Null Hypothesis Significance Testing

Nathaniel E. Helwig

University of Minnesota

1 Background and Motivation

Suppose that we have some hypothesis (i.e., proposed idea) about a random variable, and we want to explore the plausibility of our hypothesis using a sample of data collected from the population. For example, suppose that we believe that the random variable's mean μ is equal to some hypothesized value, say μ_0. Given a sample of n independent observations $x_1, \ldots, x_n$ from the population, we could estimate the population mean μ using the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Assuming that our random variable of interest is continuous, the probability that $\bar{x}$ is exactly equal to μ_0 will be zero, i.e., $P(\bar{x} = \mu_0) = 0$. In other words, given any random sample of data from the population, we would expect the sample estimate $\bar{x}$ to differ from the hypothesized population mean μ_0 to some degree. But how small of a difference is "small enough" such that we should assume that the hypothesis $\mu = \mu_0$ is reasonable? And how large of a difference is "too large" such that we should assume that the hypothesis $\mu = \mu_0$ is unreasonable?

Statistical tests, also known as "significance tests" or "null hypothesis significance tests" (NHST), attempt to answer these sorts of questions. The commonly used procedure for NHST was first developed by Sir Ronald A. Fisher (1925) and further developed by Neyman and Pearson (1933). As a result, if the filed of statistics, you may hear the idea of NHST referred to as the Neyman-Pearson procedure for testing the significance of a hypothesis. The idea of NHST has caused quite a bit of controversy in the field of psychology (e.g., Nickerson, 2000) and has lead to many heated arguments. As will be discussed towards the end of this chapter, this controversy is mainly due to the traditions of misunderstanding and misusing (and sometimes outright abusing) the ideas of NHST in the field of psychology.

2 Forming a Statistical Hypothesis

Definition. In statistics, a hypothesis refers to some statement about the parameter (or parameters) of a random variable's probability distribution. A hypothesis will be denoted with the letter H followed by a colon and the statement about the parameter(s).

A variety of different distinctions can be made to describe the nature of a statistical hypothesis. One possible distinction is that between a simple versus composite hypothesis. A "simple" hypothesis is one that completely specifies the probability distribution of the random variable, whereas a "composite" hypothesis does not completely specify the random variable's distribution. For example, $H : X \sim N(\mu, \sigma^2)$ is a simple hypothesis because it specifies the family of the distribution (i.e., normal) as well as the parameter values that define the precise distribution from the family. In contrast, the hypothesis $H : E(X) = \mu$ is a composite hypothesis because it only specifies the mean of the random variable, without specifying the distribution family and/or other parameters that fully specify the distribution.

Another distinction that can be made is that between an exact versus an inexact hypothesis. An "exact" hypothesis specifies the exact value(s) of the parameter(s) of interest, whereas an "inexact" hypothesis specifies a range of possible values for the parameter(s) of interest. For example, the hypothesis $H : E(X) = \mu$ is an exact hypothesis, because it specifies the exact value of the population mean. In contrast, the hypothesis $H : E(X) \leq \mu$ is an inexact hypothesis, because it specifies that the population mean is less than or equal to some value. Note that a simple hypothesis must be exact, because you need the exact parameter values to completely specify the random variable's distribution. However, an exact hypothesis does not necessarily need to be simple, because you can make an exact hypothesis about a parameter without having to completely specify a variables distribution.

3 Testing a Statistical Hypothesis

Definition. The hypothesis that will be tested is referred to as the null hypothesis and is typically denoted by H_0. The alternative hypothesis, which is typically denoted by H_1, is the hypothesis that would be true if the null hypothesis is false. By definition, H_0 and H_1 can never both be true, i.e., $P(\{H_0 \text{ is true}\} \cap \{H_1 \text{ is true}\}) = 0$, and either H_0 or H_1 must be true, i.e., $P(\{H_0 \text{ is true}\} \cup \{H_1 \text{ is true}\}) = 1$.

As mentioned in the introductory section, the procedure of NHST involves first specifying a hypothesis that will be tested, and then using a sample of data to examine the plausibility of that hypothesis based on the evidence from the sample of data. Returning to the previous example of testing a hypothesis about a population mean, we would denote the null hypothesis that μ is equal to μ_0 using the notation $H_0 : \mu = \mu_0$. The corresponding alternative hypothesis is $H_1 : \mu \neq \mu_0$, which states that the population mean is not equal to the null hypothesized value of μ_0. In this example, the null hypothesis was exact, i.e., that μ is exactly equal to μ. However, this does not always need to be the case. Suppose that we wanted to test the null hypothesis $H_0 : \mu \leq \mu_0$ versus the alternative hypothesis $H_1 : \mu > \mu_0$. Note that these sorts of inexact hypotheses are likely to be more useful in scientific studies, where there is likely to be a hypothesized direction of an effect.

Regardless of whether the null hypothesis is exact or inexact, the general procedure for conducting a NHST remains the same (although, as will be later discussed, one specific detail of the procedure does differ depending on whether H_0 is exact or inexact). The general NHST procedure involves the following steps:

1. The null hypothesis H_0 and corresponding alternative hypothesis H_1 are stated.

2. If the null hypothesis is composite, additional assumptions are stated, which make it possible to completely determine the probability distribution of the random variable (or really the test statistic) under the assumption that H_0 is true.

3. A statistic $T = s(\cdot)$ is chosen, such that values of T can be used to evaluate the plausibility of the null hypothesis H_0. Note that the statistic T is often referred to as a "test statistic", which is typically some function of the sample estimate $\hat{\theta}$ of the population parameter θ.

4. A decision rule $\mathcal{D}$ is chosen, such that the rule specifies which values of T will be considered deviant enough to reject the fact that the null hypothesis H_0 is true. The region of values $R = \{t : t \text{ is extreme enough to reject } H_0\}$ is typically referred to as the "rejection region" for the test.

5. Given the observed data $\mathbf{x} = (x_1, \ldots, x_n)^\top$, if the observed test statistic $T_0 = s(\mathbf{x})$ falls in the rejection region R, then the null hypothesis H_0 is rejected in favor of the alternative hypothesis H_1. Otherwise, we accept (i.e., fail to reject) the null hypothesis.

***Example* 1.** As an example of the NHST procedure, suppose that we want to test the null hypothesis $H_0 : \mu = 75$ versus the alternative hypothesis $H_1 : \mu \neq 75$. (Note that we have just completed step 1 of the NHST procedure.) Also, let's assume that the random variable X is normally distributed with a standard deviation of $\sigma = 100$, which completes step 2 of the NHST procedure. As a test statistic, suppose that we define

$$T = \frac{\bar{x} - 75}{100/\sqrt{n}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the same mean. Note $T \sim N(0,1)$ when the null hypothesis H_0 is true, so large absolute values of T would be unlikely to observe if H_0 is true. This completes step 3 of the NHST procedure. As a decision rule, suppose that we define the rejection region as $R = \{t : |t| \geq 1.96\}$, i.e., we decide to reject $H_0 : \mu = 75$ if the absolute value of T is greater than or equal to 1.96. Note that if the null hypothesis H_0 is true, then the probability of observing such an extreme value of T is

$$P(|T| \geq 1.96) = P(T < -1.96) + P(T > 1.96) = \Phi(-1.96) + (1 - \Phi(1.96)) \approx 0.05$$

given that $T \sim N(0,1)$ when the null hypothesis H_0 is true. This completes step 4 of the NHST procedure. For the final step, we need to collect some data and calculate $\bar{x}$ (and the corresponding observed test statistic T_0), and determine whether the observed test statistic is in the rejection region. Suppose that our observed sample mean is $\bar{x} = 93$ and that we have $n = 100$ observations, which corresponds to an observed test statistic of $T_0 = 1.8$. In this case, we would fail to reject the null hypothesis given that T_0 is not in the specified rejection region, i.e., $T_0 \in R^c$.

***Example* 2.** Now suppose that we want to test the inexact null hypothesis $H_0 : \mu \leq 75$ versus the alternative hypothesis $H_1 : \mu > 75$. Note that the worst case scenario (i.e., when it would be most difficult to make the correct decision) would be when $\mu = 75$, so we will focus on this worst case scenario for forming the test statistic's (null) sampling distribution. Suppose that we make the same assumptions and use the same test statistic that was used in the previous example, which implies that $T \sim N(0,1)$ in the worst case scenario of $\mu = 75$. Note that if $\mu < 75$, then $T \sim N(\delta, 1)$ where $\delta < 0$. As a decision rule, suppose that we define the rejection region as $R = \{t : t \geq 1.65\}$, i.e., we decide to reject $H_0 : \mu \leq 75$ if the

value of T is greater than or equal to 1.65. Note that if the null hypothesis H_0 is true, then

$$P(T > 1.65) \leq 1 - \Phi(1.65) \approx 0.05$$

where the inequality (i.e., $\leq$) would be changed to an equality (i.e., $=$) if the worst case scenario of $\mu = 0.75$ was true. As before, assume that $\bar{x} = 93$ and $n = 100$, which corresponds to the same observed test statistic of $T_0 = 1.8$. In this case, we would reject the null hypothesis given that T_0 is in the specified rejection region, i.e., $T_0 \in R$. Note that we made a different decision this time—because we changed our rejection region.

4 Choosing a Decision Rule

In the previous two examples, the decision rules that were used to determine rejection regions were stated without providing much intuition—however the careful reader likely noticed the common theme between the two rules. In both examples, the decision rules were chosen such that the probability of incorrectly rejecting H_0 was approximately equal to 0.05. To understand where this idea came from, it is helpful to take a step back and introduce the possible outcomes of our significance testing procedure, which are displayed in Table 1. Note that there are two different ways to make a correct decision (i.e., true negative and true positive), and there are two different ways to make an incorrect decision (i.e., an error).

Table 1: Possible outcomes of a significance test.

	H_0 is True	H_1 is True
Accept H_0	True Negative	Type II Error
Reject H_0	Type I Error	True Positive

Definition. In a significance test, a type I error occurs when we reject a null hypothesis that is true, and a type II error occurs when we fail to reject a null hypothesis that is false. The probabilities of committing the two types of errors are typically denoted by

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = P(\text{Type I Error})$$
$$\beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = P(\text{Type II Error})$$

and the power of the test is defined as $1 - \beta = P(\text{Reject } H_0 \mid H_1 \text{ is true}) = P(\text{True Positive})$.

The probability of committing a type I error is often referred to as the "significance level" of the test. The Neyman-Pearson NHST procedure involves specifying the significance level, and then choosing the decision rule that maximizes the power of the test. In the previous examples, the decision rules were determined by setting the significance level at $\alpha \approx 0.05$ for each null hypothesis. Note that the value of $\alpha = 0.05$ seems to be quite popular in applied research studies in Psychology (see Cohen, 1994, for a satirical critique). However, there is no statistical reason why one should prefer a significance level of $\alpha = 0.05$. Furthermore, there is no reason that the Neyman-Pearson idea of fixing α and maximizing $1 - \beta$ needs to be used. If one does decide to use Neyman-Pearson's idea of fixing α at a particular value, then the significance level should be chosen to meet the needs of the application. For example, if making a false positive finding (i.e., Type I Error) is subjectively viewed to be costly, then the significance level should be set quite small (e.g., $\alpha = 0.0001$). In contrast, if making a false negative finding (i.e., Type II Error) is subjectively viewed to be costly, then the significance level should be set more liberally (e.g., $\alpha = 0.1$).

5 One-Sided versus Two-Sided Tests

Definition. A statistical test is referred to as <u>one-sided</u> if the null hypothesis is inexact, given that the alternative hypothesis is of the form $H_1 : \theta < \theta_0$ (less than) or $H_1 : \theta > \theta_0$ (greater than). In contrast, a test is referred to as <u>two-sided</u> if the null hypothesis is exact, given that the alternative hypothesis is of the form $H_1 : \theta \neq \theta_0$ (two-sided).

In the previous examples, note that the decision rule differed depending on whether the null hypothesis was exact (i.e., $H_0 : \mu = 75$) versus inexact (i.e., $H_0 : \mu \leq 75$). For the exact null hypothesis, our rejection region needed to consider the possibility that the test statistic could be too small (< -1.96) or too large (> 1.96), where the values of ± 1.96 are the values that approximately cut-off $\alpha/2$ in each tail of the $N(0, 1)$ distribution—which is the assumed distribution of T under the assumption that H_0 is true. In contrast, for the inexact null hypothesis, our rejection region only needed to consider the possibility that the test statistic was too large (> 1.65), where the value of 1.65 is the value that approximately cuts-off α in the upper tail of the $N(0, 1)$ distribution—which is the assumed distribution of T under the assumption that the worst case scenario of $\mu = 75$ is true.

Definition. Given a significance level α, the value(s) of the test statistic that are on the border of the rejection region are referred to as the critical values of the test.

For an inexact null hypothesis, the rejection region is one-sided so there is a single critical value. If the inexact null hypothesis is of the form $H_0 : \theta \geq \theta_0$, then the alternative hypothesis is a "less than" alternative hypothesis, i.e., $H_1 : \theta < \theta_0$, and the critical value is T_α. Note that T_α denotes the corresponding quantile of the sampling distribution of T under the assumption that the worst case scenario of $\theta = \theta_0$ is true. If the inexact null hypothesis is of the form $H_0 : \theta \leq \theta_0$, then the alternative hypothesis is a "greater than" alternative, i.e., $H_1 : \theta > \theta_0$, and the critical value is $T_{1-\alpha}$. Finally, for an exact null hypothesis, the rejection region is two-sided so there are two critical values, which are $T_{\alpha/2}$ and $T_{1-\alpha/2}$. For the previous examples, the critical value was $T_{1-\alpha} = 1.65$ for the one-sided test, and the critical values were $T_{\alpha/2} = -1.96$ and $T_{1-\alpha/2} = 1.96$ for the two-sided test.

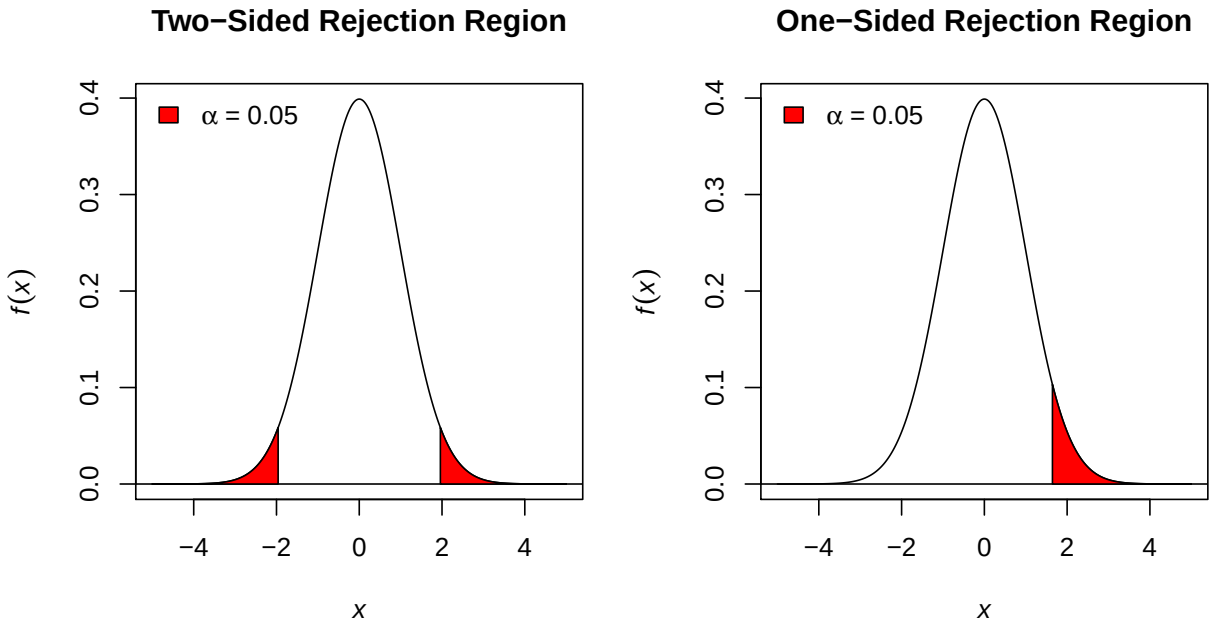

Figure 1: Visualization of the two-sided and one-sided rejection regions with $\alpha = 0.05$. Note that in both figures, the red shaded region contains 5% of the area under the $N(0, 1)$ density. For the one-sided rejection region, the figure depicts the "greater than" alternative.

6 p-Value of a Test

Definition. The p-value of a statistical test refers to the probability of observing a test statistic value as or more extreme than the observed test statistic T_0 under the assumption that the null hypothesis is true. What it means to be "as or more extreme" will depend on the direction of the alternative hypothesis:

- If $H_1 : \theta < \theta_0$, then $p = P(T < T_0 \mid H_0 \text{ is true})$
- If $H_1 : \theta > \theta_0$, then $p = P(T > T_0 \mid H_0 \text{ is true})$
- If $H_1 : \theta \neq \theta_0$, then $p = P(T < -|T_0| \mid H_0 \text{ is true}) + P(T > |T_0| \mid H_0 \text{ is true})$

where p denotes the p-value corresponding to T_0.

The p-value provides another way to talk about the decision of a significance test. Given a specified significance level α, the null hypothesis is rejected if $p < \alpha$, where p is the p-value associated with T_0. This is because $T_0 \in R$ whenever $p < \alpha$, i.e., there is a one-to-one relationship between the p-value being less than the significance level and the observed test statistic being in the rejection region. Of course, this implies that if $p > \alpha$, then the test statistic is not within the rejection region, i.e., $T_0 \in R^c$. Note that in the special case when $p = \alpha$, the test statistic is on the border of the rejection region, i.e., T_0 is equal to a critical value of the test. However, in practice you don't really need to worry about this possibility when you're working with continuous distributions, because the probability of observing a test statistic that is exactly equal to a critical value is zero.

7 Relation to Confidence Intervals

In some (low quality) statistical textbooks and journal publications, you may encounter the claim that there is a one-to-one correspondence between p-values and confidence intervals. In particular, you are likely to hear a claim along the lines of "if the null hypothesis value θ_0 is within the $100(1-\alpha)\%$ confidence interval, then the null hypothesis $H_0 : \theta = \theta_0$ would be retained using the significance level α". Unfortunately, this claim is not entirely accurate, given that there is a key distinction between how p-values and confidence intervals are calculated. As a reminder, a confidence interval is computed using the quantiles of the

sampling distribution of a parameter estimate—without being attached to any null hypothesis. In contrast, a p-value is computed using the sampling distribution of the test statistic under the assumption that the null hypothesis is true. In some cases, these two sampling distributions will be equivalent to one another; however, in other cases (such as proportion tests), the sampling distribution will differ depending on whether or not the null hypothesis is assumed to be true. Note that the classic example of when these two will be equal is when you are testing a hypothesis about the mean of a normal distribution—which seems to be the standard hypothesis that is tested in applied studies. But, more generally, the sampling distribution of a statistic (without any specified null hypothesis) will not necessarily be the same as the sampling distribution of the statistic under the assumption that H_0 is true.

8 Asymptotic and Nonparametric Tests

In step 2 of the NHST procedure, it was stated that you need to make additional assumptions to completely determine the probability distribution of the random variable (or the test statistic) under the assumption that H_0 is true. Note that this statement is true if you are conducting a standard parametric test, but there are other options for testing hypotheses. As a reminder, in the previous chapter we discussed three different methods for forming confidence intervals: parametric, asymptotic, and nonparametric. Analogues of these three options exist for testing hypotheses:

- *Parametric*: If the null hypothesized sampling distribution of T can be completely derived, then a parametric significance test can be conducted.
- *Asymptotic*: If the null hypothesized sampling distribution of T can the asymptotically derived, then an (asymptotically) approximate significance test can be conducted.
- *Nonparametric*: If the null hypothesized sampling distribution of T can be conditionally derived (given the observed data), then a nonparametric significance test, also known as a randomization test or permutation test, can be conducted.

Note that asymptotic tests are often used when (n is large and) maximum likelihood estimation is used to estimate the parameters, given that MLEs are asymptotically normally distributed. Nonparametric tests are powerful alternatives that can perform well across a variety of data generating distributions and sample sizes (see Helwig, 2019b,a).

9 Common Sense Considerations

9.1 Overview

As mentioned in the introduction, the concept of NHST has become a controversial topic in the field of psychology (and other social sciences). Throughout the remainder of this chapter, I will outline the various abuses, muses, and misunderstandings of NHST that seem to have caused this controversy. I will begin by discussing outright abuses of the NHST framework, which anyone should agree are problematic—but, sadly, several researchers are guilt of these abuses. Then I will address two issues that misinformed researchers often use to critique the NHST framework. In both cases, I will point that these seemingly unrelated critiques are centered around a common theme: psychologists do not typically use NHST in a way that will meaningfully move the science forward. I will end with a brief discussion—and encouragement—as to how the NHST framework could be used in a more useful way.

9.2 P-Hacking and Data Snooping

The pressure to find a "significant" result has lead some researchers to commit serious abuses of the NHST framework. More specifically, some researchers are so concerned with finding a "statistically significant" result that they are willing to go to great lengths to do so. As a reminder, a valid application of the NHST framework requires the user to specify their hypothesis and assumptions (steps 1 and 2), choose a test statistic and decision rule (steps 3 and 4), and then calculate the test statistic for the observed data. If done correctly, this procedure allows the user to control the type I error rate (i.e., the probability of making a false positive finding) at a desired level. However, in practice, many researchers do not implement NHST in this fashion. Instead, they change their hypothesis, assumptions, test statistic, and/or decision rule in order to obtain a final result that is "statistically significant" at the conventional $\alpha = 0.05$ level. Such dubious practices are referred to as p-hacking or data snooping. Note that if steps 1–4 of the NHST framework are adapted to produce a statistically significant result for a given sample of data, the probability of a false positive finding is likely to be much higher than the desired $\alpha = 0.05$ level. This is evident from the "reproducibility crisis" that exists in psychology (and other social sciences). In my opinion, many unreproducible results are likely to have been false positive findings.

9.3 Practical Significance

One critique of NHST that is routinely repeated in psychology is the idea that "statistical significance is not the same thing as practical (or clinical) significance". To understand the heart of this argument, consider the standard error of the sample mean $\text{SE}(\bar{x}) = \sigma/\sqrt{n}$, which was the denominator of the test statistic used in our examples. As the sample size n gets infinitely large, the standard error of the sample mean $\bar{x}$ approaches zero, i.e., as $n \to \infty$, we have that $\text{SE}(\bar{x}) \to 0$. Thus, for a large enough sample size, we would expect to reject the null hypothesis $H_0 : \mu = \mu_0$ for any value of μ_0 that is not the true population mean μ. In practice, the difference between the hypothesized μ_0 and the true μ may be so small that it makes little practical difference whether we believe the mean is μ or μ_0. If that is the case, finding a statistically significant result (i.e., rejecting $H_0 : \mu = \mu_0$) does not translate into a practically meaningful result. However, I do not see why this is a problem, and I think that this "critique" is rather silly. You should not fault the NHST framework because it works as intended, i.e., rejects an incorrect H_0 given enough power to do so. And, in reality, most applications of NHST in psychology are vastly underpowered, so worrying about the hypothetical situation of having perfect power (i.e., $\beta = 0$) doesn't seem too practical.

9.4 The Nil Hypothesis

The primary problem regarding the use of NHST in psychology (and other social sciences) is the fact that most applications are testing the wrong null hypothesis. More specifically, in many applications of NHST, the null hypothesis being tested is not formed using any useful knowledge about the data. Instead, researchers often test what is referred to as the "nil hypothesis", which refers to a null hypothesis of form $H_0 : \theta = 0$. Note that the value of the parameter that is specified in the null hypothesis should be your best guess of the population parameter—not some arbitrary value (such as 0) that is the default in whatever statistical software is being used. Most applications of NHST in psychology are backwards, such that the null hypothesis is arbitrary (e.g., $H_0 : \theta = 0$) and the alternative hypothesis is what the experimenter believes to be true (e.g., $H_1 : \theta \neq 0$). Note that this is completely backwards because rejecting a null hypothesis that you know is false will do nothing to move the science of psychology forward. The correct way to use NHST is to specify a null hypothesis that you think is true, and then design a (well powered) study to test the null hypothesis.

References

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist 49*(12), 997–1003.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Helwig, N. E. (2019a). Robust nonparametric tests of general linear model coefficients: A comparison of permutation methods and test statistics. *NeuroImage 201*, 116030.

Helwig, N. E. (2019b). Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *WIREs Computational Statistics 2*, e1457.

Neyman, J. and E. S. Pearson (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society 29*(4), 492–510.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods 5*(2), 241–301.