2020

# Clear-Sighted Statistics: Module 13: Introduction to Null Hypothesis Significance Testing (NHST)

Edward Volchok
*CUNY Queensborough Community College*

*Clear-Sighted Statistics: An OER Textbook*

## Module 13: Introduction to Null Hypothesis Significance Testing (NHST)

> …I shall not require of a scientific system that it shall be capable of being singled out once and for all, in a positive sense; but I shall require that its logical form shall be such that it can be singled out, by means of empirical tests, in a negative sense: *it must be possible for an empirical scientific system to be refuted by experience*.[1]
>
> -- Karl R. Popper

### I. Introduction

Karl Popper, the twentieth-century philosopher of science, argued that any system of ideas that cannot be falsified—refuted or nullified—by empirical tests is not a science. For Popper and the framers of Null Hypothesis Significance Testing (NHST), scientists do not prove the "truth" of scientific propositions; they use empirical evidence to falsify or disprove them. The fictional detective, Sherlock Holmes, made this point to his sidekick Dr. Watson in the movie *Dressed to Kill*. He said, "The truth is only arrived at by the painstaking process of eliminating the untrue."[2] NHST is a widely used method of falsifying scientific propositions. It does not, however, verify propositions.

In this module, we will use the NHST framework laid out in Module 13, Introduction to Null Hypothesis Significance Testing, to conduct one-sample hypothesis tests.

After completing this module, you will be able to:

- Determine which of the three one-sample tests is appropriate.
- Conduct a one-sample z-test for the mean, μ.
- Conduct a one-sample z-test for the proportion, π.
- Conduct a one-sample t-test for the mean, μ.
- Calculate effect size, ES.
- Calculate statistical power and the probability of committing a Type II.

- Determine the required sample size to achieve the desired statistical power.
- Use charts of the normal distribution and student-t distribution to make conclusions about the results of your null hypothesis significance tests.

The phrase "null hypothesis" may sound odd to people living in the twenty-first century. The word "null" means without value, effect, consequence, or significance. In NHST, the null hypothesis means that any difference between the sample statistic and population parameter is merely the result of random sampling error or that there is no important effect. The null hypothesis is a proposition that is tested through a process of nullification or falsification.

**II. The Beginnings of NHST**

In 1925, the British statistician Ronald A. Fisher published the first edition of *Statistical Methods for Research Workers*. In his ground-breaking book, Fisher laid the foundation for "significance tests."[3] In 1928, Jerzy Neyman and Egon Pearson, the son of Karl Pearson, wrote a series of papers to correct what they considered flaws in Fisher's approach.[4] They called their approach "hypothesis testing." Among their innovations were the introduction of the alternate hypothesis, and Type I and Type II errors, which they called errors of the first type and errors of the second type. Their papers provide much of the structure of NHST used today. Fisher and Neyman waged an acrimonious debate about the merits of their positions until Fisher's death in 1962.

Over the years, textbook authors have combined Fisher's *significance testing* and Neyman-Pearson's *hypothesis testing* into a unified approach even though commentators have argued that the these approaches are inherently contradictory. In particular, Fisher's

concept of p-values is not compatible with the Neyman-Pearson hypothesis test, which it has become embedded.[5] Geoff Cumming, the author of *Understanding the New Statistics*, which we will discuss in Module 19, argued that consolidation of the Fisher and Neyman-Pearson approaches is a muddled amalgam:

> To some extent students may be expected to learn one rationale and procedure (Neyman-Pearson), but see a quite different one (Fisher) modeled in the journal articles they read…. It might be tempting to regard a mixture of the two approaches as possibly combining the best of both worlds, but the two frameworks are based on incompatible conceptions of probability. The mixture is indeed incoherent, and so it's not surprising that misconceptions about NHST are so widespread.[6]

For better or worse, we will follow convention and consider the Fisher and Neyman-Pearson approaches as a unified approach. In Module 19, we will discuss the shortcomings of NHST as practiced since the middle of the twentieth century.

NHST is a set of procedures that uses sample data and probability theory to determine whether a proposition about a population based on evidence derived from a sample should be rejected. At the final step of a hypothesis test, one of two decisions will be reached:

1. There is **insufficient evidence to reject** this proposition. In such cases, any difference between the sample statistics and the population parameter is considered merely random sampling error.

2. There is **sufficient evidence to reject** this proposition. In these cases, we declare the results "statistically significant." This means that the probability that the results are due to random sampling error is very low.

### III. What a Hypothesis is Not

In the early sixteenth century under the watchful eye of an impatient Pope Julius II, the great renaissance artist Michelangelo labored non-stop for four years painting the ceiling of the Sistine Chapel in Vatican City. Michelangelo's frescoes are among the world's greatest

artistic achievements. A section of this work, shown in Figure 1, illustrates God's creation of Adam. The creation story found in *Genesis*, the first book of the *Old Testament,* is *not* a hypothesis because it contains the proposition that the Divine Being exists, formed Adam from dust, and then gave him the breath of life.[7] Because the existence of the God of the *Old* and *New Testament* and the life of Adam are not subject to empirical verification, the existence of God and Adam are matters of faith, not science.

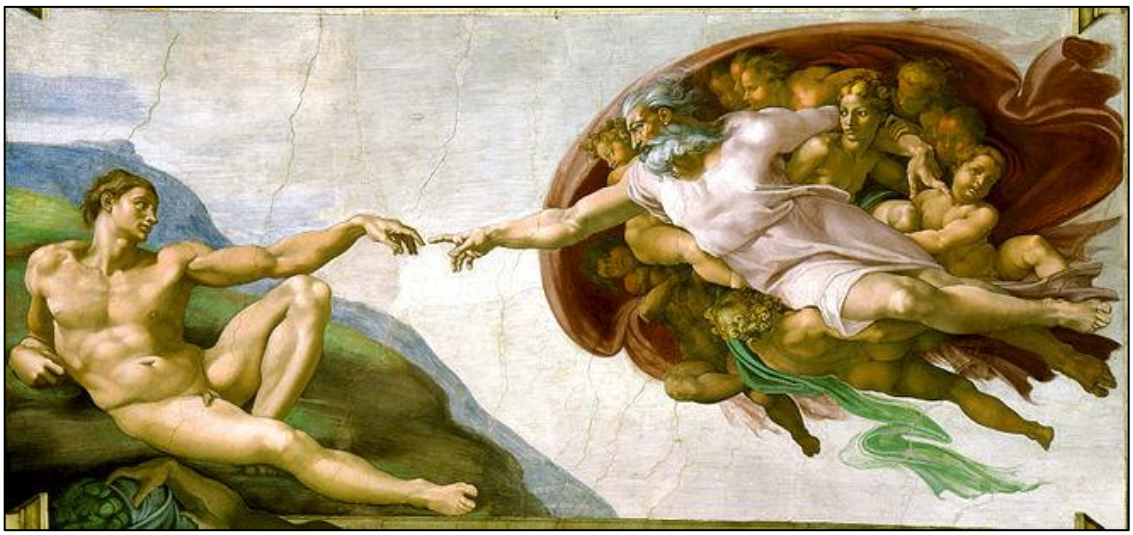


*Figure 1: God Creating Adam in Michelangelo's The Creation of Adam*

**IV. What an Hypothesis is and How It Differs from a Theory**

An hypothesis is often considered an "educated guess." This is not wrong. Hypotheses can be guesses, educated or otherwise. They can even be based on mean-spirited and ill-informed bigotry. But, hypotheses are more than everyday opinions, be they benign or malicious. An hypothesis is a preliminary proposition that provides an explanation of some phenomenon, ***which can be tested*.** In statistics, a hypothesis is a tentative statement about a population developed from a sample. Hypotheses are preliminary inferences based on limited evidence that can be refuted with hypothesis testing. This is what the movie version of Sherlock Holmes meant by the elimination of the untrue.

**The Difference Between Theories and Hypotheses**

A theory was once a hypothesis that has withstood repeated falsification and has become a well-established principle. Theories are propositions that provide a unified explanation of phenomena that have withstood repeated attempts to refute them. Common theories in the natural sciences include Albert Einstein's theory of relativity and Nicolaus Copernicus' heliocentric solar system in which the Earth revolves around the Sun. In economics we have John Maynard Keynes' theories concerning macroeconomics. In psychology there is Ivan Petrovitj Pavlov's classical conditioning theory.

The theory of natural selection was once a mere hypothesis developed by Charles Darwin, who devoted his life to collecting evidence that would elevate his hypothesis to a theory. Darwin's work inspired British statisticians in the late nineteenth and early twentieth centuries. In 1901, Darwin's younger cousin, Francis Galton[8], who created the concept of correlation, launched *Biometrika* along with Karl Pearson, and Raphael Weldon, a biologist. This journal, which still exists today, is dedicated to the statistical study of biological problems with Darwin's theory of natural selection as its starting point.[9] In the first issue of *Biometrika*, the editors asked, "…may we not ask how it came about that the founder of our modern theory of descent [Charles Darwin] made so little appeal to statistics?"[10] The goal of the journal was to give Darwin's natural selection stronger statistical support.

Like hypotheses, theories can be falsified, but this happens infrequently and the implications are far more important. When a theory is falsified, a paradigm shift may result.[11] Paradigm shifts open up new approaches to understanding phenomena that previously have not been considered. Darwin's theory of evolution based on natural

selection, which he presented in 1859 in *On the Origin of the Species*, ushered in a paradigm shift that changed our view of humanity's place in nature. After Darwin, many people began to think that humanity is a part of nature, not above it.[12]

A paradigm shift is a concept developed by the twentieth century philosopher of science Thomas S. Kuhn.[13] A paradigm shift is a scientific revolution where the core concepts underlying a scientific discipline fundamentally change. In Module 19, we will briefly consider whether the science of statistics and the notion of NHST is in the midst of a paradigm shift.

**V. A Non-Mathematical NHST**

You will recall that John W. Tukey likened descriptive statistics, or exploratory data analysis, to detective work and inferential statistics to a trial before a jury or judge.[14] A detective working in law enforcement gathers evidence about a crime. The detective may collect sufficient evidence to warrant an indictment, which is a formal initiation of a criminal case against a person. In an indictment, the person charged, the defendant, and stands trial before a jury or a judge. The central proposition of our criminal justice system is that the defendant is *presumed innocent* of the crime until the prosecutor can convince the jury *beyond a reasonable doubt* of the defendant's guilt.

A criminal trial and NHST share the same premise. In the trial, the defendant is presumed "not guilty." This is, in essence, what statisticians call the *Null Hypothesis*, $H_0$, which is pronounced either as H-zero, H-oh, H-null, or H-nought. The assumption underlying it is that the sample statistics, ($\bar{X}$, p, $s^2$, or r) equals the corresponding population parameter ($\mu$, $\pi$, $\sigma^2$, or $\rho$). In the case of chi-square tests, we compare observed distributions with expected distributions. The null hypothesis states that any observed

difference between the statistics and the parameter is merely the result of random sampling error or chance. The null hypothesis is based on the assumption that there is no difference, nothing happened, there is no effect, or no credible evidence. It is a provisional explanation that will be tested and possibly refuted. The NHST always deals with the population parameter. In Modules 14 through 19 we will deal with a variety of hypothesis tests. The Null and Alternate Hypotheses show only the population parameter, never the sample statistic. In addition, with z and t tests, the null hypothesis has one of the three equal signs, =, ≤, or ≥.

The prosecutor presents evidence to the court that he or she hopes will prove the case against the defendant beyond a reasonable doubt.[15] The prosecutor's case is what statisticians call the *alternate hypothesis*, which is represented by $H_1$ or $H_A$. The alternate hypothesis is sometimes called the *research hypothesis*. With NHST there can only be two hypotheses. Taken together, the Null and alternate hypothesis are an "either/or" proposition. The two hypotheses are mutually exclusive and collectively exhaustive. When the jury decides that the defendant is guilty beyond a reasonable doubt, the null hypothesis has been falsified and the defendant is declared guilty.

With NHST, the alternate hypothesis is accepted only when there is sufficient evidence to falsify the null hypothesis. As Ronald Fisher wrote in 1935, "…the Null Hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the Null Hypothesis."[16]

Like the null hypothesis, the alternate hypothesis deals with parameters: μ, π, $\sigma^2$, or ρ. It, too, never shows the sample statistic. In addition, with z and t tests, the

null hypothesis has one of three signs that are the opposite of the sign in the null hypothesis, ≠, >, or <. With chi-square tests, the alternate hypothesis states that the observed set of frequencies do not match the expected set of frequencies.

The verdict in a jury trial is based on the "beyond a reasonable doubt" principle, which cannot be quantified. In contrast, the decision to reject the null hypothesis is based on a quantitative measure called a level of significance, which is often called an alpha (α) level. As we discussed when we reviewed confidence levels, significance levels are found by (1 – the confidence level). Fisher introduced this concept in his book, *Statistical Methods for Research Workers*. Neyman and Pearson modified it. Today we set the significance level when we begin testing. Both Fisher and Neyman-Pearson gave tacit approval to using five percent significance levels. The 5 percent significance level is still the most commonly α used today, although, as we shall see in Module 19, the habit of using five percent significance levels has come under increased scrutiny. Sometimes a one percent significance level may be used, which makes it harder to reject the null hypothesis. When the test is a preliminary or pilot study, a 10 percent significance level may be used, which makes it easier to reject the null hypothesis.

Whenever we reject the null hypothesis, we do *not* consider the alternate hypothesis to be "true." It becomes the new null hypothesis, and is subject to falsification. When we do not reject the null hypothesis, we never say that it is "true." In fact, we never say that we "accept the Null Hypothesis." We say, "we failed to reject the Null Hypothesis." We always doubt the "truth" of our hypotheses. Similarly, in a jury trial, when the defendant is not convicted, the judgment is that the defendant is "not guilty." The defendant is never declared "innocent."

As with jury trials, null hypothesis tests are subject to two kinds of errors. Søren Kierkegaard, the nineteenth century Danish existentialist philosopher, got very close to the essence of Neyman-Pearson's Type I (alpha or α) errors and Type II (beta or β) errors. He wrote, "…one can be deceived in believing what is untrue [Type I Error], but on the other hand, one is also deceived in not believing what is true [Type II Error]."[17]

We can represent criminal trials and NHST with a 2 by 2 matrix to delineate the range of decisions regarding the null hypothesis and the two types of errors that can be made.

*Table 1: Type I and Type II Errors*

| | **Accept Null Hypothesis** | **Reject Null Hypothesis** |
|---|---|---|
| **Correct Decision** | **Jury correctly acquits the defendant. The analyst correctly fails to reject the null hypothesis.** | **Jury correctly convicts the defendant. The analyst correctly rejects the null hypothesis.** |
| **Incorrect Decision** | **Type I Error, α error, or false positive. An innocent man is convicted. The analyst wrongly rejects the null hypothesis.** | **Type II Error, β error, or false negative. An guilty man is acquitted. The analyst wrongly fails to reject the null hypothesis.** |

A jury or a research analyst can make one of two correct decisions.

- **Correct Decision** #1: Correctly fail to reject the null hypothesis. A man who did *not* commit the crime is acquitted. The difference between the population parameter and sample statistic is considered random sampling error.

- **Correct Decision #2:** Correctly convict a guilty man, or the analysts correctly decide that the parameter and statistic are not equal.

A jury or research analysts can make one of two incorrect decisions.

- **Type I Errors - False Positives:** A Type I error occurs when we reject the null hypothesis when we should have failed to reject it. In a jury trial, a Type I error occurs when the jury convicts an innocent man. A doctor who tells his patient that her diagnostic tests show she has pancreatic cancer when she does not, has committed a Type I error. The long-term probability of committing a Type I error is set by the level of significance or α. For any particular hypothesis test, however, we cannot tell if we committed a Type I error. All we know is our long-term probability of making such an error.

- **Type II Errors - False Negatives:** In a jury trial, a Type II error occurs when the jury **fails to convict a guilty man**. A doctor can commit a Type II error when he tells a patient she is in perfect health and then she suffers a fatal stroke later that day. In NHST, a Type II error is due to lack of statistical power. Statistical power is defined as the power of a test to appropriately reject the null hypothesis; which is to say, find a significant difference between the parameter and the statistic when one exists. Statistical power is found by 1 – P(β), 1 minus the probability of a Type II Error. The maximum statistical power is 1 or 100 percent and the minimum is zero. A power of 1 means that there is a 100 percent chance of finding a statistically significant event, and a zero means that it is impossible to find such an event. The generally accepted minimum level of statistical power is 80 percent. An 80 percent power means that the test has an 80% probability of correctly rejecting the null hypothesis. Low powered tests present problems even when the results are statistically significant because the difference between the sample statistics and population parameter may be a statistical accident. Low-powered tests make it difficult for other researchers to reproduce the results of the test. Over-powered tests are problematic as well. An over-powered test will find statistically significant results when the results have no practical significance.

Figure 2 shows the relationship between Type II errors and Type I errors, population variance or $\sigma^2$, and sample size or n. When the probability of a Type II error goes up, the probability of a Type I error goes down. When variance goes up, the probability of a Type II error increases, and the probability of a Type II error increases when the sample size decreases.

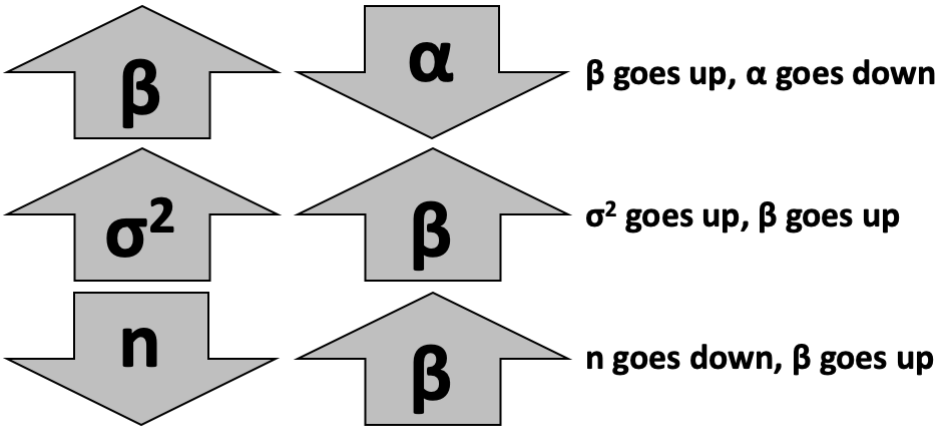



*Figure 2: The Relationship of Type II Errors to Type I Errors, Variance, and Sample Size*

Sadly, most introductory textbooks devote little, if any, attention to calculating statistical power and the probability of a Type II error. Worse still, statisticians like John Ioannidis have demonstrated that scholarly literature is plagued with underpowered studies that lead to false findings.[18] In *Clear-Sighted Statistics*, we will review how to calculate statistical power, the probability of a Type II error, and effect size (ES), which, broadly speaking, measures the strength of the relationship under investigation. We will do this using Microsoft Excel, statistical tables presented by statistician Jacob Cohen and his co-authors[19], and the popular program called G*Power, which is available for free for Windows and Macintosh computers.

Four factors affect statistical power.

1) The higher the selected significance level (α), the lower the statistical power.

2) The larger the sample size, n, the greater the statistical power.

3) The more variable the data, the lower the statistical power.

4) The smaller the effect, the lower the statistical power.

**VI. Null Hypothesis Testing: Step-By-Step**

All null hypothesis tests use six basic steps. Figure 3 shows the steps in the NHST cycle.

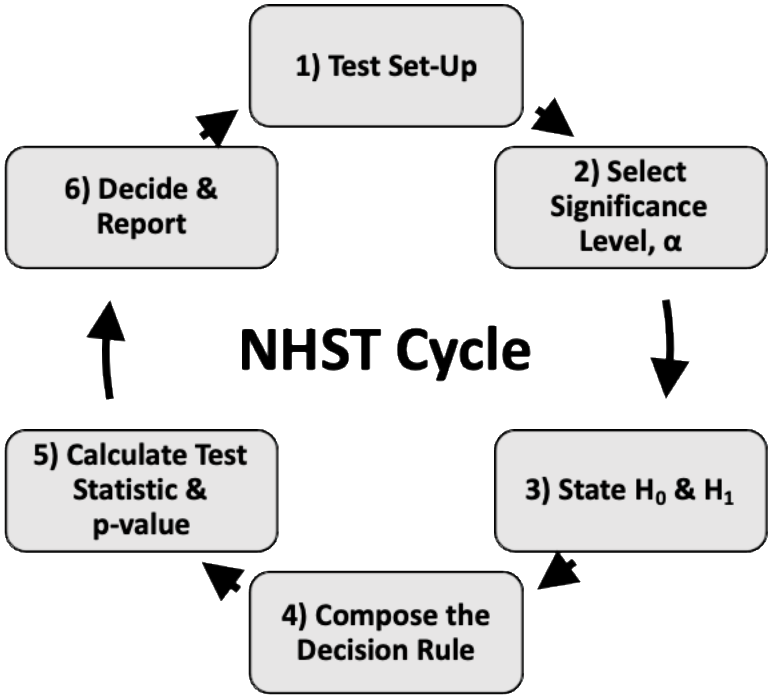



*Figure 3: NHST Cycle*

**Step 1: Test Step-Up**

As with any research project, we must state the problem in the form of a question. This will help us with the third step: Stating the null and alternate hypotheses. Once we have the *research question* articulated, we spell out the *research design*; which is to say, specify what procedures we will follow to answer the research question. A research design has six components:

1. **The Research Method:** What techniques will be used to collect the data: Secondary data, surveys, experiments, etc.

2. **The Operational Definition:** How the variables of interest will be measured.

3. **The Data Collection Process:** The way the data will be collected: Interviews, surveys, experiments, etc.

4. **Sampling Methods:** What sampling methods will be employed.

5. **Determining Sample Size:** How large a sample is needed to achieve sufficient statistical power? The sample size needs to be large enough to

uncover an effect should one actually exist. This is an issue of statistical power. Statistical power should be estimated before the data is collected. We call this an *a priori* power analysis. Conducting a power analysis after the data have been collected—*post hoc* power analysis—is generally considered less effective. Estimating the standardized effect size (ES) is an important step in determining sample size and statistical power. Effect size measures the size of the differences between the statistic and parameter without the confounding influence of sample size. In social sciences, most effect sizes are small. *Clear-Sighted Statistics* will present a preliminary introduction to the most commonly used effect sizes.

6. **Collect data**

At this stage, we may select the *test statistic*. The term test statistic means the formula for the test, or the number that results from calculating the test statistic. In the remaining modules, we will review the following test statistics:

A) One-Sample Tests
   1. z-test for the mean
   2. t-test for the mean
   3. z-test for the proportion

B) Two-Sample Tests
   1. z-test for two independent mean
   2. z-test for two independent proportions
   3. F-test for two independent variances
   4. pooled variance t-test for two independent mean
   5. unequal variance t-test for two independent means
   6. t-test for two dependent samples

C) One-Way ANOVA test for two or more independent means

D) Chi-Square Tests
   1. Goodness-of-fit test
   2. Contingency table test
   3. Test for Normality

E) Tests for Correlation and Regression

**Step 2: Select the Level of Significance, α**

We select the level of significance. The significance level is symbolized by the lower-case Greek letter alpha, α. As previously stated, the significance level is the long-term

probability of a Type I error. The significance level is used to determine whether the value of the test statistic is *statistically significant*; which is to say, the value is *not* the result of random sampling error. The null hypothesis is rejected whenever the results are statistically significant.

It is important to distinguish between *statistical significance* and *practical significance*. Practical significance refers to the magnitude of an effect. We say something has practical significance when its application changes current practices. Having statistical significance does not mean that you will have practical significance. As sample sizes increase, the hypothesis test has more power to uncover an effect. Large sample sizes can uncover miniscule effects that are statistically significant and the null hypothesis is rejected. The effect uncovered, however, may be so small that it has no practical application. In these cases, the test may be over-powered.

Closely aligned with the significance level is the *critical value*. The critical value is the value or values of the test statistic that marks the boundary between the rejection region of the probability distribution—the area where the null hypothesis is rejected—from the region where the null hypothesis is not rejected. Critical values can be z-values, t-values, F-values, or chi-square values.

**Step 3: Write the Null Hypothesis ($H_0$) and Alternate Hypothesis ($H_1$)**

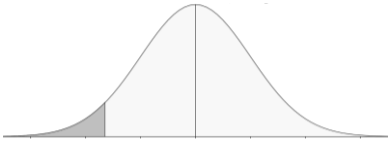
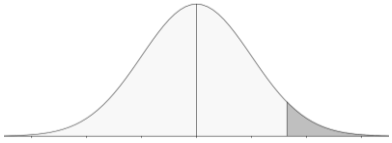
As previously stated, hypotheses are always about population parameters and there are just two mutually exclusive hypotheses: the null hypothesis and the alternate hypothesis. The null hypothesis is the hypothesis we seek to falsify. The null hypothesis is often considered a straw man the researcher seeks to reject or nullify. It states that there is no statistically significant difference or effect, The alternate hypothesis is sometimes called the

research hypothesis. The alternate hypothesis states that there is a statistically significant difference between the sample statistic and the population parameter. This means that the difference between the statistic and parameter is greater than what we would expect from sampling error.

Tests using z-values and t-values are directional. This means that the rejection region can be placed on the left or lower tail, both tails, or right or upper tail. You can tell the direction of the test by looking at the direction of the sign in the alternate hypothesis. A "less than sign," <, signals a left tail test, a "not equal sign," ≠, marks a two-tailed test. A right-tail test is used when a "greater than sign," >, is present. The rejection regions for a two-tailed test are marked by dividing the level of significance into two equal parts with half going on the right-tail and half on the left-tail. One-tailed tests have only one rejection region, which is marked by placing the entire significance level in the appropriate tail.

Drawing these curves will help you to visualize the difference between left-tailed, two-tailed, and right-tailed tests and the location of the rejection regions. Table 2 shows examples of Null and Alternate Hypotheses and the curves for left-tailed, two-tailed, and right-tailed z or t tests. The shaded areas are the rejection regions. The critical value or values are the lines that separate the rejection region from the rest of the curve.

*Table 2: Syntax for Left-Tailed, Two-Tailed, and Right-Tailed Tests*

| **Left-Tailed Test** | **Two-Tail Test** | **Right-Tailed Test** |
|---|---|---|
| **$H_0$: $\mu \geq 0$;<br>$H_1$: $\mu < 0$** | **$H_0$: $\mu = 0$;<br>$H_1$: $\mu \neq 0$** | **$H_0$: $\mu \leq 0$;<br>$H_1$: $\mu > 0$** |
|  |  |  |
| **$H_1$ contains a < sign α in left tail. α in the left tail.** | **$H_1$ contains a ≠ sign. α/2 in the left and right tails** | **$H_1$ contains a > sign. α in the right tail.** |

We will structure F-Tests and chi-square tests as right-tailed tests.

**Step 4: Write the Decision Rule**

Decision rules state the criterion for rejecting the null hypothesis, which is based on the critical value or values. The critical value is determined using the appropriate critical value tables for z, t, F, and chi-square or by using Microsoft Excel.

All decision rules follow a similar structure: "Reject the null hypothesis when [name of the test statistic] is 'less than,' 'greater than,' or 'less than or greater than' the critical value or values." Critical values may vary slightly depending upon whether you are using a printed table or Microsoft Excel. Table 3 shows the difference in z-values for left, right, and two-tailed tests using paper tables and Excel. The values used by Excel are in parentheses:

*Table 3: Critical Values for z for left, right, and two-tailed tests*

| α | Left-Tailed Test | Two-Tailed Test | Right-Tailed Test |
|---|---|---|---|
| 0.01 | -2.33 (-2.326) | -2.58 & 2.58 (-2.576 & 2.576) | 2.33 (2.326) |
| 0.05 | -1.65 (-1.645) | -1.96 & 1.96 (-1.960 & 1.960) | 1.65 (1.645) |
| 0.10 | -1.28 (-1.283) | -1.65 & 1.65 (-1.645 & 1.645) | 1.28 (1.283) |

Figure 4 shows the decision rules for left-tailed and right-tailed tests for z and t distributions using a 5 percent level of significance. For the t-test, there are 20 degrees of freedom, found by the sample size, n, minus the number of independent samples. Writing the decision rule and drawing a curve showing the rejection region, or regions, by hand will help you visualize when to reject the null hypothesis.

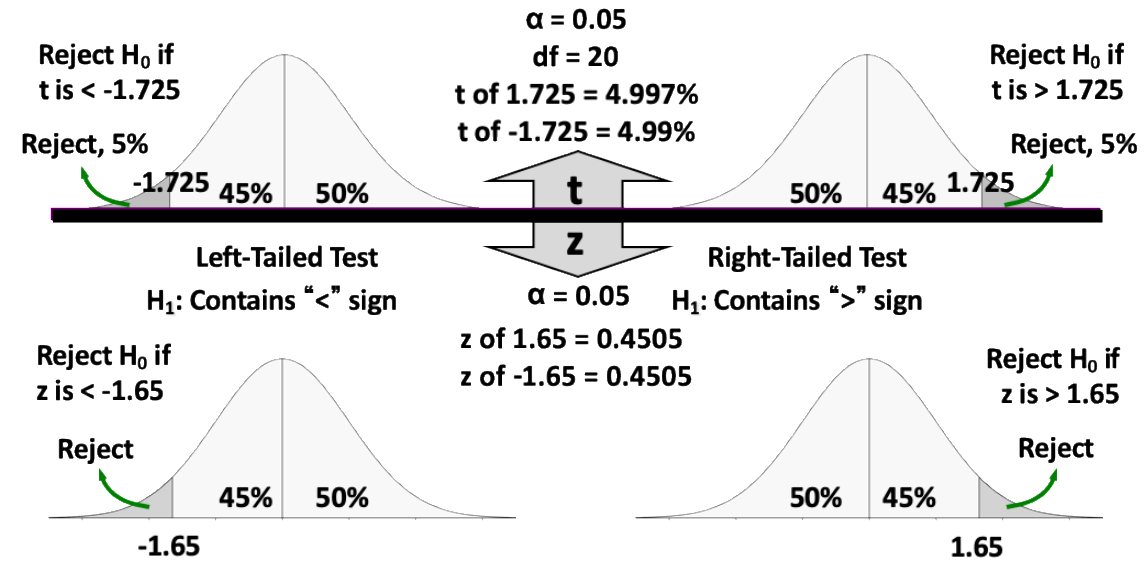



*Figure 4: Decision Rules for One-Tailed z and t tests using a 5% α*

Figure 5 shows the decision rules for two-tailed tests for z and t distributions using a 5 percent level of significance.

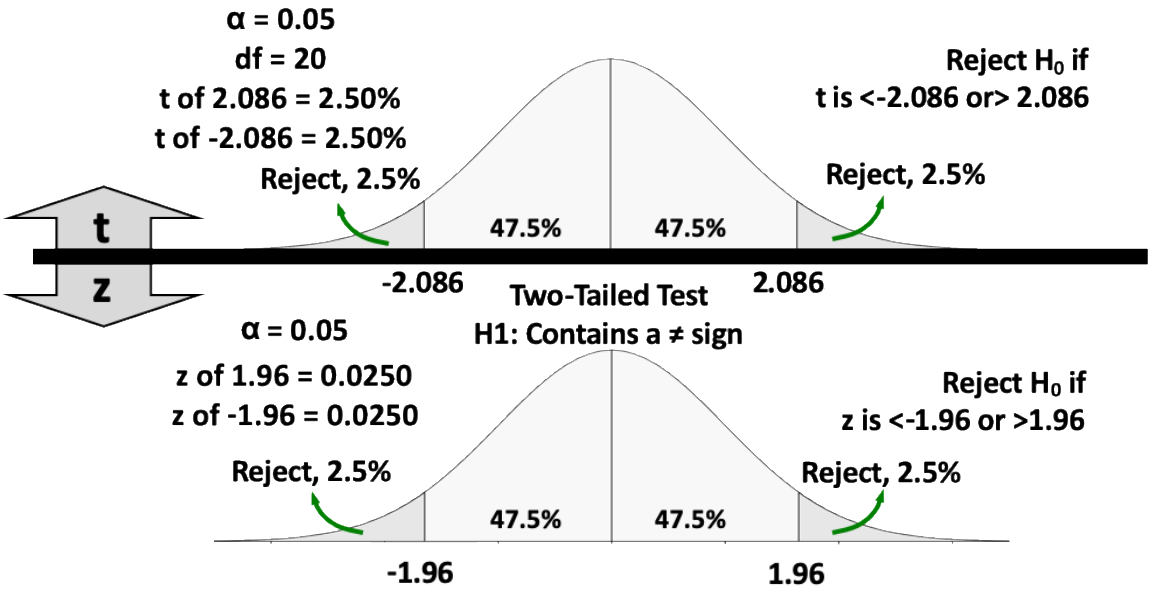



*Figure 5: Decision Rules for Two-Tailed z and t tests using a 5% α*

**Please note:** The critical values for t-tests are more extreme than the critical values for z-tests. This makes it more difficult to reject the null hypothesis for t-tests. If also means that t-tests have less statistical power than z-tests. In addition, two-tailed tests have more

extreme critical values than one-tailed tests. This also makes it harder to reject the null hypothesis for two-tailed tests, and two-tailed tests have lower statistical power than one-tailed tests.

**Step 5: Calculate the value of the test statistic**

Each null hypothesis significance test has its own test statistic or formula. For z and t tests, the test statistics are *complex fractions*. A complex fraction is a fraction where the numerator or denominator contains a fraction. The numerator measures sampling error (the sample statistic minus the population parameter): $\bar{X} - \mu$ or $p - \pi$. The denominator is the standard error of the mean, $\sigma$ or $s/\sqrt{n}$, or the standard error of the proportion, $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$.

Once we have calculated the value of the test statistic, we find the p-value using a critical value table or Microsoft Excel. The p-value, or probability value, is the probability that the result of the test statistic is due to sampling error. The p-value is a slippery concept that many people get wrong. It tells you the probability of getting a value for the test statistic that is as extreme, or more extreme, than the one you just calculated. It is a measure of how compatible your result is with the null hypothesis.

**Step 6: Make a decision regarding the $H_0$, and Report Results**

After we calculate the value of the test statistic and find the p-value, we make a decision regarding the null hypothesis and the results of the test are reported. We can make this decision on the basis of the decision rule. It is strongly recommended, however, that the decision to reject or fail to reject the null hypothesis be based on the p-value because it tells us the probability of getting a test statistic as extreme, or more extreme, than the one we found.

Here is how we interpret p-values: When the p-value is greater than the level of significance, we do not reject the null hypothesis and the higher the p-value the more confident we are in this decision. For example, a p-value of 0.051 or 0.50 would lead us to not reject the null hypothesis if the significance level were 0.05. Yet, we would be far more confident in this decision if the p-value were 0.50 than when it is only 0.051, or barely above the significance level.

When the p-value is less than or equal to the significance level, we fail to reject the null hypothesis. The smaller the p-value the more confident we are in our decision to reject the null hypothesis. At a 0.05 significance level, we reject the null hypothesis when the p-value is 0.05 or <0.001. We would not, however, have a high level of confidence in our decision to reject the null hypothesis when the p-value is 0.05, but with a p-value of <0.001 we would be very confident that our findings are statistically significant.

Figure 6 shows how to use the p-value to decide whether or not to reject the null hypothesis.

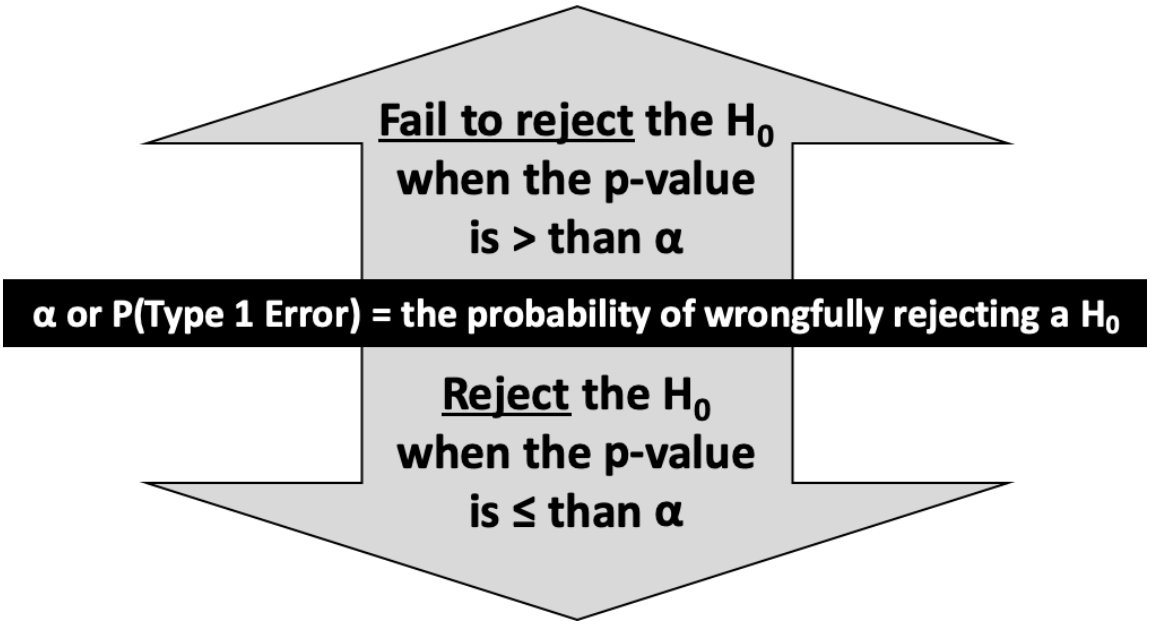



*Figure 6: How to Interpret p-values*

Over the years there has been a lot of confusion about what p-values are and how to use them. In 2016, the American Statistical Association (ASA) addressed this confusion in a statement on p-values and statistical significance.[20] The ASA listed six principles on the use of p-values:

1. P-values can indicate how incompatible the [sample] data are with a specified statistical model [or the null hypothesis].

2. P-values do not measure the probability that the studied [null] hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based solely on whether a p-value passes a specific threshold [significance level].

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. [A p-value cannot tell you whether your results have practical significance.]

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Jessica Utts, the ASA's president, concluded the ASA's statement by writing:

> The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades. But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues.[21]

We should consider p-values as a measure of how surprising our test statistic is. **But remember:** While we reject the null hypothesis when the p-value is less than or equal to the significance level, a low p-value does not tell us:

1. The alternate hypothesis is true;

2. Whether the test results have any practical significance.

**Let's review:** You may consider the NHST process a cycle. The process can require multiple testing for continuous refinement of hypotheses through the process of falsification. It is through this process that a hypothesis may eventually become a theory. The six steps are:

1. Test set-up
2. Select the level of significance, α
3. State the Null and Alternate Hypotheses
4. Compose the decision rule
5. Calculate the test statistic and p-value
6. Decide on whether or not to reject a null hypothesis and report the results

Figure 7 shows a graphic representation of the six steps in the process.

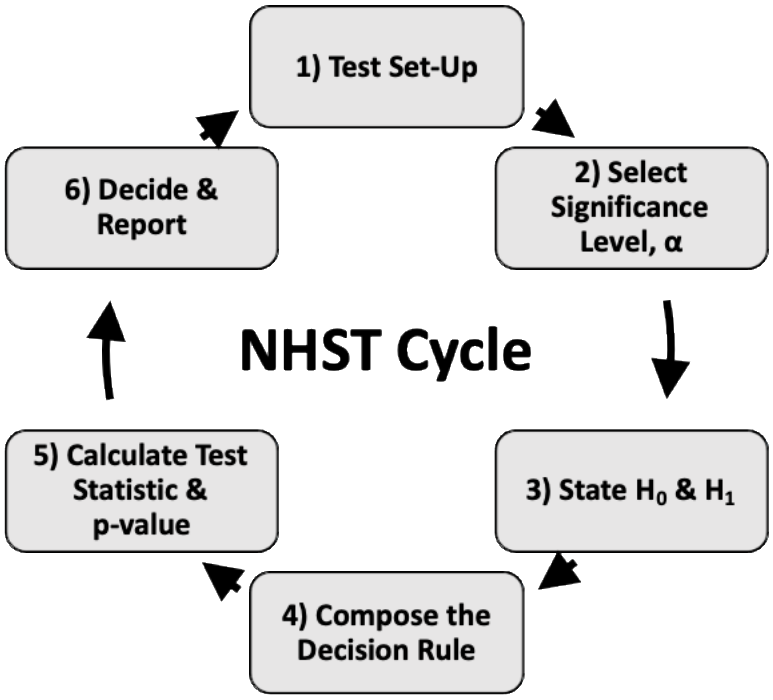



*Figure 7: The NHST Cycle*

We will use this six-step process whenever we conduct a NHST. Throughout *Clear-Sighted Statistics*, we will assume that the first step, test set-up—setting up the research design—

has been properly conducted. We will also typically conduct either an *a priori* or *post hoc* power analysis. The issue of practical significance will be discussed when appropriate.

**VII. Summary**

NHST is a process based on probability and sample statistics to determine whether the difference between the statistic and the parameter results from random sampling error. Only the null hypothesis is subject to falsification. Regarding the significance level, the researcher sets the long-term tolerance for mistakenly rejecting the null hypothesis. When the p-value is less than or equal to the significance level, the null hypothesis is rejected and the results are considered statistically significant. The p-value indicates the probability of making a Type I error, given the test results. The smaller the p-value, the more likely the results are statistically significant; which is to say, the difference between the sample statistic and population parameter is *not* due to random sampling error.

When we fail to reject the null hypothesis, we do not conclude that it is true. We also must be aware that we could be committing a Type II error. The goal of good hypothesis testing is to have sufficient statistical power. Generally speaking, we aim for at least 80 percent statistical power, or the probability of a Type II error of 20 percent or less. Type I errors are considered a more serious mistake than Type II errors, which is why researchers set the significance level, or tolerance for a Type I error, at a lower level than their tolerance for a Type II error.

**VIII. What is Next**

We have reviewed the essential features of NHST, defined key terms, and outlined the steps to conduct such tests. In Module 14, we will cover one-sample tests of hypothesis using z and t distributions. In Module 15, we will explore two sample tests of hypothesis using z

and t. We will distinguish between independent and dependent or conditional samples. We will also introduce F-distributions to determine equality of variance between two independent samples. In Module 16, the One-Way ANOVA test will be introduced. This test allows you to simultaneously compare two or more population means. This test uses the F-distribution. In Module 17, chi-square tests, the only nonparametric tests we will cover, will be reviewed. Unlike parametric tests, nonparametric tests make no assumption about the parameters in the population under investigation. Chi-square tests use the chi-square distribution. Module 18, covers linear correlation and regression. We will examine a variety of hypothesis tests used with linear correlation and regression. Figure 8 shows the type of null hypothesis significance tests we will cover in *Clear-Sighted Statistics*.

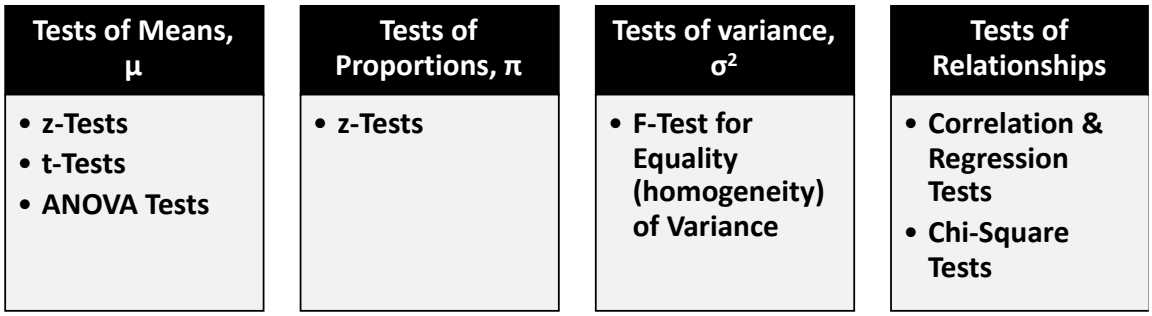



*Figure 8: Types of NHST*

**IX. Exercises**

Answers to the following questions can be found by carefully reading this module.

**Exercise 1:** What is the meaning of the word null in null hypothesis ($H_0$)?

**Exercise 2:** What is a statistical significance level, α?

**Exercise 3:** What does the term "statistically significant" mean?

**Exercise 4:** What is the difference between Type I (α) and Type II (β) errors?

**Exercise 5:** What affects the probability of Type I (α) and Type II (β) errors?

**Exercise 6:** Which errors are considered more serious? Type I or Type II errors?

**Exercise 7:** What is Effect Size (ES)?

**Exercise 8:** What is practical or clinical significance and why is it important?

**Exercise 9:** How are Type II errors and Statistical Power related and why are low powered tests a problem?

**Exercise 10:** a test be over-powered? If so, how would one know?

**Exercise 11:** What are p-values?

**Exercise 12:** P-values are widely misused. What are the American Statistical Association's key guidelines on p-values?

---

[1] Karl R. Popper, *The Logic of Scientific Discovery*, (Mansfield Centre, CT: Martino Publishing, 2014), pp. 40-41.

[2] "Dressed to Kill," *Universal Pictures*, 1946, 15:42-15:48. This movie is available on Amazon Prime.

[3] Ronald A. Fisher, *Statistical Methods for Research Workers.* (London: Oliver & Boyd, 1925). The 14th and final edition was published in 1970, 8 years after Fisher's death.

[4] Jerzy Neyman, and Ego S. Pearson, E. S. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I." *Biometrika* 20A, 1928, pp. 175–240.

[5] E. L. Lehmann, "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, Volume 88, No. 424, December 1993, p. 1248. Steven N. Goodman, "P Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate." *American Journal of Epidemiology*, Vol. 137, No. 5. 1993, pp. 485-496.

[6] Geoff Cumming, *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. (New York: Routledge, 2012), p. 25.

[7] *Genesis*, 2:7.

[8] Galton made many contributions in a wide variety of areas. He created the concept of regression toward the mean. He was among the first to use standard deviation. He promoted the use of fingerprints to identify individuals. He devised the first weather map. He was, like Karl Pearson, an early advocate of eugenics, or the idea that the human race can be improved by getting more fit people to have more children and unfit people to have fewer children. Eugenics was exposed by many people across the political spectrum in the late nineteenth and early twentieth centuries. The eugenics movement inspired the forced sterilization programs in Nazi Germany. These programs lead to the demise of this movement.

[9] "The Scope of Biometrika," *Biometrika*, Volume 1, No. 1, October 1901, pp. 1-2.

[10] "The Spirit of Biometrika, *"Biometrika*, Volume 1, No. 1, October 1901, p. 3.

[11] Thomas S. Kuhn, *The Structure of Scientific Revolutions*, (Chicago: University of Chicago Press, 2012).

[12] Tim M. Berra, "Charles Darwin's Paradigm Shift," *The Beagle, Records of the Museums and Art Galleries of the Northern Territory,* Volume 5, 2008, pp. 1-5.

[13] Thomas S. Kuhn, The Structure of Scientific Revolutions. (Chicago: University of Chicago Press, 2012).

[14] John W. Tukey, *Exploratory Data Analysis*, (Reading, MA: Addison-Wesley, 1977), pp. 1-3, 21.

[15] http://www.nycourts.gov/judges/cji/5-SampleCharges/SampleCharges.shtml. The phase "beyond reasonable doubt is a "term of art" that is difficult to define. In New York State, a statement like this is read to juries before they decide to convict or acquit a defendant:

> We now turn to the fundamental principles of our law that apply in all criminal trials– the presumption of innocence, the burden of proof, and the requirement of proof beyond a reasonable doubt.
>
> Throughout these proceedings, the defendant is presumed to be innocent. As a result, you must find the defendant not guilty, unless, on the evidence presented at this trial, you conclude that the People [who are represented by the prosecutor] have proven the defendant guilty beyond a reasonable doubt.
>
> What does our law mean when it requires proof of guilt "beyond a reasonable doubt"?
>
> The law uses the term, "proof beyond a reasonable doubt," to tell you how convincing the evidence of guilt must be to permit a verdict of guilty. The law recognizes that, in dealing with human affairs, there are very few things in this world that we know with absolute certainty. Therefore, the law does not require the People to prove a defendant guilty beyond all possible doubt. On the other hand, it is not sufficient to prove that the defendant is probably guilty. In a criminal case, the proof of guilt must be stronger than that. It must be beyond a reasonable doubt.
>
> A reasonable doubt is an honest doubt of the defendant's guilt for which a reason exists based upon the nature and quality of the evidence. It is an actual doubt, not an imaginary doubt. It is a doubt that a reasonable person, acting in a matter of this importance, would be likely to entertain because of the evidence that was presented or because of the lack of convincing evidence.
>
> Proof of guilt beyond a reasonable doubt is proof that leaves you so firmly convinced of the defendant's guilt that you have no reasonable doubt of the existence of any element of the crime or of the defendant's identity as the person who committed the crime.

[16] Ronald Aylmer Fisher, *The Design of Experiments*, (Edinburgh, UK: Oliver and Boyd, 1935), p. 19. https://archive.org/details/in.ernet.dli.2015.502684/page/n31.

[17] Søren Kierkegaard, *Works of Love: Some Christian Reflections in Form of Discourse*. Translated by Howard V. and Enda H, Hong, (New York: Harper Torchbooks, 1962), p. 23.

[18] John P. A. Ioannidis, "Why Most Published Research Findings are False," *PLoS Medicine*, August 2005, pp. 696-701. https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.0020124&type=printable

[19] Joan Welkowitz, Robert B. Ewen, Jacob Cohen, *Introductory Statistics for Behavioral Sciences, Fourth Edition.* (New York: Harcourt, Brace, Jovanovich, 1986). Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Second Edition. (New York: Psychology Press, 1988).

[20] "American Statistical Association Releases Statement on Statistical Significance and *P*-Values," March 7, 2016. https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf.

[21] "American Statistical Association Releases Statement on Statistical Significance and *P*-Values," March 7, 2016. https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf.