Null Hypothesis Significance Testing

Nathaniel E. Helwig

Associate Professor of Psychology and Statistics
University of Minnesota

October 17, 2020

Copyright © 2020 by Nathaniel E. Helwig

Table of Contents

Table of Contents

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.

- Suppose we believe that $E(X) = \mu_0$
- Use $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ to estimate μ

Unless we measure the entire population (or get REALLY lucky), we know that $P(\bar{x} = \mu_0) = 0$. But how much of a difference is okay?

- Given $\bar{x}$, should assume that the hypothesis $\mu = \mu_0$ is reasonable?
- How large of a difference is "too large" to believe that $\mu = \mu_0$?

History of Significance Testing

Statistical tests, also known as "significance tests" or "null hypothesis significance tests" (NHST), attempt to answer these sorts of questions about the statistical significance of findings.

NHST procedure was first developed by Sir Ronald A. Fisher (1925) and further developed by Neyman and Pearson (1933).

NHST has caused quite a bit of controversy in the field of psychology (e.g., Nickerson, 2000) and has lead to many heated arguments.

- Most of the controversy is due to misunderstandings

Table of Contents

Definition of a Statistical Hypothesis

In statistics, a hypothesis refers to some statement about the parameter (or parameters) of a random variable's probability distribution.

- In most cases, the parameter of interest is the expected value
- But you could make hypotheses about any parameter(s) of interest

A hypothesis will be denoted with the letter H followed by a colon and the statement about the parameter(s).

For the previous example, we could denote the hypothesis that we want to test using the notation $H : E(X) = \mu_0$.

Some Different Types of Hypotheses

Simple versus composite hypotheses:

- *Simple*: completely specifies the probability distribution of the random variable, e.g., $H : X \sim N(\mu, \sigma^2)$
- *Composite*: does not completely specify the random variable's distribution, e.g., $H : E(X) = \mu$

Exact versus inexact hypotheses:

- *Exact*: specifies the exact value(s) of the parameter(s) of interest, e.g., $H : \mu = \mu_0$
- *Inexact*: specifies a range of possible values for the parameter(s) of interest, e.g., $H : \mu \leq \mu_0$

Table of Contents

Null versus Alternative Hypothesis

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't both be true: $P(\{H_0 \text{ is true}\} \cap \{H_1 \text{ is true}\}) = 0$
- either H_0 or H_1 must be true: $P(\{H_0 \text{ is true}\} \cup \{H_1 \text{ is true}\}) = 1$

General Steps of Hypothesis Testing

1. State the null hypothesis H_0 and the alternative hypothesis H_1.
2. If H_0 is composite, state the assumptions needed to determine the distribution of the test statistic when H_0 is true.
3. A test statistic $T = s(\cdot)$ is chosen, such that values of T can be used to evaluate the plausibility of the null hypothesis H_0.
4. A decision rule $\mathcal{D}$ is chosen, which specifies the "rejection region" of the test, i.e., $R = \{t : t \text{ is extreme enough to reject } H_0\}$.
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.

Hypothesis Test Example 1

Suppose that we want to test the null hypothesis $H_0 : \mu = 75$ versus the alternative hypothesis $H_1 : \mu \neq 75$. (Step 1).

Assume that the random variable X is normally distributed with a standard deviation of $\sigma = 100$. (Step 2).

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. (Step 3).

- Note $T \sim N(0, 1)$ when the null hypothesis H_0 is true

Hypothesis Test Example 1 (continued)

Define the rejection region as $R = \{t : |t| \geq 1.96\}$. (Step 4).

- When H_0 is true $P(T < -1.96) + P(T > 1.96) \approx 0.05$

If our sample mean is $\bar{x} = 93$ and we have $n = 100$ observations, this corresponds to an observed test statistic of $T_0 = 1.8$.

We would fail to reject the null hypothesis given that T_0 is not in the specified rejection region, i.e., $T_0 \in R^c$. (Step 5).

Hypothesis Test Example 2

Now suppose that we want to test the inexact null hypothesis $H_0 : \mu \leq 75$ versus the alternative hypothesis $H_1 : \mu > 75$.

The worst case scenario (i.e., when it would be most difficult to make the correct decision) would be when $\mu = 75$.

Suppose that we make the same assumptions and use the same test statistic that was used in the previous example.

- $T \sim N(0, 1)$ in the worst case scenario of $\mu = 75$
- If $\mu < 75$, then $T \sim N(\delta, 1)$ where $\delta < 0$

Hypothesis Test Example 2 (continued)

Suppose that we define the rejection region as $R = \{t : t \geq 1.65\}$

- When H_0 is true $P(T > 1.65) \leq 1 - \Phi(1.65) \approx 0.05$
- The inequality (i.e., $\leq$) would be changed to an equality (i.e., $=$) if the worst case scenario of $\mu = 0.75$ was true

Assume that $\bar{x} = 93$ and $n = 100$, which corresponds to the same observed test statistic of $T_0 = 1.8$

We would reject the null hypothesis given that T_0 is in the specified rejection region, i.e., $T_0 \in R$

Table of Contents

Possible Outcomes of a Statistical Test

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

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).

In the previous examples, the decision rules were chosen such that the probability of incorrectly rejecting H_0 was approximately equal to 0.05.

Error Probabilities and Statistical Power

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

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

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

Some Vocabulary and Historical Context

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.

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 $\alpha = 0.05$.

Table of Contents

Exact versus Inexact Hypotheses (revisited)

A test is referred to as one-sided if the null hypothesis is inexact.

- $H_1 : \theta < \theta_0$ (less than) or $H_1 : \theta > \theta_0$ (greater than)
- Rejection region only needs to consider one direction of extremity

A test is referred to as two-sided if the null hypothesis is exact.

- $H_1 : \theta \neq \theta_0$ (two-sided)
- Rejection region needs to consider extremity in both directions

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.

Rejection Regions for One-Sided and Two-Sided Tests

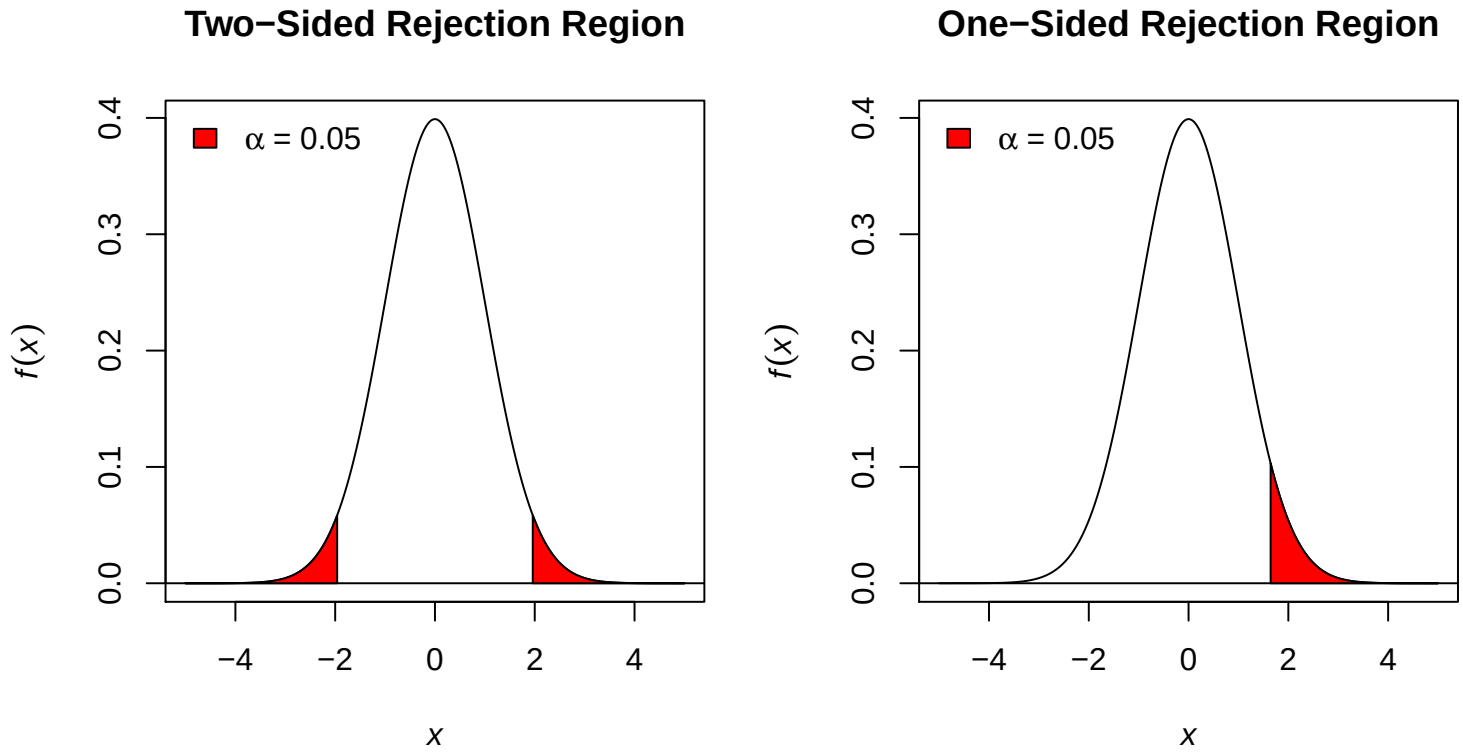

Figure 1: Visualization of the two-sided and one-sided rejection regions with $\alpha = 0.05$. Note that in both figures, the red shaded region contains 5% of the area under the $N(0, 1)$ density. For the one-sided rejection region, the figure depicts the "greater than" alternative.

P-Value of a Significance Test

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.

P-Values and Rejection Regions

Given a specified significance level α, the null hypothesis is rejected if $p < \alpha$, where p is the p-value associated with T_0.

- If $p < \alpha$, then T_0 is in the rejection region of the test
- If $p = \alpha$, then T_0 is equal to a critical value of the test

In the previous two-sided example $H_0 : \mu = 75$ versus $H_1 : \mu \neq 75$

$$p = P(T < -1.8) + P(T > 1.8) = 0.07186064 > \alpha = 0.05$$

In the previous one-sided example $H_0 : \mu \leq 75$ versus $H_1 : \mu > 75$

$$p = P(T > 1.8) = 0.03593032 < \alpha = 0.05$$

Relation to Confidence Intervals

You may encounter the claim that there is a one-to-one correspondence between p-values and confidence intervals. You may hear:

- "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"

This claim is not entirely accurate, given that there is a key distinction between how p-values and confidence intervals are calculated:

- p-values are calculated assuming that H_0 is true
- confidence intervals are not tied to any particular H_0

In some cases, the sampling distribution differs depending on whether H_0 is true (e.g., for inferences about proportions).

Table of Contents

Asymptotic and Nonparametric Tests

Analogues of the three CI options exist for testing hypotheses:

- *Parametric*: If the null hypothesized sampling distribution of T can be completely derived, then use a parametric significance test.
- *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.

Nonparametric tests are powerful alternatives that perform well in a variety of situations (see Helwig, 2019b,a).

P-Hacking and Data Snooping

As a reminder, a valid application of NHST 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

In practice, many researchers do not implement NHST in this fashion.

Instead, they change their hypothesis, assumptions, test statistic, and/or decision rule to obtain a "significant result".

- This can drastically inflate the type I error rate

Practical versus Statistical Significance

One critique of NHST that is routine in psychology is the idea that "statistical significance is not the same thing as practical significance".

To understand the heart of this argument, consider the standard error of the sample mean $\text{SE}(\bar{x}) = \sigma/\sqrt{n}$.

- As $n \to \infty$, we have that $\text{SE}(\bar{x}) \to 0$
- For large enough n, we would reject $H_0 : \mu = \mu_0$ for any $\mu_0 \neq \mu$

This "critique" is rather silly—you should not fault the NHST framework because it works as intended!

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.

In many applications of NHST, the null hypothesis being tested is not formed using any useful knowledge about the data.

- Often testing the "nil hypothesis" $H_0 : \theta = 0$

Most applications of NHST 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., $H_1 : \theta \neq 0$)

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.