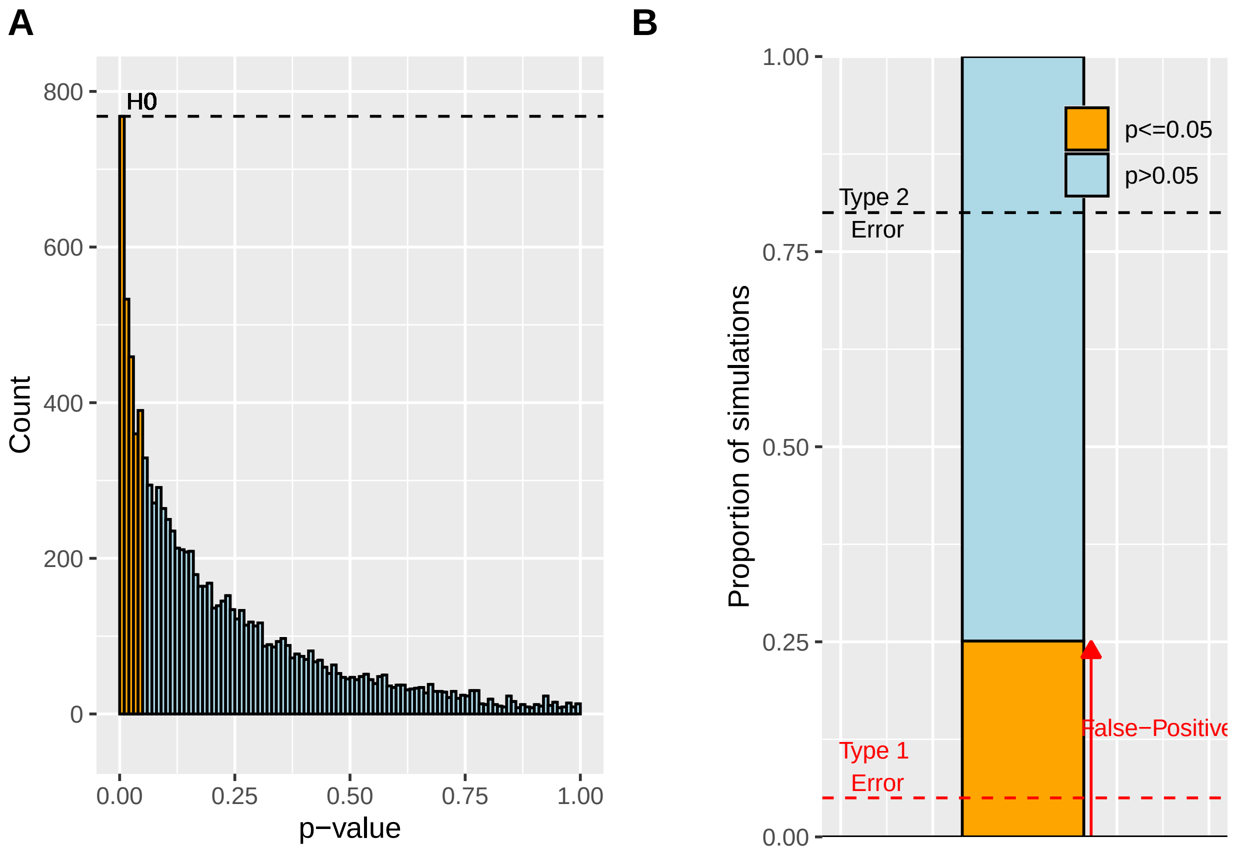


Figure 15: Figure 15: Distribution of p-values (panel A) and proportion of significant and non-significant p-values (panel B) when the simulated correlations are biased by an outlier 3 standard deviation away from the mean. True r= 0 and the sample size is 15

The interesting reader can use the Shiny App or the R code to explore the influence of population size and the distance of the outlier from the mean on the increased number of false-positive correlations caused by the outlier. Off course, this is also reflected in the value of the average correlation coefficient.

Assignment: *These analyses have been performed on the basis of r=0. The interested reader can explore the influence of such outlier on simulated correlations with R>0.*

Testing solutions

Similarly to (Pernet, Wilcox, and Rousselet 2013), we will use simulations to test whether using other types of correlations than Pearson's correlation can mitigate the influence of outliers on the false-positive error rate. Such a solution will be considered if, in the absence of correlation between the underlying variables, only 5% of simulated correlations are false-positive. Furthermore, the average correlation across simulations should be closed to the expected one.

Solution: non-parametric correlation - Spearman A first solution might be to use a non-parametric rank correlation such as Spearman (non-parametric version of the Pearson correlation). Figure 16 shows that the influence of the outlier is much reduced but that the false-positive error rate (= 8%) is still not close to the 5% level.

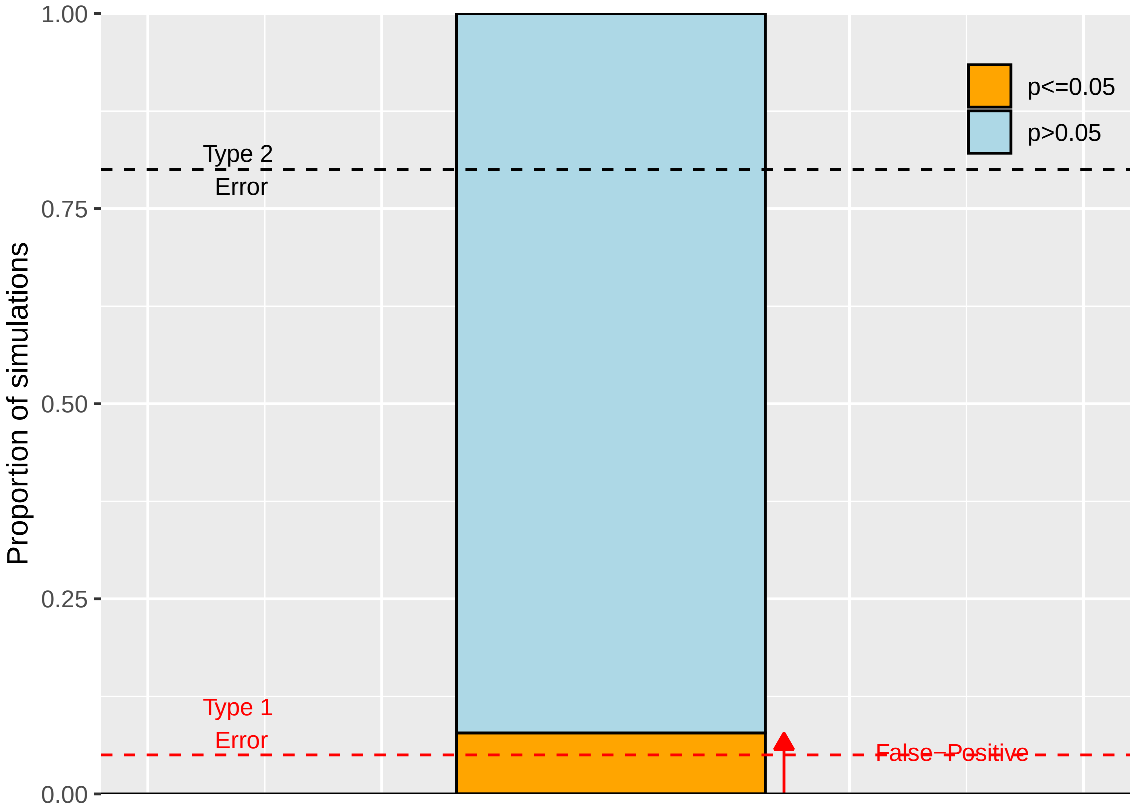


Figure 16: Figure 16: Ability of Spearman correlation to reduce the influence of an outlier on type I error. Proportion of significant and non-significant p-values for Spearman correlation when the simulated correlations are biased by an outlier 3 standard deviation away from the mean. True r= 0 and the sample size is 15

Solution: robust parametric correlation In contrast, the use of winsorized correlation (Pernet, Wilcox, and Rousselet 2013; Wilcox and Rousselet, n.d.) solves the problem as robust correlation methods are designed to handle outliers. Winsorization involves replacing extreme values (beyond a given percentile) with the value of that percentile. For instance, for a 80% winsorized correlation, all values beyond the 90th (resp. below the 10th) percentile are assigned the value of the 90th (resp. the 10th) percentile. The efficacy of winsorized correlation is confirmed by Figure 17, which shows that the influence of the outlier is eliminated and that the type I error rate (= 0) is very close to the 5% level. In R, this correlation is available in the $WRS2$ package (Mair and Wilcox 2020).

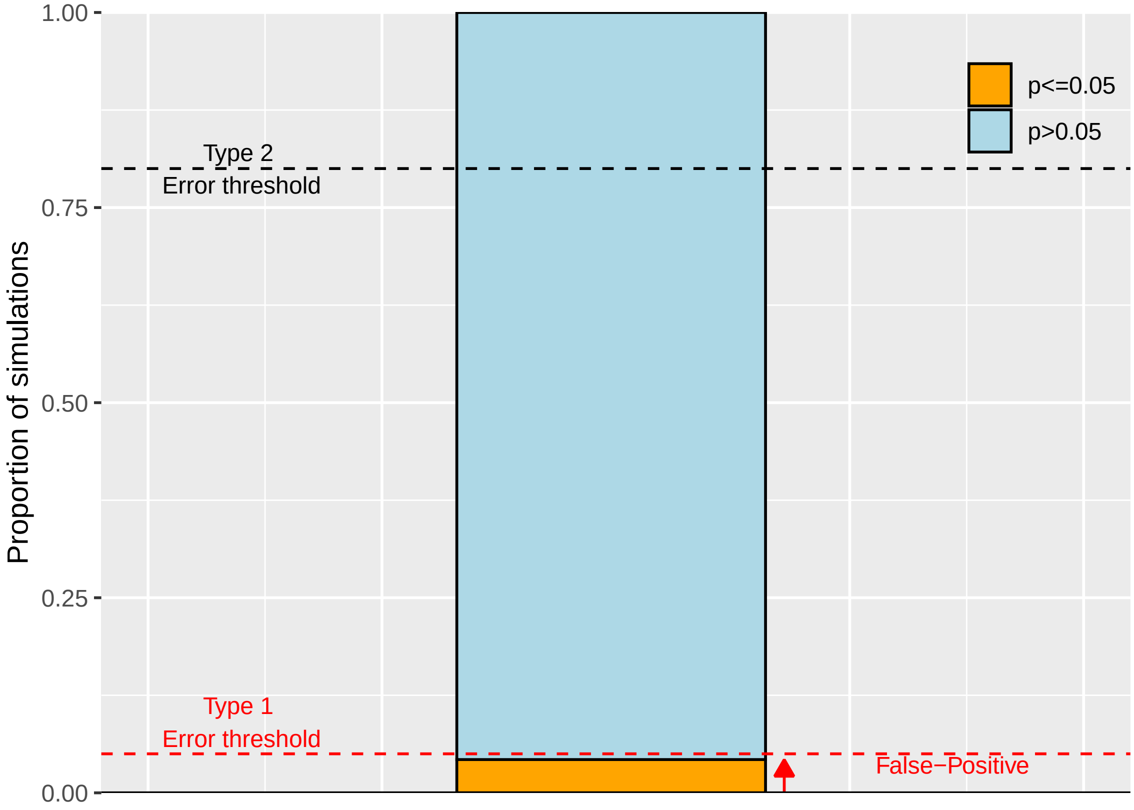


Figure 17: Figure 17: Ability of robust correlation (winsorized) to reduce the influence of an outlier on type I error. Proportion of significant and non-significant p-values for robust correlation when the simulated correlations are biased by an outlier 3 standard deviation away from the mean. True r= 0 and the sample size is 15

The subgroup problem: simulation and solution

Researchers often want to estimate correlation across different subgroups. For instance, in my own research, I want to correlate the amount of explicit motor learning with working memory capacity across both young and older participants (Vandevoorde and Orban de Xivry 2019). Yet, working memory capacity is known to decline with aging and so is the explicit component of motor learning (Vandevoorde and Orban de Xivry 2019). This does not mean that these two variables are correlated across participants. The goal of this section is to investigate the influence of subgroups on correlations because many researchers compute correlations on the pooled data. Yet correlations on heterogeneous subgroups are problematic (Hadzi-Pavlovic 2007; Hassler and Thadewald 2003; Sockloff 1975).

How sensitive are correlations to subgroups?

To generate subgroups, we first generate the x and y parameters for the 15 individuals from two uncorrelated variables (r=0). We will then split the population into two subgroups and add a value of 2 to the values of both parameters for one of the two subgroups. This produces a scatterplot such as the one presented on Figure 18.

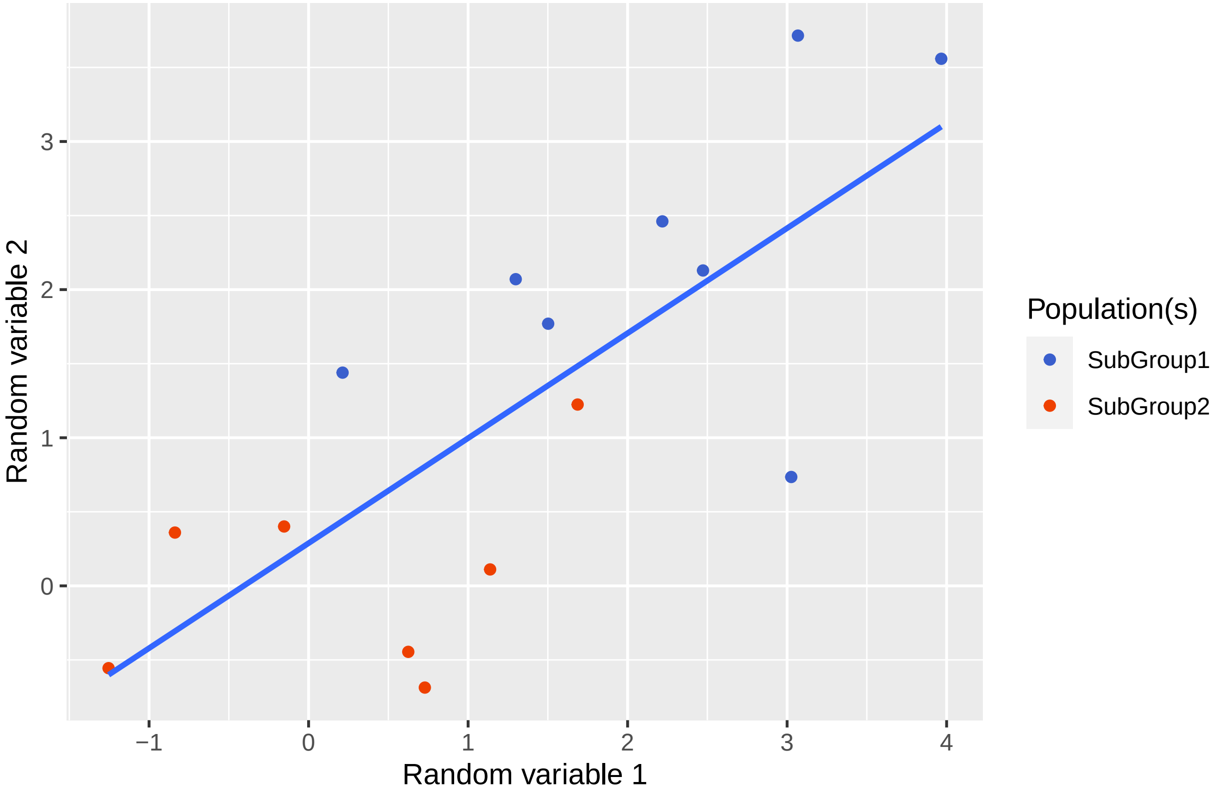


Figure 18: Figure 18: Example of correlation biased by differences between subgroups 2 standard deviation away from each other. True r= 0 and the sample size is 15

In this graph, we observe a strong correlation r=0.75 (CI: [0.38, 0.91], p=0.001). Repeating such simulation (10000 times) allows us to understand the effect of subgroups on the number of false-positive correlations. Left panel of Figure 19 illustrates that the presence of differences between the subgroups have a massive effect on the distribution of p-values for the simulated correlations. Indeed, given that we simulate two variables that are not correlated, we should expect a uniform distribution of p-values (see Figure 2)). If the two subgroups are 2 SD away, we observe 55 % of significant p-values (right panel of Figure 19), which is far above the expected 5% threshold.

Assignment: *The interested reader can test the effect of the distance between the two groups (in SD) on the percentage of false-positive correlations. Sample size also influences this percentage*

Note that for 1.8% of these significant correlations, the between-groups difference for the x and y parameters was not significant. Yet, these subgroups yielded a positive correlation. This means that the absence of significant difference across subgroups for the two parameters is not sufficient to justify the fact that the data was pooled across groups. The subgroups must be taken into account when computing the correlation.

In addition, the bigger the groups are, the smaller the difference between the subgroups must be to give rise to a significant correlation. For instance, for a sample of 15 participants, a difference of 0.5 SD between

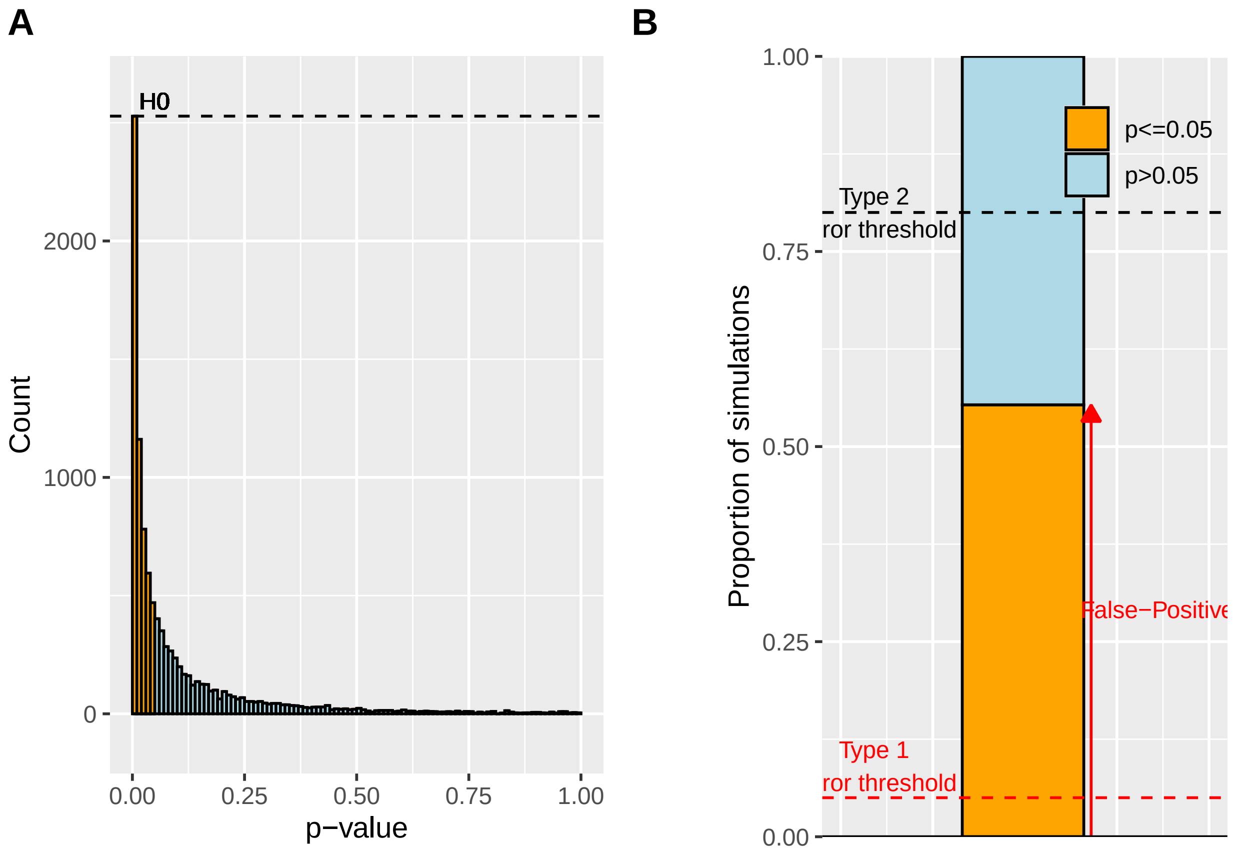


Figure 19: Figure 19: Influence of subgroup differences on the distribution of p-values (panel A) and on the proportion of significant and non-significant p-values (panel B) when the simulated correlations are biased by subgroups 2 standard deviation away from each other. True r= 0 and sample size N= 15

groups merely inflates the number of false-positive correlations (6% of correlations are significant). However, this difference becomes important for larger populations. For instance, when the total sample is N = 75, a difference of 0.5 SD between subgroups increase the type I error rate to 8.1% and this percentage increases further with sample size.

> ***Assignment:*** *How does the mean difference between the subgroups and their size affect the magnitude of the percentage of false-positive correlations.*

Solutions to the subgroup problem

We will first show that Spearman correlation is again unable to solve the problem of subgroups even though it slightly reduces the problem. Then we will show that one needs to take the subgroups into account in a regression model or via multilevel correlations in order to bring the percentage of false-positive correlations back to 5% (which is expected given that the two underlying variables are uncorrelated, r=0). Interestingly, robust regression could also be used in order to deal with both subgroups and outliers. However, the regression solution becomes tricky when there are more than two groups. An alternative solution, which works easily for many groups, is the use of multilevel correlation (Makowski et al. 2022).

Spearman is not a solution A first solution might be to use a non-parametric rank correlation such as Spearman (non-parametric version of the Pearson correlation). Figure 20 shows that the influence of subgroups is not eliminated as the false-positive error rate (58%) remains larger than the expected 5% level when the underlying variables are not correlated.

Coding subgroups as a discrete factor in a multiple regression reduces the false positive rate. To correlate parameters across different subgroups, one has to take these subgroups into account. This can be done by means of regression. In a simple linear regression $y = a * x + b$, the coefficient a corresponds to the correlation coefficient between x and y if these variables are z-normalized (mean=0 and standard deviation of 1). To z-normalize a variable, one needs to subtract the sample mean from every observation and divide the outcome by the standard deviation of the sample ($z_i = (x_i - mean(x))/SD(x)$). Note that when the data are normalized, b is always zero (in $y = a * x + b$). To take subgroups into account, one needs to modify the regression equation as follow:

$$y = a * x + c * G$$

where G is equal to -1 for one subgroup and to 1 for the other (the sum of these coefficients needs to be zero). This equation makes the hypothesis that we expect the same correlation for both groups given that there are no interaction term between x and G. As shown on Figure 21, including an additional independent variable (G) in the regression is able to reduce the false-positive rate to the 5% level and could therefore be used when data from heterogeneous subgroups are pooled for correlations.

In the absence of subgroups, false-positive correlations had the more extreme values in the distribution of correlations coefficient (see Figure 4). Yet, this is no more the case when we correct for subgroups (Figure 22). Indeed, correcting for subgroups renders some extreme correlations not significant and less extreme correlations significant. Therefore, the correction for the existence of subgroups has an effect on the value of the false-positive correlations.

Taking subgroups into account does not prevent us from detecting true correlations. Indeed, one can apply the same logic when a correlation of 0.5 is used for the simulations. In this case, we see that the average correlation over all simulations (r=0.5) is pretty close to the actual correlation value (r=0.5). This further validates the use of this method to compute correlation in the presence of subgroups. Off course, it remains also true that considering only the significant correlations remains problematic and leads to an overestimation of the actual correlation value (orange dashed line on Figure 23). It is important to note that the transition from significant to non-significant correlations presented on Figure 23 is again more blurred than the ones shown on both panels of Figure 10.

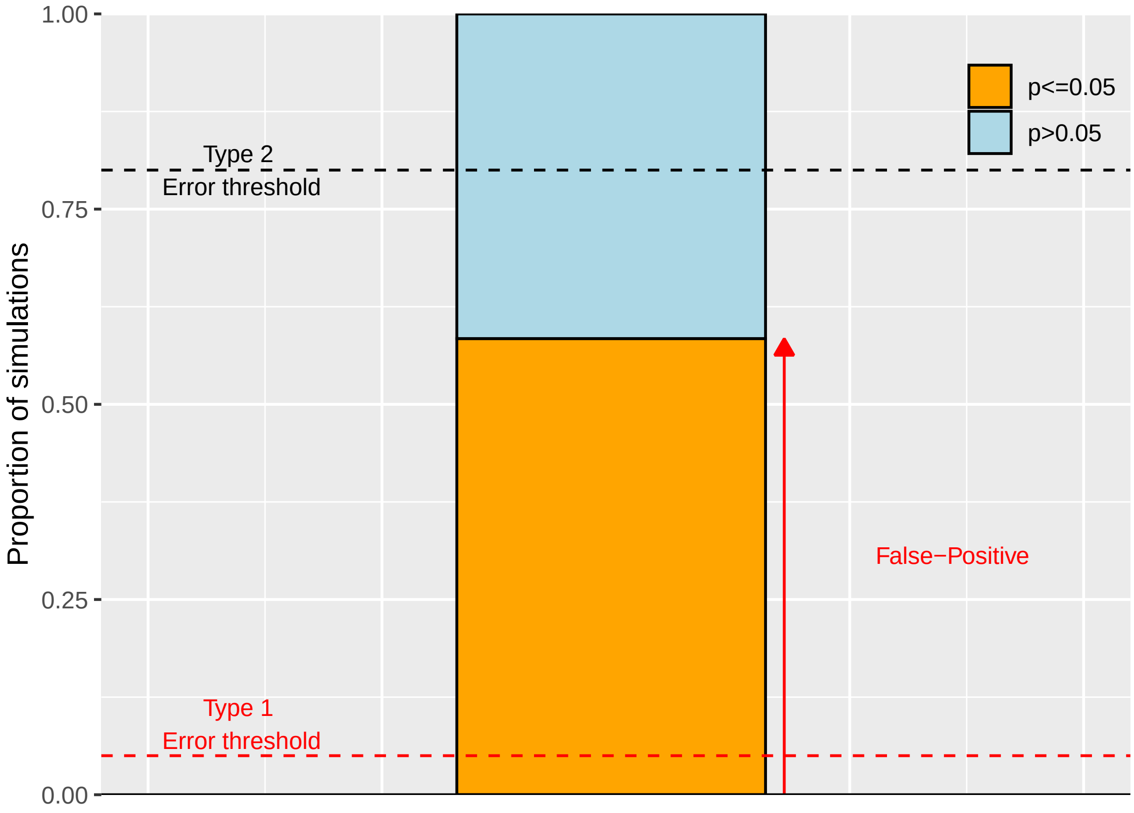


Figure 20: Figure 20: Inability of Spearman correlation to reduce the influence of subgroups on type I error. Proportion of significant and non-significant p-values for Spearman correlation when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True r= 0 and the sample size is 15

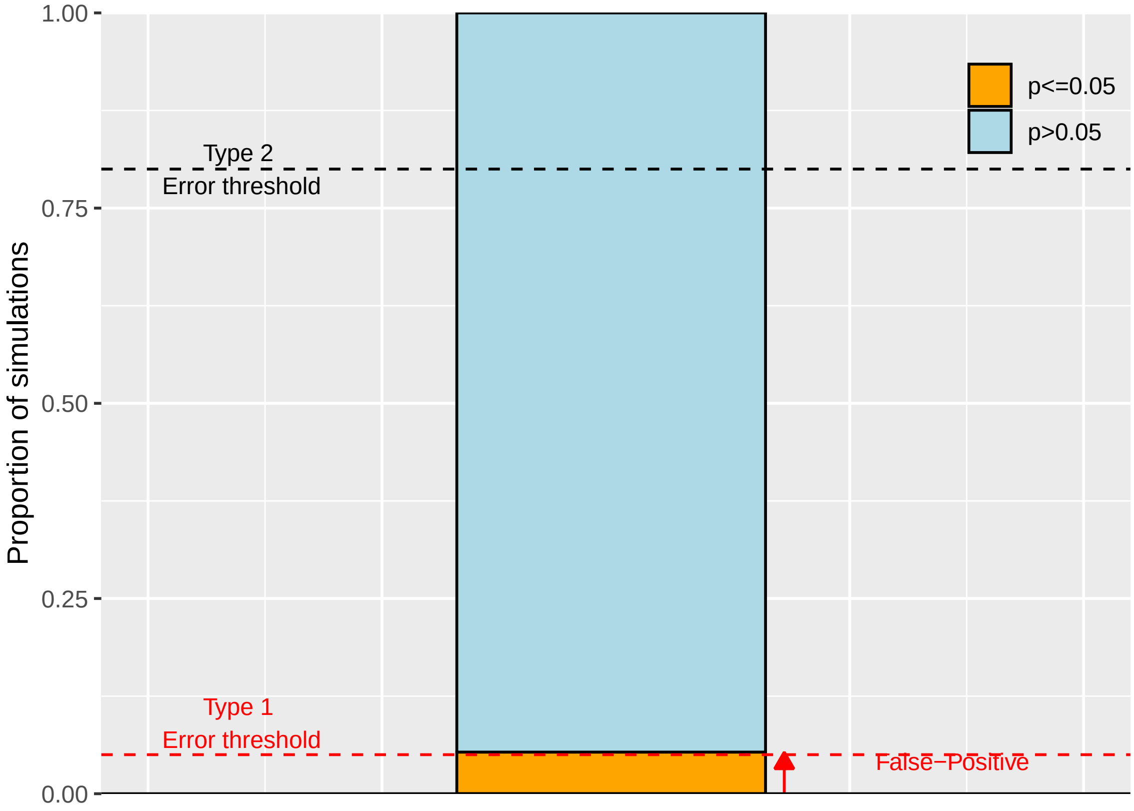


Figure 21: Figure 21: Ability of regression to reduce the influence of subgroups on type I error. Proportion of significant and non-significant p-values for standardized regression coefficient when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True r= 0 and the sample size is 15

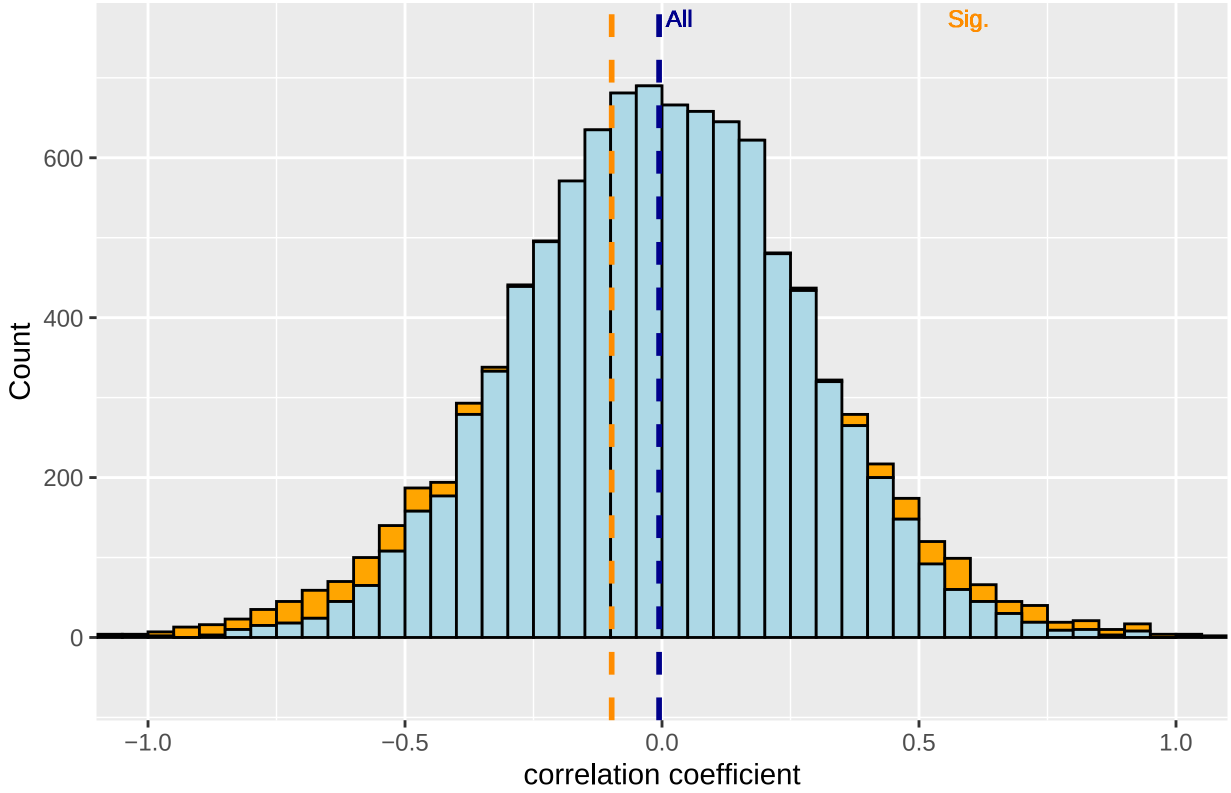


Figure 22: Figure 22: distribution of significant and non-significant standardized regression coefficients when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R is 0 and the sample size is 15

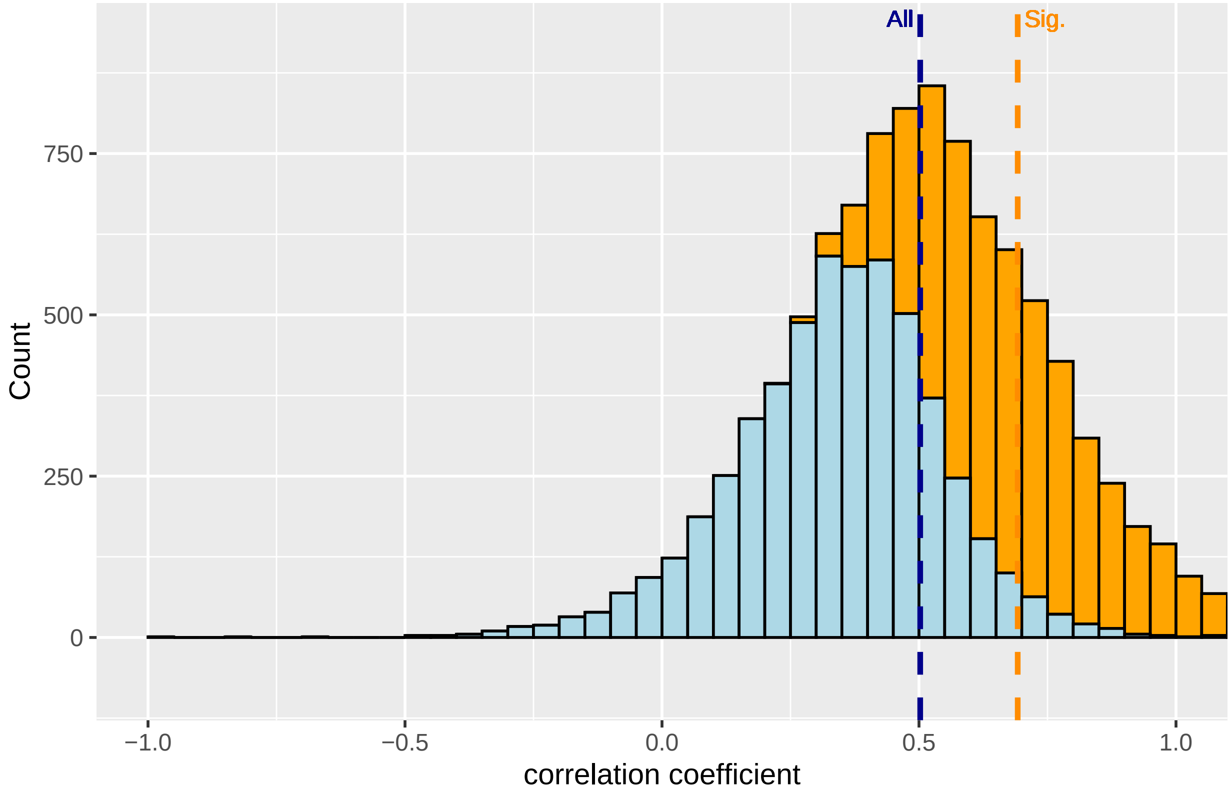


Figure 23: Figure 23: distribution of significant and non-significant standardized regression coefficients when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R is 0.5 and the sample size is 15

Multilevel correlations can take many subgroups into account and are controlling the false-positive rate. Multilevel correlations (also referred to as hierarchical or random-effects correlation) are a type of linear models that can compute a correlation for pooled data across groups while taking the different subgroups into account (Makowski et al. 2020). This is important as illustrated by the Simpson paradox (see link for details, (Blyth 1972)). As we have done for the regression, we will simulate two independent variables for two different subgroups and we will then apply the multilevel correlation to obtain the correlation magnitude and the associated p-value. We expect that 5% of the p-values will be under the p=0.05 threshold as no underlying correlation is used.

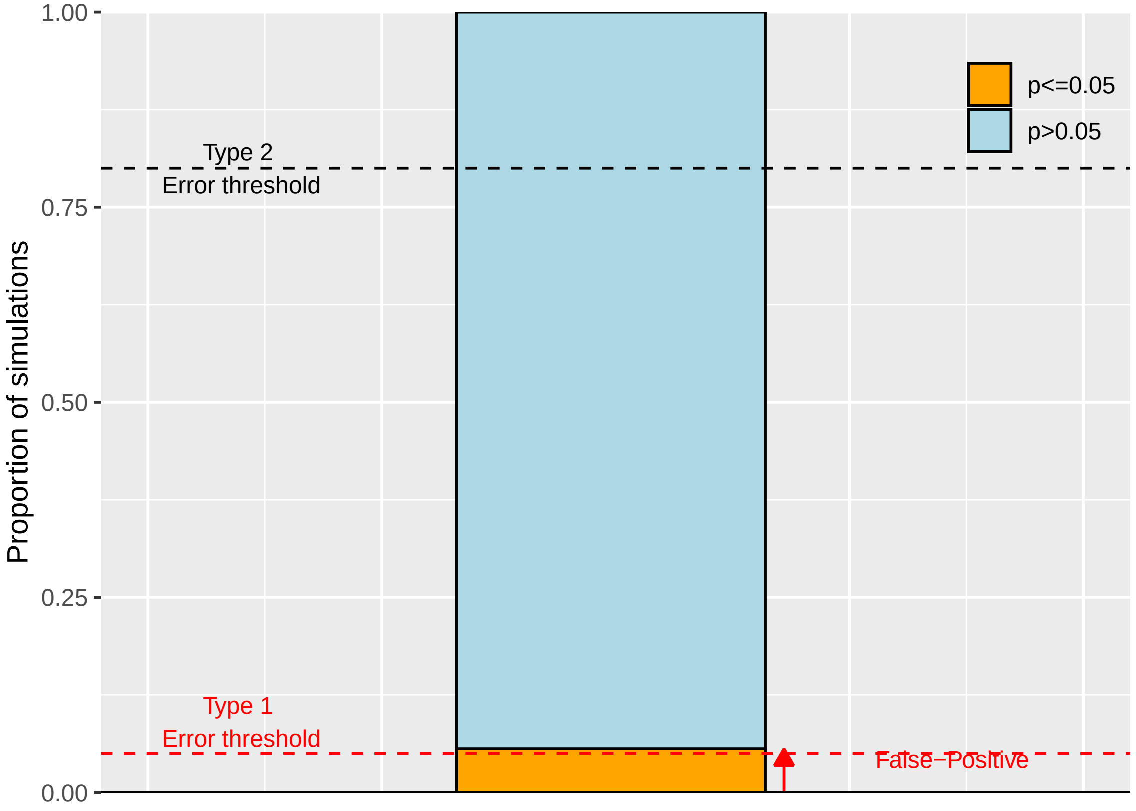


Figure 24: Figure 24: Ability of multilevel correlation to reduce the influence of subgroups on type I error. Proportion of significant and non-significant p-values for multilevel correlation when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True r= 0 and the sample size is 15

As can be seen on Figure 24, there is a clear drop in false-positive rate compared to Figure 19. Yet, this is not perfect as the false-positive rate is still slightly higher than what would be expected (5.59% compared to the expected 5%).

The pattern of correlation presented on Figure 25 is much different than that obtained for the regression one (Figure 22). In this case, the significant correlations are the most extreme ones, as was found in the absence of outliers or subgroups (Figure 4). This is also true for the values of the correlations when correlations are simulated with r=0.5 (Figure 26 compared to Figure 10.A).

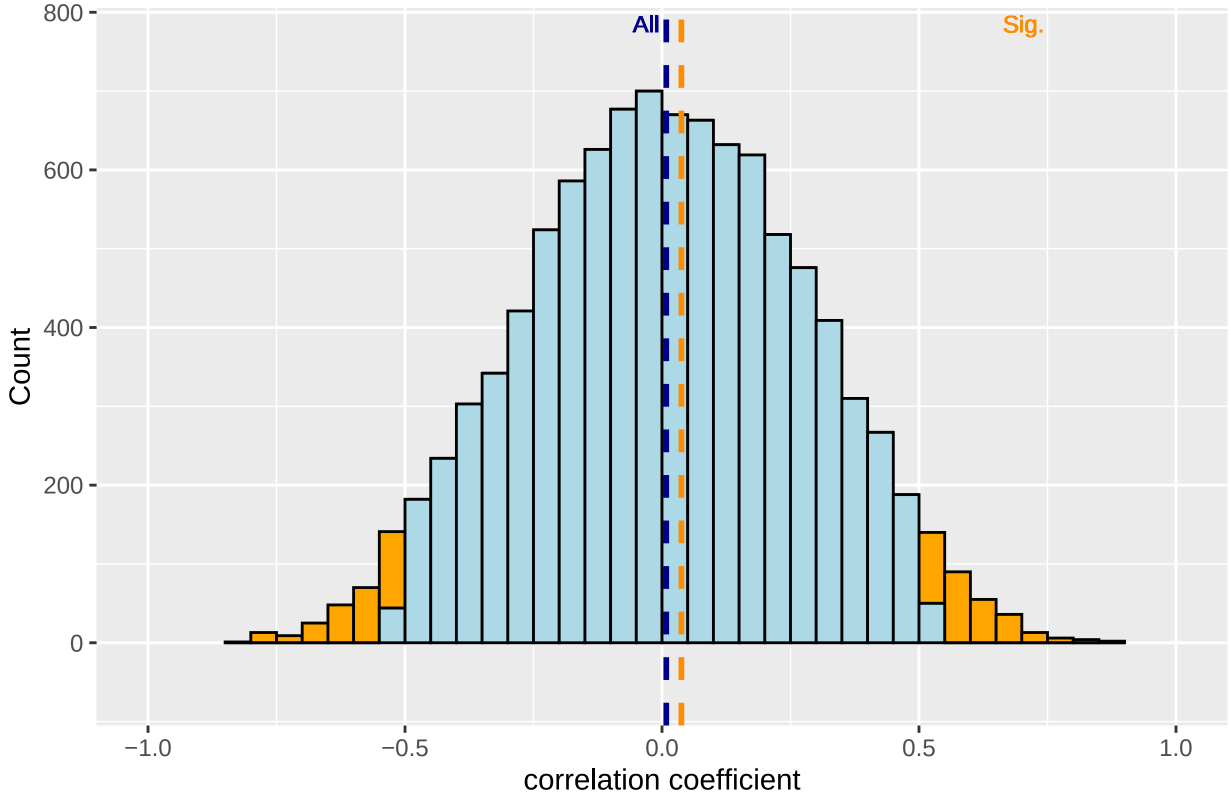


Figure 25: Figure 25: distribution of significant and non-significant multilevel correlations when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R is 0 and the sample size is 15

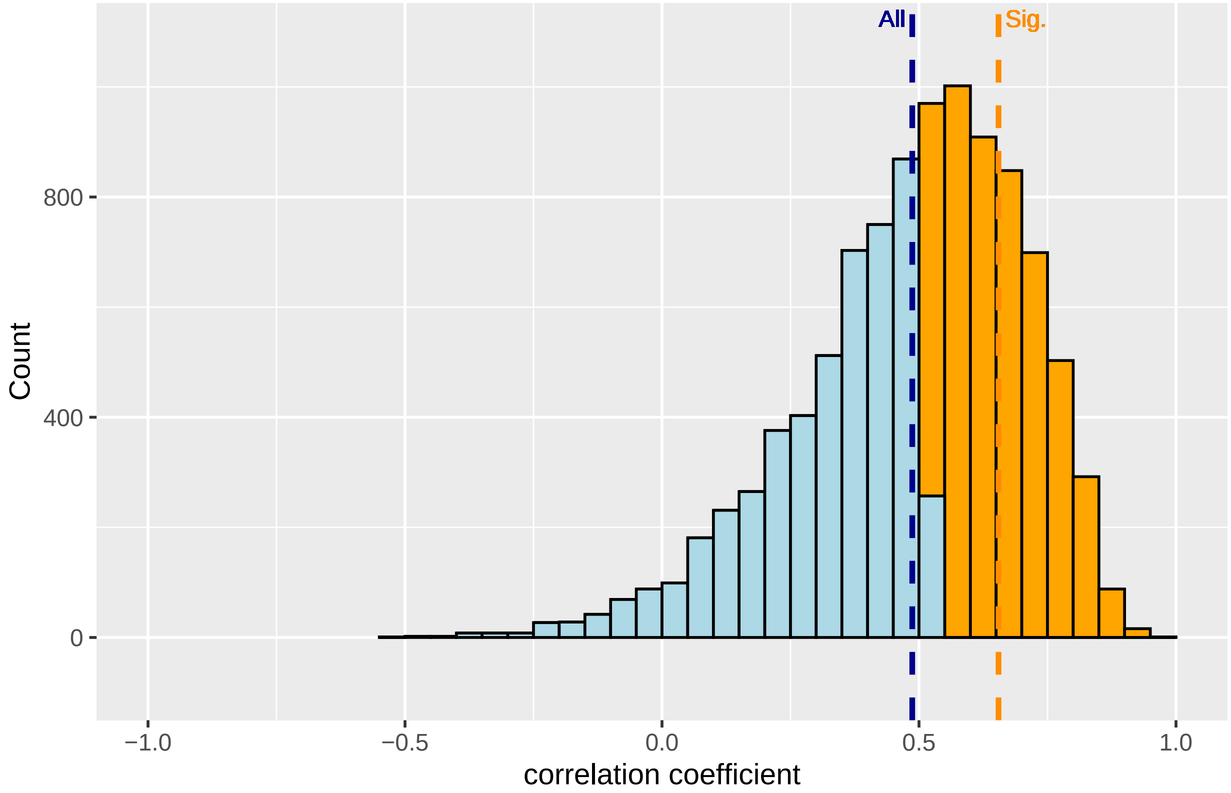


Figure 26: Figure 26: distribution of significant and non-significant multilevel correlation coefficients when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R is 0.5 and the sample size is 15

What did we learn?

Simulations stimulate statistical thinking and yield important insights on statistics. It also allows one to explore potential solutions to a data analysis problem (e.g. presence of subgroups).

1) The distribution of p-value in the absence of an effect is flat. The probability of observing p<0.05 is identical to the probability of observing p>0.95 (Benjamini and Hochberg 1995).
2) Even in the presence of an effect, some p-values can be non-significant (Daniël Lakens and Etz 2017).
3) In the presence of an effect, the proportion of significant p-values provide information about power. In the absence of an effect, it represents the number of false positives.
4) Publication bias is bad for effect size estimation (Nissen et al. 2016; Sutton et al. 2000). Both significant and non-significant effects are critical to estimate the actual effect size.
5) When power is low, the overestimation of effect size is large. This overestimation is reinforced by publication bias.
6) When power i slow, the width of the confidence interval of the individual correlation coefficient is large, which means that the uncertainty of around the actual correlation coefficient is big (Button, Ioannidis, et al. 2013b). Moreover, the effect of irregularities in the data (outliers or subgroups) is more prominent in situations where power is low.
7) Increasing the sample size has three major effects: 1) it increases power, 2) it improves accuracy of the effect size estimation at the population level and 3) it increases the accuracy of every individual correlation coefficient (width of confidence interval decreases).
8) Getting significant p-values is not difficult, one just needs to try enough different statistical tests (Bennett et al. 2011).
9) Having a mix of significant and non-significant results is normal and should become the norm. Some scientists present results that are too good to be true (Francis 2014). Anybody should be wary of scientists who only publish significant results with small sample size (Schimmack 2012).
10) Simulations can be useful to test potential solution to encountered problems. For instance, the simulations clearly demonstrated that winsorised but not Spearman correlations can handle outliers (Pernet, Wilcox, and Rousselet 2013; Rousselet and Pernet 2012) and that both multiple regression and multilevel correlations could be used when correlating two variables from data pooled from different subgroups.

This paper represents an introduction to the use of simulations to understand the concept of p-values, power and to test solutions to irregularities in the data. Yet, more advanced use of simulations is also possible. For instance, simulations can be used to demonstrate how you should decide which covariates to adjust for in the context of a randomized controlled trial (https://twitter.com/statsepi/status/1115902270888128514).

References

Abdullah, Mokhtar Bin. 1990. "On a Robust Correlation Coefficient." *Journal of the Royal Statistical Society: Series D (The Statistician)* 39 (4): 455–60. https://doi.org/10.2307/2349088.

Aggarwal, Rakesh, and Priya Ranganathan. 2016. "Common Pitfalls in Statistical Analysis: The Use of Correlation Techniques." *Perspectives in Clinical Research* 7 (4): 187–90. https://doi.org/10.4103/2229-3485.192046.

Altman, D. G., and J. M. Bland. 1995. "Absence of evidence is not evidence of absence." *BMJ (Clinical research ed.)* 311 (7003): 485. https://doi.org/10.1136/bmj.311.7003.485.

Amrhein, Valentin, and Sander Greenland. 2018. "Remove, Rather Than Redefine, Statistical Significance." *Nature Human Behaviour* 2 (1): 4. https://doi.org/10.1038/s41562-017-0224-0.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6. https://doi.org/10.1038/s41562-017-0189-z.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Bennett, CM, AA Baird, Michael B Miller, and George L Wolford. 2011. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument for Proper Multiple Comparisons Correction." *Journal of Serendipitous \Ldots* 1 (1): 1–5.

Bishop, D. V. M., Jackie Thompson, and Adam J. Parker. 2022. "Can We Shift Belief in the 'Law of Small Numbers'?" *Royal Society Open Science* 9 (3): 211028. https://doi.org/10.1098/rsos.211028.

Bishop, Dorothy. 2020. "How Scientists Can Stop Fooling Themselves over Statistics." *Nature* 584 (7819): 9–9. https://doi.org/10.1038/d41586-020-02275-8.

Bishop, Dorothy V. M. 2023. "Using Multiple Outcomes in Intervention Studies: Improving Power While Controlling Type I Errors." https://doi.org/10.12688/f1000research.73520.2.

Blyth, Colin R. 1972. "On Simpson's Paradox and the Sure-Thing Principle." *Journal of the American Statistical Association* 67 (338): 364–66. https://doi.org/10.2307/2284382.

Büttner, Fionn, Elaine Toomey, Shane McClean, Mark Roe, and Eamonn Delahunt. 2020. "Are Questionable Research Practices Facilitating New Discoveries in Sport and Exercise Medicine? The Proportion of Supported Hypotheses Is Implausibly High." *British Journal of Sports Medicine* 54 (22): 1365–71. https://doi.org/10.1136/bjsports-2019-101863.

Button, Katherine S., John P. a. Ioannidis, Claire Mokrysz, Brian a. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013b. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews. Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Button, Katherine S., John P. a. Ioannidis, Claire Mokrysz, Brian a. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013a. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews. Neuroscience* 14 (5): 36576. https://doi.org/10.1038/nrn3475.

Calin-Jageman, Robert J., and Geoff Cumming. 2019. "The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known." *The American Statistician* 73 (sup1): 271–80. https://doi.org/10.1080/00031305.2018.1518266.

Carter, Evan C., Felix D. Schönbrodt, Will M. Gervais, and Joseph Hilgard. 2019. "Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods." *Advances in Methods and Practices in Psychological Science* 2 (2): 115–44. https://doi.org/10.1177/2515245919847196.

Colquhoun, David, and College London. 2014. "An Investigation of the False Discovery Rate and the Misinterpretation of P Values." *Hazards of P Values*, 1–15. https://doi.org/10.1098/rsos.140216.

Corneille, Olivier, Jo Havemann, Emma L Henderson, Hans IJzerman, Ian Hussey, Jean-Jacques Orban de Xivry, Lee Jussim, et al. 2023. "Beware Persuasive Communication Devices When Writing and Reading Scientific Articles." *eLife* 12 (May): e88654. https://doi.org/10.7554/eLife.88654.

Cumming, Geoff. 2011. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge. https://doi.org/10.4324/9780203807002.

Francis, Gregory. 2013. "Replication, Statistical Consistency, and Publication Bias." *Journal of Mathematical Psychology*, Special Issue: A Discussion of Publication Bias and the Test for Excess Significance, 57 (5): 153–69. https://doi.org/10.1016/j.jmp.2013.02.003.

———. 2014. "The Frequency of Excess Success for Articles in Psychological Science." *Psychonomic Bulletin & Review*, 1–26.

Gagnier, Joel J., and Hal Morgenstern. 2017. "Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing." *The Journal of Bone and Joint Surgery* 99 (18): 1598–1603. https://doi.org/10.2106/JBJS.16.01314.

Gardner, M. J., and D. G. Altman. 1986. "Confidence Intervals Rather Than P Values: Estimation Rather Than Hypothesis Testing." *Br Med J (Clin Res Ed)* 292 (6522): 746–50. https://doi.org/10.1136/bmj.

292.6522.746.

Gigerenzer, Gerd. 2018. "Statistical Rituals: The Replication Delusion and How We Got There." *Advances in Methods and Practices in Psychological Science* 1 (2): 198–218. https://doi.org/10.1177/2515245918771329.

Goodman, Steven N. 1999. "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy." *Annals of Internal Medicine* 130 (12): 995. https://doi.org/10.7326/0003-4819-130-12-199906150-00008.

Gopalakrishna, Gowri, Gerben ter Riet, Gerko Vink, Ineke Stoop, Jelte M. Wicherts, and Lex M. Bouter. 2022. "Prevalence of Questionable Research Practices, Research Misconduct and Their Potential Explanatory Factors: A Survey Among Academic Researchers in The Netherlands." *PLOS ONE* 17 (2): e0263023. https://doi.org/10.1371/journal.pone.0263023.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31 (4): 337–50. https://doi.org/10.1007/s10654-016-0149-3.

Hadzi-Pavlovic, Dusan. 2007. "Correlations III: Heterogeneous Data." *Acta Neuropsychiatrica* 19 (3): 215–16. https://doi.org/10.1111/j.1601-5215.2007.00219.x.

Hassler, Uwe, and Thorsten Thadewald. 2003. "Nonsensical and Biased Correlation Due to Pooling Heterogeneous Samples." *Journal of the Royal Statistical Society. Series D (The Statistician)* 52 (3): 367–79.

Havlicek, Larry L., and Nancy L. Peterson. 1976. "Robustness of the Pearson Correlation Against Violations of Assumptions." *Perceptual and Motor Skills* 43 (3_suppl): 1319–34. https://doi.org/10.2466/pms.1976.43.3f.1319.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13 (3): e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Hubbard, Raymond, and R. Murray Lindsay. 2008. "Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing." *Theory & Psychology* 18 (1): 69–88. https://doi.org/10.1177/0959354307086923.

Joober, Ridha, Norbert Schmitz, Lawrence Annable, and Patricia Boksa. 2012. "Publication Bias: What Are the Challenges and Can They Be Overcome?" *Journal of Psychiatry & Neuroscience : JPN* 37 (3): 149–52. https://doi.org/10.1503/jpn.120065.

Lakens, Daniel, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, et al. 2018. "Justify Your Alpha." *Nature Human Behaviour* 2 (3): 168–71. https://doi.org/10.1038/s41562-018-0311-x.

Lakens, Daniël, and Alexander J. Etz. 2017. "Too True to Be Bad: When Sets of Studies With Significant and Nonsignificant Findings Are Probably True." *Social Psychological and Personality Science* 8 (8): 875–81. https://doi.org/10.1177/1948550617693058.

Lane, Anthony, Olivier Luminet, Gideon Nave, and Moïra Mikolajczak. 2016. "Is There a Publication Bias in Behavioral Intranasal Oxytocin Research on Humans? Opening the File Drawer of One Lab." *Journal of Neuroendocrinology*, no. 16: n/a–. https://doi.org/10.1111/jne.12384.

Lee Rodgers, Joseph, and W. Alan Nicewander. 1988. "Thirteen Ways to Look at the Correlation Coefficient." *The American Statistician* 42 (1): 59–66. https://doi.org/10.1080/00031305.1988.10475524.

Lyu, Xiao-Kang, Yuepei Xu, Xiao-Fan Zhao, Xi-Nian Zuo, and Chuan-Peng Hu. 2020. "Beyond Psychology: Prevalence of p Value and Confidence Interval Misinterpretation Across Different Fields." *Journal of Pacific Rim Psychology* 14. https://doi.org/10.1017/prp.2019.28.

Mair, Patrick, and Rand Wilcox. 2020. "{Robust Statistical Methods in r Using the WRS2 Package}" 52.

Makin, Tamar R, and Jean-Jacques Orban de Xivry. 2019. "Ten Common Statistical Mistakes to Watch Out for When Writing or Reviewing a Manuscript." Edited by Peter Rodgers, Nick Parsons, and Nick Holmes. *eLife* 8 (October): e48175. https://doi.org/10.7554/eLife.48175.

Makowski, Dominique, Mattan S. Ben-Shachar, Indrajeet Patil, and Daniel Lüdecke. 2020. "Methods and Algorithms for Correlation Analysis in R." *Journal of Open Source Software* 5 (51): 2306. https://doi.org/10.21105/joss.02306.

Makowski, Dominique, Brenton M. Wiernik, Indrajeet Patil, Daniel Lüdecke, and Mattan S. Ben-Shachar. 2022. "{{Correlation}}: Methods for Correlation Analysis." https://CRAN.R-project.org/package=correlation.

Martins, Rui Manuel da Costa. 2018. "Learning the Principles of Simulation Using the Birthday Problem." *Teaching Statistics* 40 (3): 108–11. https://doi.org/10.1111/test.12164.

Mlinarić, Ana, Martina Horvat, and Vesna Šupak Smolčić. 2017. "Dealing with the Positive Publication Bias: Why You Should Really Publish Your Negative Results." *Biochemia Medica* 27 (3). https://doi.org/10.11613/BM.2017.030201.

Nissen, Silas Boye, Tali Magidson, Kevin Gross, and Carl T Bergstrom. 2016. "Publication Bias and the Canonization of False Facts." Edited by Peter Rodgers. *eLife* 5 (December): e21451. https://doi.org/10.7554/eLife.21451.

O'Hara, Michael. 2019. "Teaching Hypothesis Testing with Simulated Distributions." *International Review of Economics Education* 30 (January): 100138. https://doi.org/10.1016/j.iree.2018.05.005.

Orban de Xivry, Jean-Jacques, Marcus Missal, and Philippe Lefèvre. 2009. "Smooth Pursuit Performance During Target Blanking Does Not Influence the Triggering of Predictive Saccades." *Journal of Vision* 9 (11): 7.1–16. https://doi.org/10.1167/9.11.7.

Pernet, Cyril R., Rand Wilcox, and Guillaume A. Rousselet. 2013. "Robust Correlation Analyses: False Positive and Power Validation Using a New Open Source Matlab Toolbox." *Frontiers in Psychology* 3 (JAN): 1–18. https://doi.org/10.3389/fpsyg.2012.00606.

Ravn, Tine, and Mads P. Sørensen. 2021. "Exploring the Gray Area: Similarities and Differences in Questionable Research Practices (QRPs) Across Main Areas of Research." *Science and Engineering Ethics* 27 (4): 40. https://doi.org/10.1007/s11948-021-00310-z.

Rousselet, Guillaume A., and Cyril R. Pernet. 2012. "Improving Standards in Brain-Behavior Correlation Analyses." *Frontiers in Human Neuroscience* 6. https://doi.org/10.3389/fnhum.2012.00119.

Schimmack, Ulrich. 2012. "The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles." *Psychological Methods* 17 (4): 551–66. https://doi.org/10.1037/a0029487.

Sockloff, Alan L. 1975. "Behavior of the Product-Moment Correlation Coefficient When Two Heterogeneous Subgroups Are Pooled1." *Educational and Psychological Measurement* 35 (2): 267–76. https://doi.org/10.1177/001316447503500204.

Song, Fujian, Lee Hooper, and Yoon K. Loke. 2013. "Publication Bias: What Is It? How Do We Measure It? How Do We Avoid It?" *Open Access Journal of Clinical Trials*. https://www.dovepress.com/publication-bias-what-is-it-how-do-we-measure-it-how-do-we-avoid-it-peer-reviewed-article-OAJCT. https://doi.org/10.2147/OAJCT.S34419.

Sullivan, Gail M., and Richard Feinn. 2012. "Using Effect Size—or Why the P Value Is Not Enough." *Journal of Graduate Medical Education* 4 (3): 279–82. https://doi.org/10.4300/JGME-D-12-00156.1.

Sutton, A J, S J Duval, R L Tweedie, K R Abrams, and D R Jones. 2000. "Empirical Assessment of Effect of Publication Bias on Meta-Analyses." *BMJ : British Medical Journal* 320 (7249): 1574–77.

Szucs, Denes, and John P. A. Ioannidis. 2017. "When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment." *Frontiers in Human Neuroscience* 11. https://doi.org/10.3389/fnhum.2017.00390.

Taylor, Richard. 1990. "Interpretation of the Correlation Coefficient: A Basic Review." *Journal of Diagnostic Medical Sonography* 6 (1): 35–39. https://doi.org/10.1177/875647939000600106.

Tintle, Nathan, Beth Chance, George Cobb, Soma Roy, Todd Swanson, and Jill VanderStoep. 2015. "Combating Anti-Statistical Thinking Using Simulation-Based Methods Throughout the Undergraduate Curriculum." *The American Statistician* 69 (4): 362–70. https://doi.org/10.1080/00031305.2015.1081619.

Tversky, Amos, and Daniel Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin*, 105–10.

Vandevoorde, Koenraad, and Jean-Jacques Orban de Xivry. 2019. "Motor Adaptation but Not Internal Model Recalibration Declines with Aging." *Neurobiology of Aging*.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. 4th ed. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Vrieze, Jop de. 2021. "Large Survey Finds Questionable Research Practices Are Common." *Science* 373 (6552): 265–65. https://doi.org/10.1126/science.373.6552.265.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Wilcox, Rand R., and Guillaume Rousselet. n.d. "An Updated Guide to Robust Statistical Methods in

Neuroscience." https://doi.org/10.31234/osf.io/kcjfe.

Zimmerman, Donald W. 1994. "A Note on the Influence of Outliers on Parametric and Nonparametric Tests." *The Journal of General Psychology* 121 (4): 391–401. https://doi.org/10.1080/00221309.1994.9921213.