6: Introduction to Null Hypothesis Significance Testing

Acronyms and symbols

P	P value
p	binomial parameter "probability of success"
n	sample size
H_0	the null hypothesis
H_a	the alternative hypothesis

P value

Statistical inference is the act of generalizing from sample (the data) to a larger phenomenon (the population) with calculated degree of certainty. The prior chapter introduced the most important form of inference: estimation. This chapter introduces the second form of inference: null hypothesis significance tests (NHST), or "hypothesis testing" for short.

The main statistical end product of NHST is the *P* value, which is the most commonly encountered inferential statistic and most frequently misunderstood, misinterpreted, and misconstrued statistics in the biomedical and public health literature.[1] Most teachers of statistics do not fully understand *P* values. Not even specialist scientists can easily explain them.

Since the process of NHST revolves around the *P* value, let us start with its definition, which is easiest to remember with this notation:

$$P \text{ value} \equiv \Pr(\text{data or data more extreme} \mid H_0 \text{ true})$$

where

Pr ≡ probability
| ≡ "given" or "conditional upon"
H_0 ≡ the null hypothesis

Thus, the *P* value answers the question "If the null hypothesis were true, what is the probability of observing the current data or data that is more extreme?" Note that the *P* value is NOT the probability that the hypothesis (or any other hypothesis) is right or wrong. In fact, it assumes the null hypothesis is right!

In light of these facts, there are actually two classical schools of thought on how best to use the *P* value: the Fisher and Neymann-Pearson schools.[2] There is also a Bayesian way to interpret the *P* value, but that presents a whole other set of dilemmas.

As a starting point, we will consider the *P* value as a calculated index which, as it gets smaller-and-smaller, provides stronger-and-stronger evidence against the null hypothesis.

[1] There is a large body of literature about the misinterpretation of *P* values. My favorite is: Cohen J. (1994). The earth is round (p < .05). *American Psychologist*, 49, 997-1002.

[2] For an introduction to the distinct between these interpretations, see this video. You can get to the video by Googling "Gerstman p value Youtube."

Example of a NHST

The first step of NHST is to convert the **research question** into null and alterative hypotheses. Thus, the research question must be concisely articulated before starting this process.

- The **null hypothesis (H_0)** is a statement of “no difference,” “no association,” or “no treatment effect.”
- The **alternative hypothesis, H_a** is a statement of “difference,” “association,” or “treatment effect.”

H_0 is assumed to be true until proven otherwise. However, H_a is the hypothesis the researcher hopes to bolster.

Take as an example a treatment that is said to be 25% effective. A researcher claims she has a new treatment with improved efficacy. It is essential that we *articulate* the research question into “plain English.” In null (no difference) form, the new treatment is NO more effective than the existing treatment (25% effective). In alternative (“difference”) form, the new treatment is more effective than the existing treatment.

We can see that this question is about a proportion (25% vs. not 25%). Thus, the parameter to be inferred is similar to binomial proportion p. Under the null hypothesis, H_0: $p = .25$. This is the most important part of setting up the NHST.

The researcher tests the new treatment in 3 patients. This experiment lends itself to the binomial distribution since it is based on series of trials which can each outcome can be characterized as a success or failure. *Since the null hypothesis of “no difference" is assumed to be true* until proven otherwise, the number of successes in the experiment should follow a binomial pmf with $n = 3$ and $p = 0.25$. This exact pmf was introduced in Chapter 4 and is also shown here:

X **Number of successes**	**Pr(*X* = *x*)** **Probability**	**Pr(*X* ≤ *x*)** **Cumulative Probability**
0	0.4219	0.4219
1	0.4219	0.8438
2	0.1406	0.9844
3	0.0156	1.0000

We consider *two possible* outcomes of the experiment.

If all 3 patients in this experiment responded to the new treatment, P value = (data or data more extreme| H_0 true) = Pr(X=3) = 0.0156. This observation would be rare if the true success rate was only 25%. Thus, the evidence against H_0 would be deemed to be significant.

If 2 of the 3 patients in the experiment responded, P value = (data or data more extreme| H_0 true) = Pr(X = 2) or Pr(X=3) = 0.1406 + 0.0156 = 0.1562. In this “2 out of 3” case, the P value is 0.1562 indicates that this would not be unusual if the probability of success was actually 0.25. Thus, the evidence against H_0 is deemed non-significant.

In neither case can we say that the evidence of efficacy is conclusive. However, the “3 out of 3” (P = .0156) provides some evidence of a “real difference” and is worthwhile of follow-up, where the “2 out of 3” (P = .1562) evidence is weak.

The Exact Binomial Test

The exact binomial test is suited to test a binomial proportion from a single sample. A step-by-step analysis of the exact binomial test is presented.

Step 1. Review the research question and identify the null hypothesis. Read the research question. Verify that we have a single sample that addresses a binomial proportion. Identify the value of binomial parameter *p* when there is truly "no difference." Write the null hypothesis in this form:

$$H_0: p = \text{the value of } p \text{ if } H_0 \text{ is true}$$

Calculate the sample proportion $(\hat{p})$ to see how much it differs from the value proposed by the null hypothesis.

Step 2. In lieu of a test statistic, determine the binomial pmf that applies under H_0. Since this test is based on exact probabilities, there is no test statistic *per se*. Instead, list the pmf that applies when H_0 is true.

$$\text{When } H_0 \text{ is true, } X\sim\text{b}(n, p)$$

where *n* is the sample size and *p* is the assume value of *p* when H_0 is true.

Step 3: Determine the *P* value. The *P* value is the probability of observing the data or data more extreme. When we are looking for an increase in the number of successes *P* value = $\Pr(X \geq x)$ where *x* is the observed number of successes. When we are looking for an increase in the number of successes *P* value = $\Pr(X \leq x)$.

Step 4: Interpret results in narrative form. Note the sample proportion, direction of the observed difference (increase or decrease), and P value. When the *P* value is small (say, less than .10), the evidence against the null hypothesis cannot easily be explained by chance ("statistical significance").

Illustration

Suppose a treatment has an expected success rate of 0.25. We observe successful treatment in 2 out of 3 patients. Is this observation worthy of note, i.e., is it statistically significant?

Step 1. H_0: $p = .25$. Note that $\hat{p} = \frac{2}{3} = .6667$.

Step 2. Under H_0, $X\sim\text{b}(3,.25)$

X	**Pr(*X* = *x*)**
0	0.4219
1	0.4219
2	0.1406
3	0.0156

Step 3. $P = \Pr(X \geq 2) = \Pr(X = 2) \text{ or } \Pr(X=3) = 0.1406 + 0.0156 = 0.1562$

Step 4: The difference between the observed proportion (2 of 3) and the expected proportion under the null hypothesis is explicable by chance, i.e., not statistically significant (P = .1562).

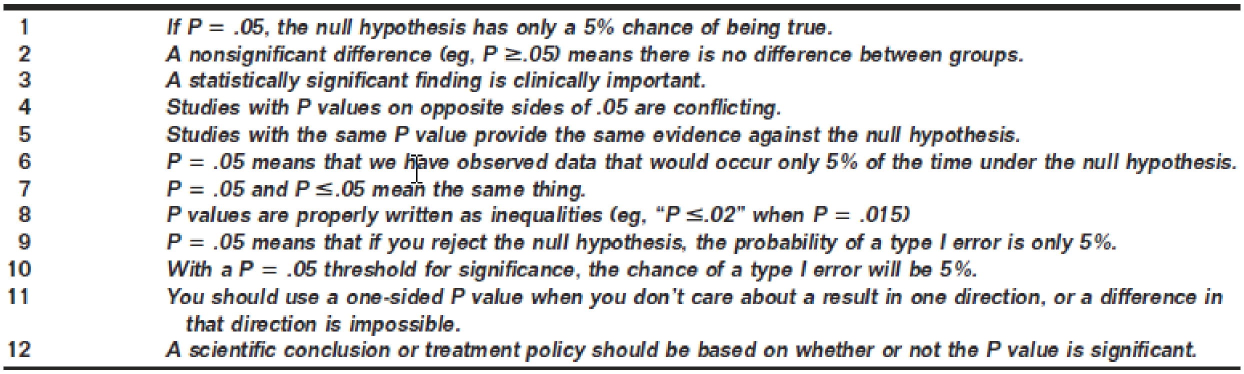
Misconceptions about *P* values

The interpretation of *P* values is a minefield. It is therefore very important to start out with the proper understanding. As the movie “The Big Short” noted, “It ain’t what you don't know that gets you into trouble. It's what you know for sure that just ain’t so” (falsely attributed to Mark Twain).[3]

“It is not the fault of researchers that the *P* value is difficult to interpret correctly. The man who introduced it as a formal research tool, the statistician and geneticist R.A. Fisher, could not explain exactly its inferential meaning. He proposed a rather informal system that could be used, but he never could describe straightforwardly what it meant from an inferential standpoint. In Fisher’s system, the *P* value was to be used as a rough numerical guide of the strength of evidence against the null hypothesis.”[4]

One common *mistake* in using the *P* value is to declare a result as “significant” if the *P* value is less than .05. We want to avoid this common and costly mistake. In addition, here are twelve additional **misconceptions** of the *P* values that we wish to avoid (Goodman 2008, footnote 5):

Table 1 Twelve *P*-Value Misconceptions



1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, “P ≤.02” when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don’t care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Future sections of this chapter will introduce the one-sample z test for a mean and the one-sample z test for a proportion.

[3] An actual quote from Tolstoy (1893): “The most difficult subjects can be explained to the most slow-witted man if he has not formed any idea of them already; but the simplest thing cannot be made clear to the most intelligent man if he is firmly persuaded that he knows already, without a shadow of a doubt, what is laid before him.”

[4] Goodman S. A (2008) Dirty Dozen: Twelve *P*-Value Misconceptions. Seminars in Hematology, 45(3), 135-40.