PRZEGLĄD PSYCHOLOGICZNY, 2021, TOM 64, Nr 4, 177–193
DOI: 10.31648/przegldpsychologiczny.7889

Statistical power of a test – an analysis of a test's power, its role in the research methodology and the interpretation of (non-)significance in a low- (high-) powered test

Lilianna Jarmakowska-Kostrzanowska*
Institute of Psychology
Nicolaus Copernicus University in Toruń
0000-0003-3644-006X

ABSTRACT

Aim
This study has two main aims – to present the statistical power of a test and to discuss the main problems in analyses of a test's power with the use of a new-old tool. The applied tool is new because it marks a recent addition to a researcher's standard toolbox, but it is old because has been long recognized in statistics. The technical aspects of a power analysis in relation to the p-value were also discussed.

Hypotheses
The power analysis and statistical significance are concepts that originate from two different approaches to null hypothesis statistical testing (NHST). The lack of conformity between different approaches to the NHST paradigm creates problems in the interpretation of test results.

Conclusions
The required sample size can be determined in a power analysis, but the results of a power test are not easy to interpret. There are no clear rules for interpreting a statistically non-significant result in a high-powered test or a significant result in a low-powered test. A test's power does not confirm the statistically significant result, nor does it disprove the null hypothesis when the result is not statistically significant.

Keywords: statistical significance, p-value, power analysis, power of a test

* Correspondence address: Lilianna Jarmakowska-Kostrzanowska, Institute of Psychology, Nicolaus Copernicus University in Toruń, Gagarina Street 39, 87–100 Toruń. Poland. E-mail: lkostrzanowska@umk.pl.

In quantitative research, the number of observations (sample size) is regarded as a standard concern. Despite the above, the size of a population sample had been largely disregarded by researchers for many years. The size of a population sample consisted of as many people as could be solicited for a study. This attitude changed with the onset of the replication crisis (Ioannidis, 2005) which raised concerns about the credibility of research findings disseminated in the literature.

Science is based on replication. If the results of an original research study cannot be replicated, such findings are likely to be artifacts or overestimations. This was the case with the facial feedback hypothesis (Strack, Martin, Stepper, 1988) which postulated that subjects experienced a subjective increase in happiness when they stretched zygomaticus muscles in the pen-in-mouth test. This hypothesis was based on research results that could not be replicated (Wagenmakers et al. 2016). Similar observations were made in a study which investigated the cognitive ability of subjects who were asked to read text printed in a disfluent font. The original study demonstrated that students made fewer errors when presented with text printed in a hard-to-read font. However, the results of this experiment could not be replicated either (Meyer et al., 2015; Sirota et al., 2020). This is not the only study where the original findings were more promising than their replicates (cf. Klein et al., 2014).

It is generally believed that a negligent attitude towards sample size was one of the causes of the replication crisis. According to some researchers, the power analysis can offer a solution to this problem (Cumming, 2011; Murphy, Myors and Wolach, 2014). The power analysis is not a new concept, but it has gained popularity only in recent years, and it continues to generate new problems and challenges. Statistical significance and a test' power have been and will be used simultaneously to interpret the results. This article deals with a test' power, its implications in statistical analyses, and the interpretation of significant and non-significant results in tests with low or high power.

DEFINITIONS OF STATISTICAL POWER

The new tool can generate problems due to absence of a cohesive definition of a test's statistical power. Various definitions have been proposed in the literature. According to the authors of this concept, the power of a test is the ability to avoid type II errors (Neyman, 1977). Other authors, in particular in medical research, have argued that a high-powered test decreases the probability of a false negative result. In turn, Cohen (1988, p. 1) defined a test's power as the probability of obtaining a statistically significant result.

The first definition is based on the concepts developed by Neyman and Pearson, and readers who are unfamiliar with this approach risk interpreting *ignotum per ignotum* (the unknown by the more unknown). The second definition draws an analogy between a statistical test and a medical test, but it offers an intuitive explanation for readers who are not well versed in statistics. The third

definition is clearly worded, but it has a flaw that is difficult to eliminate (Mayo, 2018, p. 324). This definition prevents a correct interpretation of research results. Revisiting statistical hypothesis testing in psychology sheds light on this flaw and its consequences, and it will be discussed in detail in the next section. These concepts will be reviewed to provide the reader with a better understanding of the role played by power analysis in the existing set of statistical techniques.

Statistical power and statistical significance

Let us begin by making a simple observation that a scientist conducting empirical research has to expect variation in the results of the experiment. This implies that even a correctly built model containing several variables cannot be used to construct the deterministic equation in the form of y = f(x) that is known from chemistry or physics. A psychologist cannot claim that each person who experiences frustration will eventually become aggressive.

Psychologists rely on the null hypothesis significance testing (NHST) procedure to analyze their results (Wolski, 2017, Jarmakowska-Kostrzanowska, 2016; Gigerenzer, 2004). The NHST concept has evolved from two distinct approaches: Fisher's approach and the frequentist approach developed by Neyman and Pearson. Each approach proposes its own concepts and, more importantly, methods of interpreting the results. Despite many differences, both approaches concede that hypotheses are not verified by directly viewing data in a spreadsheet, but by transforming data with the use of a test statistic. Since data are random, the results of a test statistic are also random. The probability distribution of the results has to be inferred to draw conclusions about the underlying population. Certain assumptions regarding the tested hypotheses have to be made. If we assume that the null hypothesis is true and there is no relationship between the variables, the distribution of the test statistic is known. For example, Student's t-test has *n-1* degrees of freedom on the assumption that there are no differences between groups (*n* is the number of observations).

The one sample t-test will be used in this article for the purpose of simplicity. This test is typically applied to determine whether the sample mean is consistent with the theoretical value, for example, whether systolic pressure meets the ideal value of 120 Hg.

In Figure 1 (p. 180), the curves presented in both panels represent the Student's t-distribution for a population of $n = 50$ when the null hypothesis is true. The curves differ in the way t-distribution is interpreted in the discussed statistical approaches. The t-distribution in Neyman's approach is presented on the left, and the t-distribution in Fisher's approach is presented on the right. In the first approach, where the alternative hypothesis is regarded as the opposite of the null hypothesis, a researcher preexperimentally sets the cutoff point of errors when he/she thinks that he/she discovers some effect, whereas in reality it does not exists. This is known as a type I error, where a true null hypothesis postulating the absence of a relationship is incorrectly rejected.

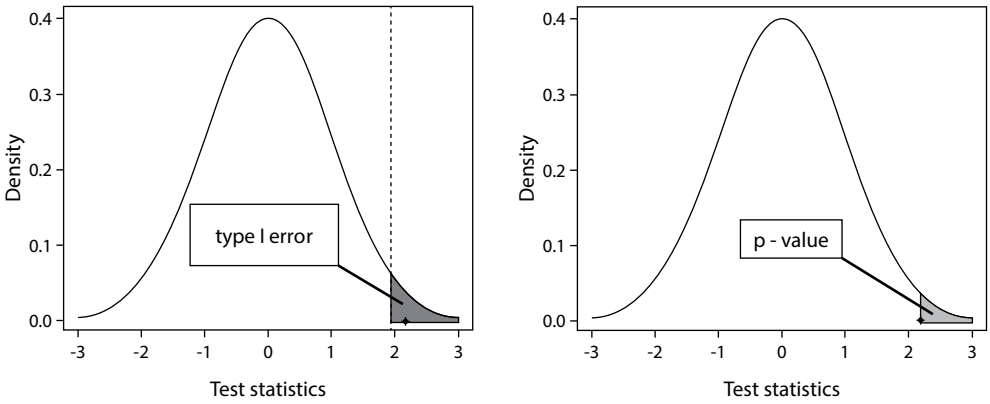

Figure 1. Test statistic and t-distribution in the frequentist approach and Fisher's approach.

Note: The frequentist approach is presented on the left. The probability of a type I error is determined before the analysis, and it marks the region where the test statistic is rejected (gray area under the curve marked with a dashed line), regardless of location of the black dot representing the obtained value. Fisher's approach is presented on the right. The position of the test statistic (black dot) is important because it is used to calculate the p-value. Source: own elaboration.

The error rate should not exceed 5%, and the critical value of the test statistic is marked with a dashed line. For $n = 50$ observations, the critical value of the test statistic is rounded off to 1.68. When this value is higher, the null hypothesis is rejected in favor of the alternative hypothesis. In itself, the value of the test statistic (represented by the black dot) is not important, what is important is whether the test statistic is localized in the gray-shaded area. This is why Neyman's approach is referred to as the fixed-alpha approach (Huberty, 1993). The probability of a type I error is not presented on the right side of Figure 1 because this concept does not exist in Fisher's approach. In this case, the value of the test statistic (represented by the black dot) plays a very important role because it is used to calculate the p-value, i.e. the area to the right of the black dot. If the gray-shaded area accounts for less than 5% of the area under the curve, the value of the test statistic is statistically significant. For this reason, Fisher's approach is referred to as the p-value approach (Huberty, 1993).

The choice of a specific approach does not play a role until the value of the test statistic is calculated. From the user's point of view, these approaches are indistinguishable: they rely on the same test statistics, similar terminology and reference values of statistical tools[2].

The reason why researchers tend to interpret Fisher's significance level as Neyman's type I error (Hubbard and Bayarri, 2003) becomes apparent when both t-distribution curves are overlaid. The question that remains to be answered is

[2] The significance level in Fisher's approach and a type I error in Neyman's approach both equal 5% .

whether a similar error can be made the other way around based on Cohen's argument that a type II error is inversely related to the significance level.

Let us go back to the left panel in Figure 1. A type II error (and the power of a statistical test) is an important element of Neyman-Pearson's approach. A type II error occurs when the researcher incorrectly assumes that there is no effect whereas it exists in the population and incorrectly rejects the true alternative hypothesis. This is a key concept in Neyman's approach, but it is not represented by any elements in Figure 1. So how is a type II error calculated? A curve illustrating the probability of t-distribution, which corresponds to the true alternative hypothesis, is also missing in Figure 1.

Every test statistic has two distributions, one of which applies when the null hypothesis is true, and other – when the alternative hypothesis is true. To plot a curve for the second distribution, a noncentrality parameter has to be introduced to the probability distribution. Researchers who rely solely on the NHST paradigm are not familiar with concepts such as the noncentrality parameter, type II error or the power of a statistical test. The noncentrality parameter does not have to be introduced to a probability distribution curve for testing the null hypothesis.

The noncentrality parameter is a value that determines the shape of the entire curve (Kelley, 2013), and it is dependent on other values, including the number of observations and the value postulated by the alternative hypothesis. The number of observations is a parameter with a specific value, whereas the alternative hypothesis in NHST is usually non-specific (e.g. H_1: $\mu \neq 0$ or H_1: $\mu > \mu_1$ or H_1: $\mu_1 \neq \mu_2$). As a result, a single curve is unlikely to represent the entire range of values that meet the statement of inequality in the alternative hypothesis. Therefore, the proposed parameter should have a specific form, for example H_1: $\mu = 0.2$ or H_1: $\mu_1 - \mu_2 = 0.5$. The second curve, known as the t-distribution curve for the true alternative hypothesis, can be plotted only when the noncentrality parameter is a specific value. The use of a specific value of a desired parameter, referred to as a point value in mathematical terminology, has important implications in a power analysis. The researcher is faced with the task of selecting the appropriate value from the entire range of possible values.

In the discussed example, let us examine Figure 2 on the assumption that H_1: $\mu = 0.5$ (an arbitrarily selected value). Figure 2 (p. 182) is a typical graphic representation of the relationships between classical statistical concepts – type I and II errors – and the relationship between power and a type II error. Type I (α) and type II (β) errors are not bound by a simple arithmetic dependence. Both fields are calculated based on different curves, and they are not summed up to a specific value. These fields are located on both sides of the critical value of the test statistic, and they are bound by a negative relationship. The higher the beta error, the lower the alpha error, and the higher the alpha error, the lower the beta error.

A simple arithmetic mapping exists between power and a type II error because it occupies an area under the same curve. This dependence represents power = 1 – beta. The larger the field corresponding to a type II error, the lower the power of a statistical test.

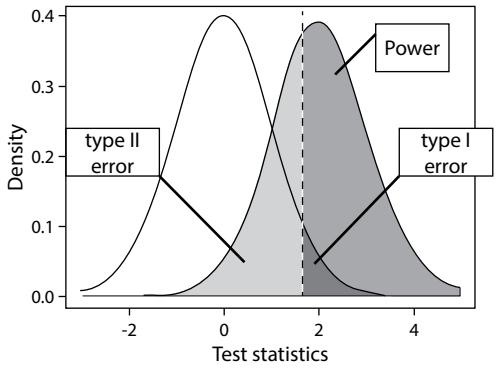

Figure 2. The relationship between a type I error (alpha) and a type II error (beta)

Note: Distribution of the test statistic when the null hypothesis is true (first from the left) and when the alternative hypothesis is true (when H1: $\mu = 0.5$; first from the right). Source: own elaboration.

The reason why the conflicting terminology in both statistical approaches (as indicated in Cohen's definition of power) is technically feasible should be explained. Let us examine the right side of the dashed line. The values of the test statistic which lead to the rejection of the null hypothesis in Neyman's approach, but are regarded as statistically significant in Fisher's approach, are distributed on the horizontal axis. The curve located above these values contains a field denoting the probability of a result within this range of values. Therefore, Neyman's power represents the probability of a statistically significant result in Fisher's approach, which is consistent with Cohen's definition of power (1988, p. 1). Therefore, the two approaches can be easily conflated by overlapping the curves which illustrate the concepts in each approach.

Technical aspects could be disregarded – the term "significance" was also used by Neyman, although in his own sense. However, power and statistical significance are concepts that belong to different approaches. The power of a statistical test is a frequentist concept that has been developed by Neyman. It rests on the assumption that if a given effect exists decisions are made in a large percentage of experiments, but not in a single study. In turn, Fisher's p-value does not require a high number of replications of the same experiment. In this case, a researcher decides whether to disprove null hypothesis based on the results of a single experiment. It could be argued that philosophical disputes should be left to philosophers, but the philosophical perspective is exactly what generates problems in the interpretation of data, for example in decisions on whether a significant result in a high-powered test validates the alternative hypothesis.

Conditional power

Cohen's definition of power cited at the beginning of the article has yet another flaw. In Figure 1 (p. 180), the curve corresponding to type II errors cannot be

plotted without a clearly stated alternative hypothesis. In the analyzed example, H_1: $\mu = 0.5$. If other values were adopted, for example H_1: $\mu = 0.2$ or H_1: $\mu = 0.7$, the curve would have a different shape (refer to Figure 3).

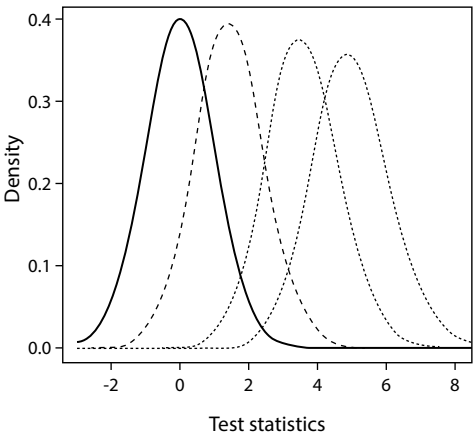

Figure 3. Changes in the shape of t-distribution curves for a true alternative hypothesis with different location parameters

Note. Increasing values of parameter μ influence the distribution of the test statistic for a true alternative hypothesis. The above affects not only the shape, but also the location of the distribution curve. Source: own elaboration.

Let us return to Figure 2. The researcher can control a test's power by shifting the dashed line to the left or right, which increases or decreases the value of a type I error. The above implies that a test's power is also influenced by sample size. Therefore, power is not a property of ~~in~~ a statistical test itself. A test can have a power of 30% or 99%, depending on the number of times the researcher is willing to make a mistake by rejecting a true null hypothesis (probability of type I errors) and a true alternative hypothesis (probability of type II errors), as well as the size the sample and the value of the searched parameter proposed in the alternative hypothesis. A test has a power of 80% under highly specific circumstances. Cohen's definition does not account for the conditional power of a specific data set.

Power analysis – significance of analysis

The number of observations, type I and type II errors, and the value postulated by the alternative hypothesis form a system of four equations with four unknowns. In a power analysis, the fourth variable is estimated when the values of the remaining three variables are known. These calculations can be performed with the use of free G*Power or SPSS (version 27 and higher) software tools.

Let us return to Figure 2. A type II error is represented by the area under the distribution curve on the left side of the dashed line. This area should be as small as possible. To move this area to the right, the researcher controls the factors that influence the noncentrality parameter, i.e. the number of observations, the probability of a type I error or the value of the parameter proposed in the alternative hypothesis.

Effect size

The effect size is the last concept that should be explained in this discussion. In a power analysis, effect size replaces the direct value of the examined parameter. An elaborate scientific definition of effect size was proposed by Kelley and Preacher (Kelley and Preacher, 2012). For the needs of this article, *effect* will be defined as a phenomenon or a relationship, and *size* will be defined as a measure of that phenomenon's magnitude.

A clearly formulated alternative hypothesis is a hypothesis that assigns a specific value to the expected effect size. For example, the standardized difference between two means (Cohen's d) equals 0.5, and Pearson's correlation coefficient (r) equals 0.7. The researcher has to select the appropriate effect size, which is not an easy task. Various methods have been proposed for selecting the effect size. Therefore, the researcher has to make an educated guess about the optimal effect size in a given population based on the available data. This decision has important implications when interpreting the results because the effect size has both theoretical (calculated in the power analysis) and practical (calculated based on empirical data in the experiment) relevance.

Standard threshold values of type I and type II errors in a power analysis

To calculate sample size (n) in a power analysis, the remaining three values have to be determined: effect size (which generates the previously discussed problems), type I error and type II error. By default, the probability of a type I error is 5%. In psychological research, a test is conventionally set to achieve 80% power – as Cohen advises (Cohen, 1988). The above implies that if a given effect exists, a type II error will reach 20%. In other words, the researcher will make an error in 20% of the cases. In one of five cases, the researcher will reject a true alternative hypothesis postulating that a given effect exists in the population, and will incorrectly assume that the effect equals zero.

Underpowered and overpowered studies

Taking a point value of .80 as a desired power of a test, studies fall into two categories – those with a power of less than 80% and those with a power higher than 80%. The first category of studies are underpowered, which implies that

the study has insufficient power to detect a given effect size. An underpowered study can be compared to a microscope which has a near focus for observing large objects. The microscope can be used to view a 15 cm-long earthworm, but not a roundworm with a length of several millimeters.

If the conventional standard is 80% power, to what extent can a study's power be reduced without significantly compromising the magnitude of the effect size? A study with 78% power is nearly as effective as a study with 80% power[3]. If a given phenomenon exists, the probability that it will be discovered is relatively high in both studies. However, a study with 50% power is as effective as tossing a coin. Thus, it is difficult to clearly set an interval around 80% of acceptable values.

Studies with a power much lower than 80% are seriously underpowered. Overpowered tests with a power much higher than 80% occupy the other end of the spectrum.

If the replication crisis was caused by studies that lacked sufficient power to detect the analyzed phenomenon, what risks are carried by studies with a power significantly higher than 80%? The power of a statistical test can be increased by analyzing a larger sample, but not all psychological studies involve online questionnaires. Research studies can adversely impact the tested subjects' well-being, which is an ethical cost, but they can be also fraught with financial and logistic problems (transport of equipment). If research costs can be minimized, a study's power should be maintained at 80%, and this threshold should not be exceeded. However, excessive power is not a problem of a statistical nature.

INTERPRETATION OF RESULTS: STATISTICAL SIGNIFICANCE AND STATISTICAL POWER

For many years, statistical significance was the only tool for determining whether the null hypothesis is true or false. Power analysis is a relatively new addition to a statistician's toolbox. Does additional information about a test's power enable the researcher to infer the truth about the null hypothesis or the alternative hypothesis? If a test has 80% power and the result is statistically significant, the null hypothesis is highly likely to be false. By the same token, a statistically non-significant result in a high-powered test implies that the null hypothesis is true. Four scenarios can be analyzed in this case.

Statistical significance and low statistical power

Common sense dictates that low-powered tests should produce only statistically non-significant results (in other words, the value of the test statistic should fall

[3] To paraphrase Cohen (Cohen, 1990), "surely, God loves 81% as much as 79%".

within a range that supports the null hypothesis). Such tests do not have sufficient power to detect the analyzed phenomenon. However, in some cases, the value of the test statistic falls within a range that rejects the null hypothesis, and the p-value exceeds the .05 threshold. The above suggests that even a low-powered test can pick up on an effect that is present. How do we interpret a statistically significant result in a low-powered test? This question can be rephrased as follows: why is the null hypothesis rejected in a low-powered test?

The following example can be analyzed to answer the above questions. For a Student's t-test with power 80% and Cohen's d of 0.1, the size of the studied population has to exceed $n = 787$. If the size of the sample is reduced to $n = 50$ for reasons of time, the test's power is reduced to 10% (Figure 4).

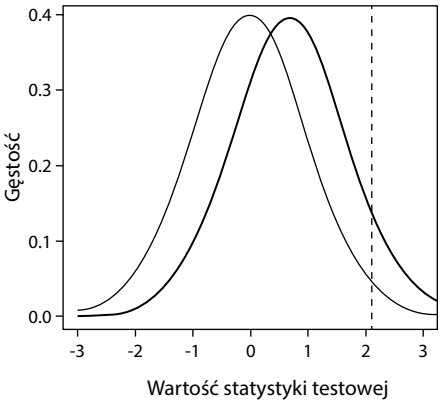

Figure 4. Distribution of the test statistic for a true null hypothesis and a true alternative hypothesis in a low-powered test.

Note: The distribution of the test statistic is marked with a thin line when the null hypothesis postulating the absence of an effect is true. The distribution curve is marked with a thick line when the alternative hypothesis postulating an effect size of 0.1 is true. The value of the test statistic which marks the beginning of the set of critical values is marked with a dashed line. Source: own elaboration.

The distribution curve is marked with a thin line when the null hypothesis is true. According to the discussed principles, a distribution is controlled by the noncentrality parameter when the alternative hypothesis is true. In turn, the noncentrality parameter corresponds to the effect size and the sample size. In this case, the effect size is 0.1, and the sample is small ($n = 50$). Therefore, when the alternative hypothesis is true, the entire distribution curve for the probable values of the test statistic is situated in the proximity (thick line) of the curve corresponding to the true alternative hypothesis.

The dashed line represents the critical value of the test statistic, in this case 2.10 (read from the t-distribution table). In most part, the area under both curves is located on the left side of the dashed line, which implies that the value of the test statistic is less than 2.10 in most cases. However, the optimal curve

is difficult to select because both curves are located close to each other. The values typical of a null hypothesis are nearly as probable as the values typical of an alternative hypothesis. However, in some cases, the value of the test statistic moves to the right side of the dashed line. In both cases (true null hypothesis or true alternative hypothesis), these values will be less typical because they are located farther away from both distribution curves. In both cases, this location requires a test statistic with a high value.

Which factors contribute to a high value of the test statistic? A large test statistic denotes a high difference between mean values and smaller variation in results. However, the observed distribution is not caused by real-world differences between population means, but by variation in the results (sampling error). Therefore, a result is statistically significant because the effect size has been overestimated. Even if the researcher is right and the study reveals an effect size, rarely encountered values also can exceed the .05 threshold. However, researchers who consider only results that have reached statistical significance have to deal with a problematic side-effect known as the "winner's curse" (Gelman, 2019)[4]. The result is statistically significant and can be published, but overestimation of the effect size can lead to a biased view on the phenomenon. The winner's curse is thus difficult to reverse. Initially promising results cannot be replicated. The decline effect, namely a decrease in effects over time, is one of the most interesting explanations of this phenomenon. The winner's curse affects not only the research team, but all those who base their knowledge on the results of overestimated studies or research alone. This is because they look at the phenomenon through distorted lenses.

Statistical significance and high (excessive) statistical power

A statistically significant outcome in a high-powered test is a dream come true for every scientist. Why? Because such a result indicates that the analyzed variables are bound by a relationship and the statistical test is capable of picking up on an effect size. The researcher might think that he/she can rely on the high power of a statistical test to assume that the alternative hypothesis was validated.

We will attempt to find a flaw in the assumption that a significant result validates the alternative hypothesis simply because the study has high power. The first counterargument is that from a purely scientific point of view, no single study is capable of proving the research hypothesis. The study should

[4] The power posing research conducted by Carney et al. (2010) is an example of a study that was probably biased due to the winner's curse phenomenon. The authors argued that people can boost self-confidence simply by assuming a powerful superhero posture every day (cf. Ranehill et al., 2015). Science is no longer confined to the academia, and many people without a scientific background are promoting the results of research studies. For example, a popular fitness trainer has been encouraging her clients to adopt power pose in their daily training to feel and behave more assertively (Lewandowska, 2018).

be replicated to determine whether a similar result can be obtained. However, this argument does not address the essence of a power analysis. Let us return to the moment before the study, when the researcher has to come up with the anticipated effect size to calculate the required sample size. In this case, a vague alternative hypothesis (e.g. H_1: $\mu \neq 0$) is replaced by a specific hypothesis (e.g. H_1: $\mu = 0.1$). So what is actually confirmed by a significant result in a high-powered test? The absence of equality in the original alternative hypothesis or the exact value of the effect size that was used in the power analysis to calculate the population sample? This question can be rephrased as follows: does a significant result in a high-powered test confirm that the effect size in a population is exactly as planned in the power analysis? Or does it merely indicate that an effect, even if infinitesimal (in this case $\mu = 0.1$), exists? This is an important question with no simple answer.

Let us analyze a different scenario, where a significant result is obtained in a study whose power considerably exceeds 80%, for example 95% or 99%. According to the adopted 80% benchmark of properly powered study, such a study is clearly overpowered. How do we interpret statistical significance in an overpowered study? To answer this question, we need to take a closer look at the relationship between p-value and sample size. Such a relationship does not exist when the null hypothesis is true, but it is strong when the alternative hypothesis is true. If the variables in the studied population are not correlated, p-value and effect size are independent, and an increase in population size does not affect p-value. The result will not be even less significant if a larger population is studied. However, when a correlation exists between variables in the examined population, the significance of all, even trivial, effects increases with sample size. The fact that sample size tends to be larger in high-powered tests raises concerns about overpowered studies (Utts, 2005; Brzeziński, 1997, p. 337). We will demonstrate that this is not a statistical paradox, but a problem in interpreting statistical significance in a high-powered test.

Figure 5 clearly explains why statistical significance increases with sample size. Let us assume that the researcher is expecting a small effect size. Figure 5 is similar to Figure 3, but it differs in sample size. The number of observations increases, the location and shape of distribution curves change, and the curves move apart. An increase in sample size does not change the effect size in the analyzed population, and the researcher still expects that Cohen's *d* will equal 0.5.

When sample size changes, the distribution of the test statistic for the alternative hypothesis will also change by moving away from the distribution for the null hypothesis.

Distribution curves shift every time a size effect is detected in the population, which implies that the real value of this parameter is non-zero.

It appears that the structure of classical statistical tests is the root cause of the described problem, i.e. a test's statistical significance increases with a rise in sample size. Meanwhile, the distribution of the test statistic is changed by both an increase in sample size and an increase in effect size.

The concerns associated with large samples and overpowered studies can be rationally justified only when statistical significance is equated with practical

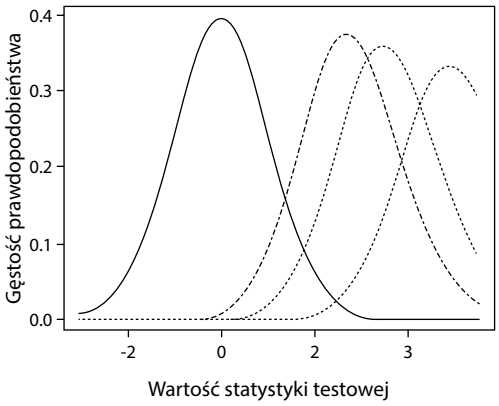

Figure 5. Gradual shift of noncentral distribution to the right despite a fixed size effect of Cohen's d equal to 0.5

Note: An increase in sample size affects the distribution of the test statistic both when the null hypothesis and the null hypothesis are true. The distribution becomes narrower for a true null hypothesis, but its location does not change. The distribution for the alternative hypothesis changes in both shape and location solely due to the change in sample size which affects the noncentrality parameter. This phenomenon is responsible for the illusion that all effects (even the most trivial) are significant. Source: own elaboration.

significance. Researchers who adopt this interpretation (Goodman, 2008) will also have reservations about large samples. In studies with large samples, it is the effect size (not affected by the sample size) that is useful to assess the importance of the phenomenon.

Otherwise, collecting a large sample might be regarded as a bad solution.

Statistical non-significance and low statistical power

The previous section demonstrated that low statistical power hinders the interpretation of a statistically significant result. However, a statistically non-significant result does not resolve the researcher's dilemma. If a study has low power, for example 20%, then 80% of the results will fall within the region of acceptance of the null hypothesis.

Figure 3 indicates that in a low-powered test, distribution curves are located close to one another when the null hypothesis and the alternative hypothesis are true. In some cases, the distance between curves is too small to infer whether the result is statistically significant because the null hypothesis is true and the test has insufficient power to identify an effect. The introduction of a large enough sample would overturn the argument that the result would be statistically significant if more people were studied.

Such results have two flaws. Firstly, statistically non-significant results are not published, which leads to the file drawer problem. The replication crisis is

associated not only with the effects of low-powered tests, but also with a skewed picture of the phenomenon. The majority of published results are statistically significant, whereas findings that do not cross the 5% significance threshold are not disseminated. The above applies to statistically non-significant results in both low-powered and high-powered tests.

The second problem is that some studies are doomed to have low statistical power and generate statistically non-significant results. In some scientific disciplines, observations are difficult to get (animal studies, studies of rare diseases), and the anticipated effect size is small. The researcher is at risk of obtaining a non-informative and statistically non-significant result or making a biased overestimate (winner's curse). When a large population cannot be examined (because it physically does not exist), the researcher can rely on the effect size or non-classical statistical methods, such as Bayesian inference.

Statistical non-significance and high statistical power

The fourth scenario combines a non-significant result (e.g. $p = .32$) with high statistical power (e.g. 90%). The researcher adopts the following line of reasoning: *the study has high statistical power; therefore, it can detect the anticipated effect. The result is non-significant; therefore, the null hypothesis postulating the absence of an effect must be true*. Thus, the conclusions that follow from statistical non-significance would be validated by the study's high statistical power.

The power of a statistical test can be easily confused with the sensitivity of a medical test. However, a statistical test is governed by different principles than a medical test. To fully understand the implications of a non-significant result, we have to go back to the moment when the researcher estimates the effect size for a power analysis. The researcher can never be certain that the effect size has been estimated correctly for a given population. It can be overestimated and assume a too high value. Let us assume that the effect size is 0.57 seconds. This value is unknown to the researcher. The researcher expects the effect size to be 1 second. In a power analysis, instead of including more subjects in the study, the researcher will decrease the size of the studied population to match the overestimated effect size. In other words, while looking for a splinter, the study has sufficient resolution to identify a wooden beam. The above will increase the probability of a type II error; the test will no longer have 80% power, and the third scenario will apply.

SUMMARY

The replication crisis is an ongoing methodological crisis that affects all fields of science. The power analysis was introduced to overcome these problems. The popularity of the power analysis is also on the rise because researchers are increasingly often expected to justify the size of the collected samples in published articles. A correctly performed power analysis has to be conducted before the

study (Nakagawa, Foster, 2004). Sample size must be considered when interpreting the significance of the results. However, statistical significance and power analysis are two separate problems. Statistical significance measures the extent to which the data fit the null hypothesis (Wasserstein, Lazar, 2016). In turn, the power analysis is conducted to control the probability of making a type II error in the long-term perspective (Neyman, 1977). The power analysis belongs to the frequentist paradigm. In the long-term, the researcher can only control the incorrect rejection of a true alternative hypothesis; therefore, he or she is unable to determine whether the result is erroneous.

Significant or non-significant results are difficult to interpret in studies with small or large effect sizes. Four scenarios were analyzed in this article, but none of them provides unambiguous instructions for interpreting the results. Similarly to interpretations of statistical significance based on the probability of a type I error (Hubbard and Bayarri, 2003), the formulation of conclusions about the truth of a null hypothesis based on the p-value and the probability of a type II error is yet another attempt to integrate incompatible statistical approaches.

REFERENCES

Brzeziński, J. (1997). *Metodologia badań psychologicznych*. Wydawnictwo Naukowe PWN.

Carney, D. R., Cuddy, A. J. C., Yap, A. J. (2010). Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. *Psychological Science*, *21*(10), 1363–1368. DOI: https://doi.org/10.1177/0956797610383437.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. DOI: https://doi.org/10.1037/0003-066X.45.12.1304.

Cumming, G. (2011). Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge. ISBN 9780415879682.

Gelman, A. (2019, January 4). Yes, it makes sense to do design analysis ("power calculations") after the data have been collected. Statistical Modeling, Causal Inference, and Social Science.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606. DOI: https://doi.org/10.1016/j.socec.2004.09.033.

Hubbard, R., Bayarri, M. J. (2003). Confusion Over Measures of Evidence (p 's) Versus Errors (α 's) in Classical Statistical Testing. *The American Statistician*, *57*(3), 171–178. DOI: https://doi.org/10.1198/0003130031856.

Huberty, C. J. (1993). Historical Origins of Statistical Testing Practices: The Treatment of Fisher versus Neyman-Pearson Views in Textbooks. *The Journal of Experimental Education*, *61*(4), 317–333. DOI: https://www.jstor.org/stable/20152384.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124.

Jarmakowska-Kostrzanowska, L. (2016). W statystycznym matriksie: kontrowersje wokół testowania istotności hipotezy zerowej (null hypothesis significance testing, NHST) oraz p-wartości. *Psychologia Społeczna*, *4*(39), 458–473. DOI: https://doi.org/10.7366/1896180020163906.

Kelley, K. (2013). Effect Size and Sample Size Planning. W: *The Oxford Handbook of Quantitative Methods* (Vol. 1, pp. 206–222).

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. DOI: http://dx.doi.org/10.1027/1864-9335/a000178.

Lewandowska, A. (2018, February 18). Power pozycja! Healthy Plan by Ann. https://hpba.pl/power-pozycja/.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (1st ed.). Cambridge University Press. DOI: https://doi.org/10.1017/9781107286184.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., Thompson, V. A., Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16– e30. DOI: https://doi.org/10.1037/xge0000049.

Murphy, K.R., Myors, B., i Wolach, A. (2014). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Routledge.

Nakagawa, S., Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, *7*(2), 103–108. DOI: https://doi.org/10.1007/s10211-004-0095-z.

Neyman, J. (1977). Frequentist Probability and Frequentist Statistics. *Synthese*, *36*(1), 97–131. JSTOR. DOI: 10.1007/BF00485695.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., Weber, R. A. (2015). Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science*, 26(5), 653–656. DOI: https://doi.org/10.1177/0956797614553946.

Sirota, M., Theodoropoulou, A., Juanchich, M. (2020). Disfluent fonts do not help people to solve math and non-math problems regardless of their numeracy. *Thinking & Reasoning*, 1– 18. DOI: https://doi.org/10.1080/13546783.2020.1759689.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768–777. DOI: https://doi.org/10.1037/00223514.54.5.768.

Utts, J. M. (2005). *Seeing through statistics* (3rd ed). Thomson, Brooks/Cole.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. DOI: https://doi.org/10.1177/1745691616674458.

Wasserstein, R. L., Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician, *70*(2), 129–133. DOI: https://doi.org/10.1080/00031305.2016.1154108.

Wolski, P. (2017). Istotność statystyczna III. Od rytuału do myślenia statystycznego. *Rocznik Kognitywistyczny*, *9*(2016). DOI: https://doi.org/10.4467/20843895RK.16.007.6413.