Simulate to stimulate statistical thinking

Simulating correlations to get insights into statistics and their robustness to irregularities in the data results

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 or on social media. Yet, the problem with p-values is that people do not understand them properly or do not understand the circumstances under which a p-value can be used (Colquhoun and London 2014; Gagnier and Morgenstern 2017; Greenland et al. 2016). 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 and many problems that have been discussed in meta-science (the science of science). This part adresses the significance problem linked to p-values.

But p-values on their own have very little value if they are not coupled to the effect size of an effect (Hubbard and Lindsay 2008). Indeed, while we sommetimes 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. 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 estimation problem. 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 do not cherish not significant effect and rarely published them while significant effects are published very easily. We will thus use simulation to understand how publication bias is a threat for the estimation of the magnitude of the effect size.

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 should be expected (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 were based on the correlation between two variables. Correlation was chosen as an example of statistical test because it is very frequently used in scientific papers. 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) and to find the best technique to account for irregularities in the data (Rousselet and Pernet 2012). This paper is partially inspirated by the blogposts of Prof. Dorothy Bishop: [Blog 1](http://deevybee.blogspot.com/2017/11/how-analysis-of-variance-works.html), [Blog 2](https://deevybee.blogspot.com/2017/11/anova-t-tests-and-regression-different.html), [Blog3](https://deevybee.blogspot.com/2017/12/using-simulations-to-understand.html), [Blog 4](http://deevybee.blogspot.com/2017/12/using-simulations-to-understand-p-values.html), [Presentation](https://www.slideshare.net/deevybishop/introduction-to-simulating-data-to-improve-your-research)

# Learning about statistics from simulated correlations.

Let’s imagine the following experiment. Researchers from Klow in Syldavia are willing to measure height and working memory capacity in a population of young healthy participants. Because of their limited budget, they are able to measure these parameters in a population of 15 participants. For each individual participant , they have two observations: where is the height of the participant and is his/her working memory capacity. Then, the researchers decided to test the correlation between these two variables for their population. There is no reason whatsoever to expect a correlation between height and working memory capacity. This should be zero. We can simulate this experiment numerically by generating 15 random numbers with mean of 0 and standard deviation of 1 corresponding to the simulated height of each participant and 15 different random numbers with mean of 0 and standard deviation of 1 corresponding to the simulated working memory capacity of each participant. A mean of 0 and a standard deviation of 1 were chosen for simplicity but all the arguments below hold if ones 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

In summary, to generate two independent variables (with zero correlations), we will apply the following algorithm: 1. Pick randomly a number from a Gaussian distribution (mean = 0 and standard deviation =1) and assign this number to the height of this participant () 2. Pick randomly a number from a Gaussian distribution (mean = 0 and standard deviation =1) and assign this number to the working memory capacity of this participant () 3. Repeat for every individual in the population (N= 15)

By doing so, we obtain the following matrices:

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 wil 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). Interstingly, while 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 observing such a correlation coefficient is not so unlikely under the hypothesis that there is no correlation (the null hypothesis). In other words, we cannot conclude from this data that the variables height and working memory capacity are related in this population (and we cannot conclude that these are not correlated either).

## `geom\_smooth()` using formula 'y ~ x'

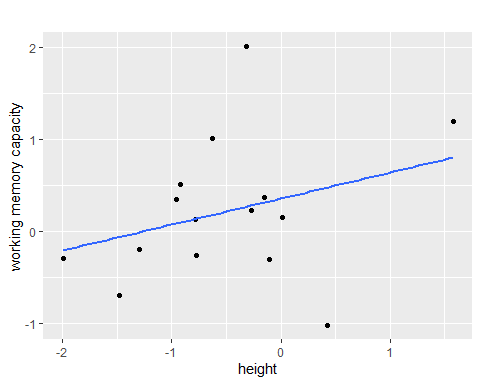


Figure 1: Absence of correlation from two randomly generated samples from two uncorrelated variables

Now, this is the data from a single city. Let’s now simulate what would happen if the same correlation was computed for 10^{4} different towns with enough inhabitants. In this case, we should repeat the process 10^{4} times in order to know whether the result obtained for Figure 1 is a special case or not. That is, for every simulated city (N=10^{4}), 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 10^{4} towns.

## simulating multiple correlation 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).

### doing that in R

Simmulating randomly generated parameters for a given number of participants can be done in with the function from the package. To draw 15 samples from a normal distribution, we can do

library(Hmisc) # for rcorr  
library(MASS) # for mvnorm  
set.seed(12)# to make sure everybody gets the same results  
# number of individuals in the population/sample  
PopSize = 15  
# 15 randomly generated number representing the height of the 15 individuals in the population  
x=mvrnorm(PopSize,0,1,empirical = FALSE)  
# 15 randomly generated number representing the working memory capacity of the 15 individuals in the population  
y=mvrnorm(PopSize,0,1,empirical = FALSE)  
# correlation between the x and y variables  
R<-rcorr(x,y)  
# display correlation coefficient  
R$r[1,2]  
# display corresponding p-values  
R$P[1,2]

We will then repeat this process 10^{4} times to obtain a reliable distribution of the p-values. To do so, we will use the ability of the mvnorm 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:

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 function to build two uncorrelated variables at the same time. This is repeated 10^{4} times (Nsim= 10^{4}) in the code below:

library(Hmisc) # for rcorr  
library(MASS) # for mvnorm  
set.seed(123)# to make sure everybody gets the same results  
Nsim=10000 # number of simulations  
PopSize = 15 #number os elements in each sample  
Sigma <- matrix(c(1,0,0,1),2,2)# covariance matrix  
res <- data.frame(p=rep(NA,Nsim),R=rep(NA,Nsim),Sig=rep(NA,Nsim)) # pre-allocating space for the dataframe  
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 second to the samples of the y parameter (e.g. working memory capacity)  
 D <- mvrnorm(n = PopSize, rep(0, 2), Sigma, empirical = FALSE)   
 R<-rcorr(D[,1],D[,2]) # correlation between the randomly generated x and y   
 res[i,"R"] <- R$r[1,2] # storing the correlation value for future use  
 res[i,"p"]<-R$P[1,2]# storing the p-value for future use  
 res[i,"Sig"]<-res[i,"p"]<0.05 # tag p-values smaller than the type II error threshold (here 0.05)  
}

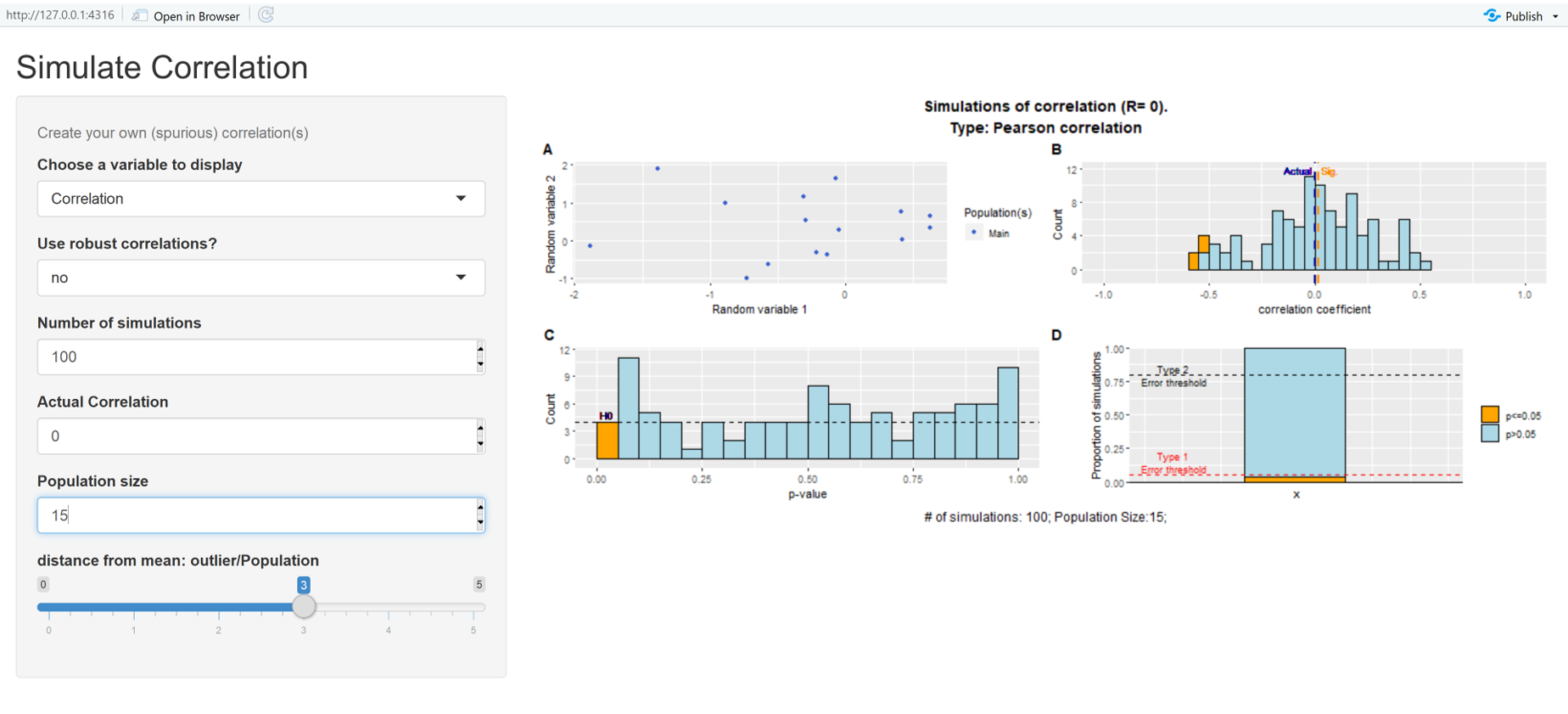
In the code above the R-dataframe stores the results of the correlation between x (first colunm of D - ) and y (second column of D = ) for each iteration of the simulations. The value of the correlation coefficient () and the associated p-value () are stored in this dataframe. One can use this dataframe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

library(ggplot2)   
# plotting the distribution of p-values  
Fcolor <- c("lightblue","orange")  
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 

## 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.

## Warning: Use of `res$SigBin` is discouraged. Use `SigBin` instead.

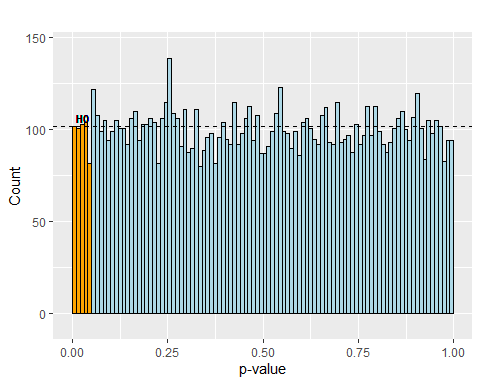
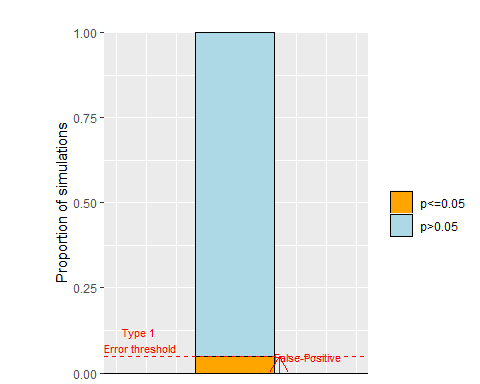


Figure 2: Distribution of p-values when no effect is present. Each simulated p-value corresponds to the correlation between samples taken from two uncorrelated variables

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 II 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 variables x and y (that where measured on 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.

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

 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 (and power) on the magnitude of the false-positive correlations

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).

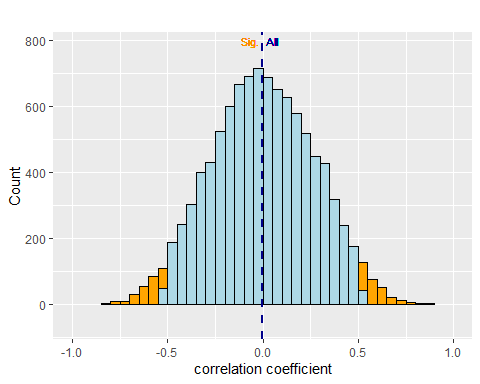


Figure 4: distribution of the correlations coefficient when R= 0 and N= 15

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 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 veracity of the effect but they at least give an better idea of the investigated correlation.

In other words high-correlations on a limited number of datapoints 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)).

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, it is easys to obtain high false positive correlations with low sample size and much more difficult to do so with high sample size.

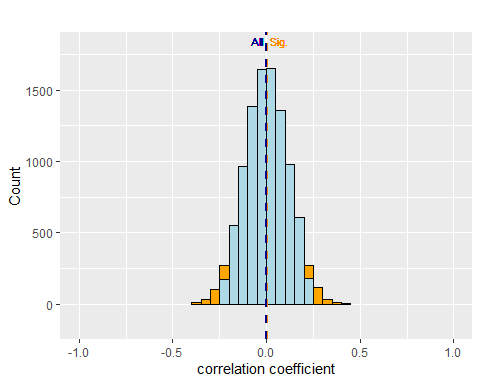


Figure 5: distribution of the correlations coefficient when R= 0 and the sample size is 75

The small number fallacy has important consequences for the publishability of findings. If two researchers are (unknowingly) reporting a false-positive correlation with N=15 and R=0.7 for researcher A and with N=75 and R=0.4 for researcher B, which one will more easily publish in a high-impact journal (whatever that means…)? Therefore, any scientist should be warry on exploratory correlations with low N.

***Assignment:*** *How big needs the population size to be to have all false positive correlations with a smaller than 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 10^{4} simulations, there is 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 NaN trials. 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:*** *Simulate repeatedly 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

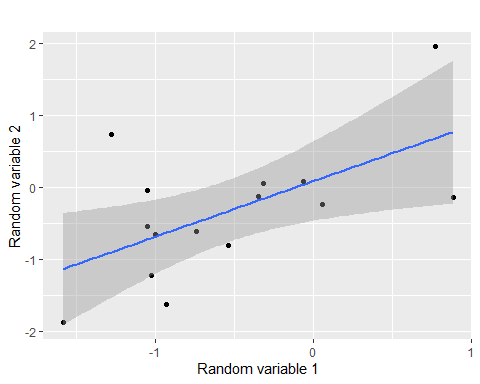
It is also interesting to simulate cases where the correlation is not zero. While simulating correlations with R=0 thaught us about false-positive correlations (type I error), the importance of sample size and the multiple comparison problem, simulating corelation with can teach us about power (and type 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 function 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:

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 below. 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.

 Together, these two parameters measured in 15 individuals are correlated with a correlation coefficient R= 0.595 (CI: [0.12, 0.85], p-value = 0.0192). 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

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 as we did earlier when no correlation was present and the distribution of p-values was completely flat (Figure 2). In contrast, When there is an underlying correlation between the two variables, the distribution of p-values becomes skewed with more p-values under 0.05.

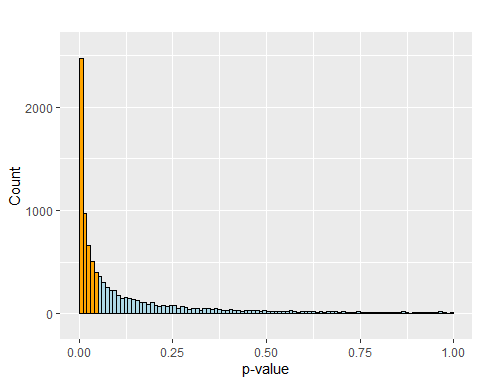


Figure 7: distribution of p-values when the simulated R is 0.5 and the sample size is 15

In this graph, it appears very clearly that p-values represents a continuous quantity and that there is nothing special about the 0.05 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 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 sizes and for different population size. 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).

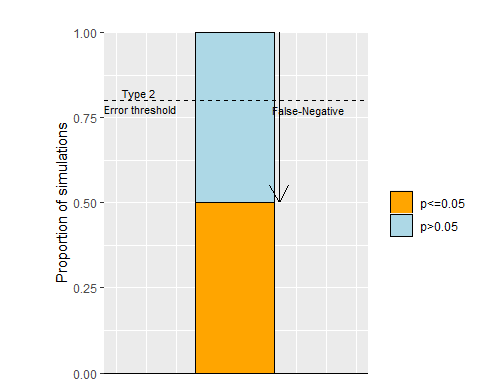
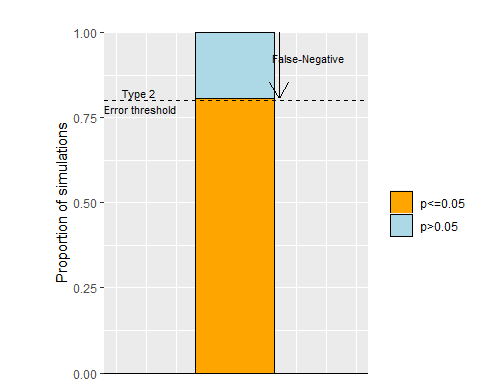


Figure 8: proportion of significant and non-significant p-values when the simulated R is 0.5 and the sample size is 15

As indicated on Figure 9, the expected power of an experiment is 80% (or 90%). Yet, this is here not the case. To detect a correlation R=0, 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 manipulations demonstrate 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 now in advance how big the effect size (e.g. the strengthh the correlation) will be.

***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 11, we can see that while many simulated correlations are around the value of 0.5, the value of the obtained correlations are spread over a large interval. Indeed, the simulated correlations exhibited a large 95% confidence interval: between 0.0098 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 ([0.2,0.74]). 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.

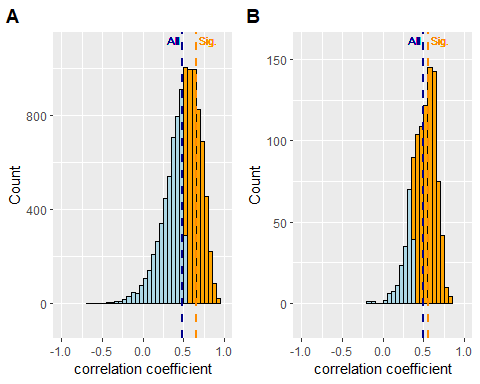


Figure 11: distribution of significant and non-significant correlation coefficients when the simulated R is 0.5 and the sample size is 15 (panel A) and 28 (panel B)

The accuracy of the correlation can also be quantified by the confidence interval associated with each correlation in particular. The 95% confidence interval give you a range of plausible values for the correlation for this particular sample/population. This interval has a 95% chance of containing the true effect size (here, the true value fo the correlation). To check this definition, we commputed the CI for each correlation presented on 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 10).

As a correlate of this demonstration, it is interesting to investigate what the probability that the correlation of a second experiment will fall into the CI of the correlation of a first experiment. For instance, the CI for the correlation presented in Figure 6 was [0.12, 0.85]. What is then the chance that a second experiment with the same sample size provides a correlation that falls in this CI. 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 this is the case (see panel B of Figure 10).

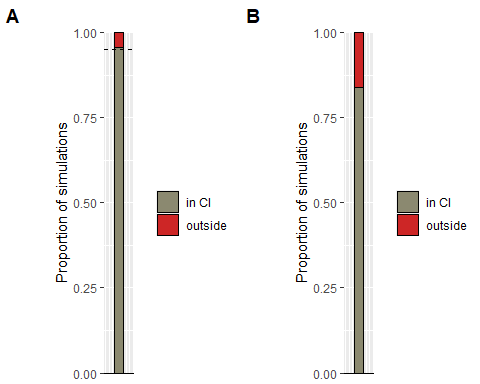
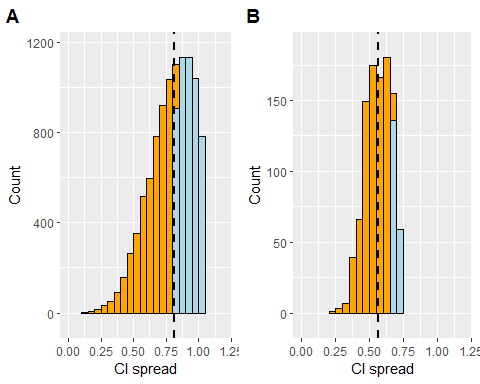


Figure 10: proportion of confidence intervals including the actual effect size of R= 0.5 (left panel) or the correlation coefficient of the next simulation (right panel). Sample size used is 28 but this is true for any sample size

In Figure 12, we compare the width of the confidence interval of the simulated correlations with N= 15 and N=28, respectively. 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 11). 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.

### understanding that publication bias inflates effect size

Figure 11 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 11, with N= 28), this average correlation (mean R=0.552) overestimates the actual correlation value (R= 0.5). In contrast, the average correlation over all samples (darkblue dashed vertical line on the graphs above) is much closer to the actual simulated correlation (with N= 28, average correlation is 0.499). Furhermore, 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 average correlation from all studies is R= 0.488) If we look at the mean correlation from only significant studies, it is always larger than the actual population average correlation. With low power, significant effect size are even more inflated.

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

### How can results be too good to be true?

It is clear from Figure 11 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 such series of 5 experiments 1000 times and the counted the proportion of significant correlation.

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 2.9% 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.*

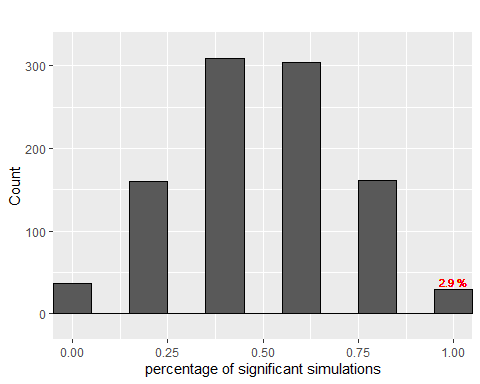


Figure 13: percentage of significant simulations among 5 simulations for R = 0.5 and sample size = 15

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 are use questionable practice to always obtain significant p-values. I personally find it depressing to assist to presentations with series of only significant studies with small sample size. This cannot reflect the reality (Lane et al. 2016).

Note that the percentage of series of 5 simulations with 5 significant studies increases with power and sample size (N=15: 2.9%; N=28: 30.7%). Yet, even with 80% power, these remain the minority (69.3% of the series of 5 simulations contain at least one non-significant p-value).

In contrast, the probability of observing only significant p-values decreases with the number of experiments considered in one series. For instance, when considering series of 10 experiments, the probability of observing 10 significant p-values is only 10.9%, 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 too good to be true (Francis 2014).

## Understanding p-values: summary and final assignments

In this 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 = 1000) 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 in order to make a statement about the veracity of an effect. In other words, making statements about effect is only possible if we get away from the law of small numbers (Tversky and Kahneman 1971). In other words, we need larger samples (because they provide more reliable estimates) and several experiments (because they allow us to choose between two hypotheses, between the null effect or the existence of an effect).

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.

### final assignments

I have describe p-values distribution when there is an effect (R=0.5) or when there isn’t (R=0). Now, let’s test whether the attentive readers can solve the following problem (Gigerenzer 2018): Consider a simple example: 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 ressearcher 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.

*If the alternative hypothesis is true, then the probability of getting another significant result equals the statistical power of the test, that is, 50%. If, however, the null hypothesis is true, the probability of getting another significant result would still be only 5%.*

You read two papers, each reporting three preregistered analyses. Which effect seems more reliable? A: p=0.024, p=0.034 and p=0.031 B: p=0.25, p=0.032 and p=0.0001 source: <https://twitter.com/EJWagenmakers/status/889987997046910978>

*I would trust the paper B because it is very unlikely to have three p-values just under 0.05 (paper A) while having one non-significant p-value in three experiments is likely even if the power is 80%*

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?

# how irregularities in the data can yield spuriously low p-values and how to test that a proposed solution works.

Looking at false-positives or power is only correct if the studied p-values really reflect a true 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 (CITE). We will therefore focus irregularities in the data. More specifically, we will look at how irregularities in the data can yield spurious correlations. Correlations are very sensitive to violations of the 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.

## 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). This will be done for both x and y parameters of this individual. By doing so, we obtain a graphic with one point detached from the other points (Figure 14).

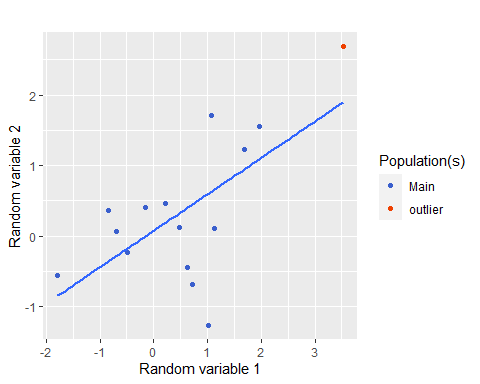


Figure 14: example of a correlation biased by an outlier (in red) when the simulated R is 0 , the sample size is 15 and the outlier is 3 standard deviation away from the mean

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=0.0096). However, it is unclear whether more correlations will turn out to be significant in the presence of an outlier. As outlined above, the p-value associated with a single correlation does not provide any information on the error 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 sensisitive are correlations to one outlier?

## Warning: Use of `resOut$SigBin` is discouraged. Use `SigBin` instead.

We will again repeatedly generate 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). Left panel of Figure 15 illustrates that both of these expectations are falsified by the data. Therefore, a single outlier in a population of 15 points increase the type I error tremendously. That is, there are 26 % significant p-values (right panel of Figure 15)!

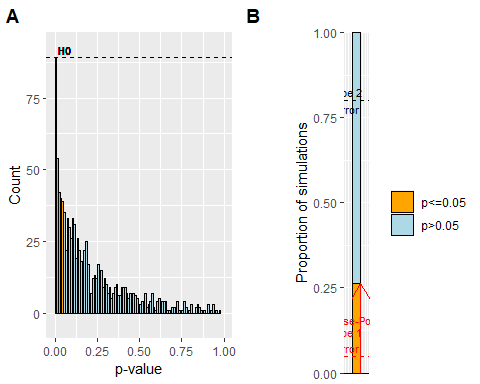


Figure 15: Distribution of p-values (panel A) and proportion of significant and non-significant p-values (panel B) when the simulated correlations are biased by an outlier 3 standard deviation away from the mean. True R= 0 and the sample size is 15

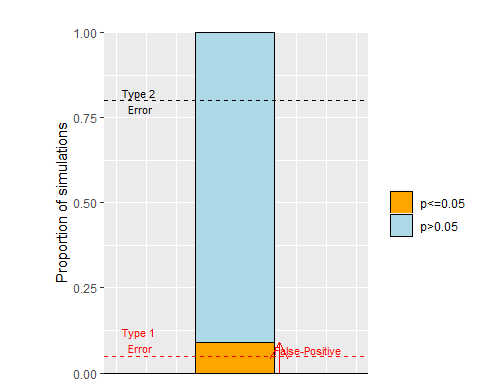
The interesting reader can use the Shiny App or the R code to explore the influnce of population size and the distance of the outlier from the mean on the increase in number of false-positive correlations caused by the outlier. Off course, this is also reflected in the value of the average correlation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

***Assignment***: *These analyses have been performed on the basis of R=0. The intersted 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 to be good 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 (= 9%) is still not close to the 5% level. 

### Solution: robust parametric correlation

In contrast, the use of winsorized correlation ((Pernet, Wilcox, and Rousselet 2013)) solves the problem as robust correlation methods are designed to handle outliers. This is confirmed by Figure 17, which shows that the influence of the outlier is eliminated and that the type I error rate (= 0.051) is very close to the 5% level. In R, this correlation is available in the package (Mair and Wilcox 2018).

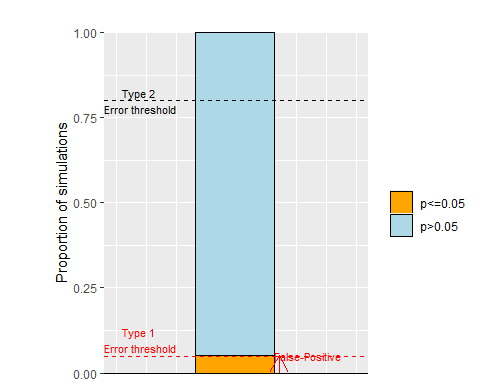


Figure 17: Ability of robust correlation (winsorized) to reduce the influence of an outlier on type I error. Proportion of significant and non-significant p-values for robust correlation when the simulated correlations are biased by an outlier 3 standard deviation away from the mean. True R= 0 and the 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 wanted to correlate the amount of explicit motor learning with working memory capacity across both young and old 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). Yet, 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 sensisitive are correlations to subgroups?

To generate subgroups, we first generate the x and y parameters for the 15 individuals from two uncorrelated variables (R=Rval). 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.

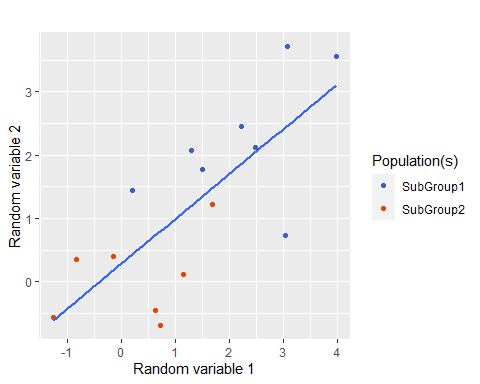


Figure 18: Example of correlation biased by differences between subgroups 2 standard deviation away from each other. True R= 0 and the sample size is 15

## Warning: Use of `resSub$SigBin` is discouraged. Use `SigBin` instead.

In this graph, we observe a strong correlation R=0.75 (CI: [0.38, 0.91], p=0.0013). Repeating such simulation (1000 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. If the two subgroups are 2 SD away, we observe 54 % 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*

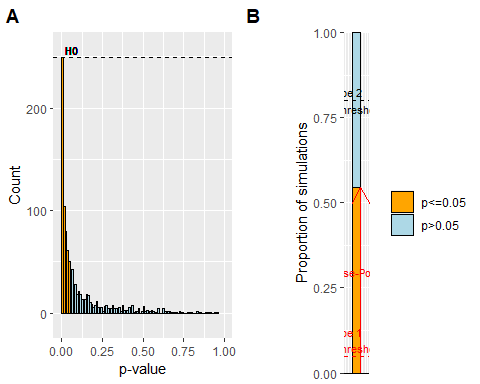


Figure 19: Influence of subgroup differences on the distribution of p-values (panel A) and on the proportion of significant and non-significant p-values (panel B) when the simulated correlations are biased by subgroups 2 standard deviation away from each other. True R= 0 and sample size N= 15

Note that for 2% of these significant correlations, the between-groups difference for the x and y parameters was not significant. Yet, these subgroups yielded a positive correlation.

In addition, the bigger the groups get, the smaller the difference between the subgroups must be to give rise to a significant correlation. For instance, for a sample of 15, a difference of 0.5 SD between group merely inflates the number of false-positive correlations (6% of correlations are significant). However, this difference becomes important for larger populations. For instance, a difference of 0.5 SD between subgroups increase the type I error rate to 8 and this number increases with sample size.

***Assignment:*** *The mean difference between the subgroups and their size also affect the magnitude of the false-positive correlations, which will tend to be larger because of the subgroups.*

### Solution to the subgroup problem

We will first show that Spearman correlation is again unable to solve the problem of subgroups eventhough it slightly reduces the problem. Then we will show that one needs to take the subgroups into account in a regression model 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)

### Spearmann 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 (57%) remains larger than the expected 5% level when the underlying variables are not correlated.

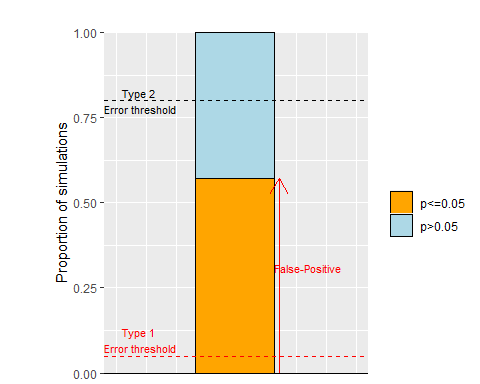


Figure 20: Inability of Spearman correlation to reduce the influence of subgroups on type I error. Proportion of significant and non-significant p-values for Spearman correlation when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R= 0 and the sample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

### Taking subgroups into account

In order 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 , the coefficient corresponds to the correlation coefficient between x and y if these variables are z-normalized. To z-normalize a variable, one needs to substract its mean from every observation and divide the outcome by the standard deviation of the sample (). Note that when the data are normalized, is always zero. To take subgroups into account, one needs to modify the regression equation as follow:

where is equal to zero for one subgroup and to one for the other. This equation makes the hypothesis that we expect the same correlation for both groups given that there are no interaction term between and . As shown on Figure 21, including a term in the regression coefficient is able to reduce the false-positive rate to the 5% level and should therefore be used when data from heterogeneous subgroups are pooled for correlations.

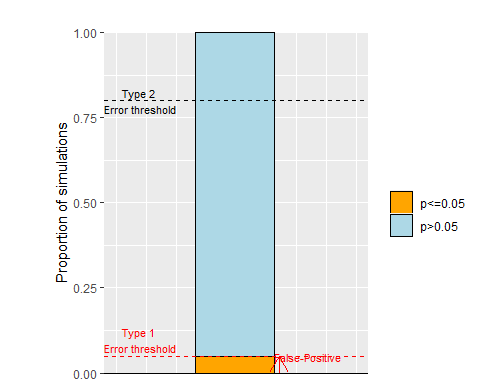


Figure 21: Ability of regression to reduce the influence of subgroups on type I error. Proportion of significant and non-significant p-values for standardized regression coefficient when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R= 0 and the sample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

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.

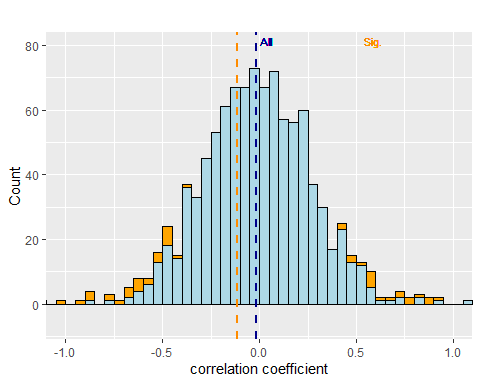


Figure 22: distribution of significant and non-significant standardized regression coefficients when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R is 0 and the sample size is 15

Taking subgroups into account does not prevent us from detecting correlations when they are present. 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 futher 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 11.

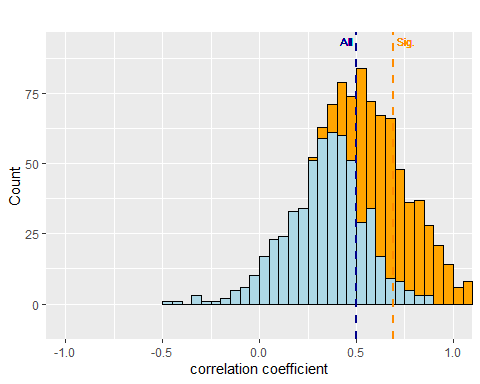


Figure 23: distribution of significant and non-significant standardized regression coefficients when the simulated correlations are biased by subgroups 3 standard deviation away from each other. True R is 0.5 and the sample size is 15

# What have we learned?

Simulations stimulate statistical thinking and yield important insights on statistics.

1. The distribution of p-value under no correlation 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, 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 because of publication bias is large, the with of the CI of the indivdiual correlation coefficient is large, the uncertainty of around the correlation coefficient is large, the effect of irregularities in the data (outliers or subgroups) is also prominent (Button et al. 2013).
6. 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 CI decreases).
7. Getting significant p-values is not difficult, one just needs to try enough different statistical tests (Bennett et al. 2011).
8. Having a mix of significant and non-significant results is normal and should become the norm. Some scientists present resuls that are too good to be true. Anybody should be warry of scientists who only publish significant results with small samples (Schimmack 2012).
9. 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).

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>.

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.

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

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>.

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>.

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>.

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>.

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>.

Mair, Patrick, and Rand Wilcox. 2018. *WRS2: Wilcox Robust Estimation and Testing*.

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>.

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 Sizeor 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](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC27401).” *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*.

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>.

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>.