

8.1: The null and alternative hypotheses

Introduction

Classical statistical parametric tests — **t-tests** (one sample t-test, independent sample-t-test), **analysis of variance (ANOVA)**, **correlation**, and **linear regression**— and nonparametric tests like χ^2 (**chi-square: goodness of fit and contingency table**), share several features that we need to understand. It's natural to see all the details as if they are specific to each test, but there's a theme that binds all of the classical statistical inference in order to make claim of "statistical significance."

- a calculated **test statistic**
- **degrees of freedom** associated with the calculation of the test statistic
- a **probability value** or **p-value** which is associated with the test statistic, assuming a **null hypothesis** is "true" in the population from which we sample.
 - Note that as discussed in (Chapter 8.2), this is not strictly the interpretation of p-value, but a shorthand for how likely the data is to fit the null hypothesis. P-value alone can't tell us about "truth."
- in the event we reject the null hypothesis, we provisionally accept the **alternative hypothesis**.

Statistical Inference in the NHST Framework

By inference, we mean to imply some formal process by which a conclusion is reached from data analysis of outcomes of an experiment. The process at its best leads to conclusions based on evidence. In statistics, evidence comes about from the careful and reasoned application of statistical procedures and the evaluation of probability (Abelson 1995).

Formally, statistics is rich in inference process. We begin by defining the **classical frequentist**, aka Neyman-Pearson approach, to inference, which involves the pairing of two kinds of statistical hypotheses: the null hypothesis (H_0) and the alternate hypothesis (H_A). Whether we accept the null hypothesis or not is evaluated against a decision criterion, a fixed **statistical significance level** (Lehmann 1992). Significance level refers to the setting of a **p-value threshold** before testing is done. The threshold is often set to Type I error of 5% (Cowles & Davis 1982), but researchers should always consider whether this threshold is appropriate for their work (Benjamin et al 2017).

This inference process is referred to as Null Hypothesis Significance Testing, NHST. Additionally, a probability value will be obtained for the test outcome or test statistic value. In the Fisherian **likelihood** tradition, the magnitude of this statistic value can be associated with a probability value, the p-value, of how likely the result is given that the null hypothesis is "true". (Again, keep in mind that this is not strictly the interpretation of p-value, it's a shorthand for how likely the data is to fit the null hypothesis. P-value alone can't tell us about "truth", per our discussion in Chapter 8.2.)

Note:

About **-logP**. P-values are traditionally reported as a decimal, like 0.000134, in the **closed (set) interval** (0,1) — p-values can never be exactly zero or one. The smaller the value, the less the chance our data agree with the null prediction. Small numbers like this can be confusing, particularly if many p-values are reported, like in many genomics works, e.g., GWAS studies. Instead of reporting vanishingly small p-values, studies may report the **negative log₁₀ p-value**, or **-logP**. Instead of small decimal numbers, large numbers are reported; the larger, the more chance our data is against the null hypothesis. Thus, our p-value becomes 3.87 -logP.

R code

```
-1*log(0.000134, 10)
[1] 3.872895
```

Why log₁₀ and not some other base transform? Just that log₁₀ is convenient — powers of 10.

The **antilog** of 3.87 returns our p-value:

```
> 10^(-1*3.872895)
[1] 0.0001340001
```

For convenience, here is a partial p-value -logP transform table.

P-value	-logP
0.1	1
0.01	2
0.001	3
0.0001	4

On your own, complete the table for -logP values of 5 through 10. See [Question 7 below](#).

NHST Workflow

We presented in the introduction to Chapter 8 without discussion a simple flow chart to illustrate the process of decision. Here, we repeat the flow chart diagram and follow with descriptions of the elements.

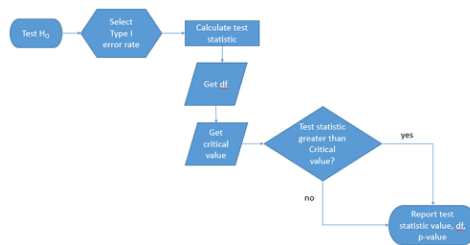


Figure 8.1.1: Flow chart of inductive statistical reasoning.

What's missing from the flow chart is the very necessary caveat that interpretation of the null hypothesis is associated with two kinds of error, Type I error and Type II error. These points and others are discussed in the following sections.

We start with the hypothesis statements. For illustration we discuss hypotheses in terms of comparisons involving just two groups, also called **two-sample tests**. **One-sample tests**, in contrast, refer to scenarios where you compare a sample statistic to a population value. Extending these concepts to more than two samples is straightforward, but we leave that discussion to Chapters 12 – 18.

Null hypothesis

By far the most common application of the null hypothesis testing paradigm involves the comparisons of different treatment groups on some outcome variable. These kinds of null hypotheses are the subject of Chapters 8 through 12.

The **Null hypothesis** (H_0) is a statement about the comparisons, e.g., between a sample statistic and the population, or between two treatment groups. The former is referred to as a **one-tailed test** whereas the latter is called a **two-tailed test**. The null hypothesis is typically “no statistical difference” between the comparisons.

For example, a one-sample, two-tailed null hypothesis.

$$H_0 : \bar{X} = \mu$$

and we read it as “there is no statistical difference between our sample mean and the population mean.” For the more likely case in which no population mean is available, we provide another example, a two-sample, two-tailed null hypothesis:

$$H_0 : \bar{X}_1 = \bar{X}_2$$

Here, we read the statement as “there is no difference between our two sample means.” Equivalently, we interpret the statement as both sample means estimate the same population mean.

$$H_0 : \bar{X}_1 = \bar{X}_2 = \mu$$

Under the **Neyman-Pearson approach** to inference we have two hypotheses: the null hypothesis and the alternate hypothesis. The null hypothesis was defined above.

Note:

Tails of a test are discussed further in chapter 8.4.

Alternative hypothesis

Alternative hypothesis (H_A): If we conclude that the null hypothesis is false, or rather and more precisely, we find that *we provisionally fail to reject the null hypothesis*, then we *provisionally accept the alternative hypothesis*. The view then is that something other than random chance has influenced the sample observations. Note that the pairing of null and alternative hypotheses covers all possible outcomes. We do not, however, say that we have evidence for the alternative hypothesis under this statistical regimen (Abelson 1995). We tested the null hypothesis, not the alternative hypothesis. Thus, it is incorrect to write that, having found a statistical difference between two drug treatments, say aspirin and acetaminophen for relief of migraine symptoms, it is not correct to conclude that we have proven the case that acetaminophen improves improves symptoms of migraine sufferers.

For the one-sample, two-tailed null hypothesis, the alternative hypothesis is

$$H_A : \bar{X} \neq \mu$$

and we read it as “there is a statistical difference between our sample mean and the population mean.” For the two-sample, two-tailed null hypothesis, the alternative hypothesis would be

$$H_A : \bar{X}_1 \neq \bar{X}_2$$

and we read it as “there is a statistical difference between our two sample means.”

Alternative hypothesis often may be the research hypothesis

It may be helpful to distinguish between technical hypotheses, scientific hypothesis, or the equality of different kinds of treatments. Tests of technical hypotheses include the testing of statistical assumptions like **normality assumption** (see Chapter 13.3) and **homogeneity of variances** (Chapter 13.4). The results of inferences about technical hypotheses are used by the statistician to justify selection of parametric statistical tests (Chapter 13). The testing of some scientific hypothesis like whether or not there is a positive link between lifespan and insulin-like growth factor levels in humans (Fontana et al 2008), like the link between lifespan and IGFs in other organisms (Holtzenberger et al 2003), can be further advanced by considering multiple hypotheses and a test of nested hypotheses and evaluated either in Bayesian or likelihood approaches (Chapter 16 and Chapter 17).

How to interpret the results of a statistical test

Any number of statistical tests may be used to calculate the value of the **test statistic**. For example, a one-sample t-test may be used to evaluate the difference between the sample mean and the population mean (Chapter 8.5) or the independent sample t-test may be used to evaluate the difference between means of the control group and the treatment group (Chapter 10). The test statistic is the particular value of the outcome of our evaluation of the hypothesis and it is associated with the p-value. In other words, given the assumption of a particular probability distribution, in this case the t-distribution, we can associate a probability, the p-value, that we observed the particular value of the test statistic and the null hypothesis is true in the reference population.

By convention, we determine **statistical significance** (Cox 1982; Whitley & Ball 2002) by assigning ahead of time a decision probability called the **Type I error rate**, often given the symbol α (alpha). The practice is to look up the **critical value** that corresponds to the outcome of the test with degrees of freedom like your experiment and at the Type I error rate that you selected. The **Degrees of Freedom** (DF , df , or sometimes noted by the symbol ν), are the number of independent pieces of information available to you. Knowing the degrees of freedom is a crucial piece of information for making the correct tests. Each statistical test has a specific formula for obtaining the independent information available for the statistical test. We first were introduced to DF when we calculated the sample variance with the **Bessel correction**, $n - 1$, instead of dividing through by n . With df in hand, the value of the test statistic is compared to the critical value for our null hypothesis. If the test statistic is smaller than the critical value, we fail to reject the null hypothesis. If, however, the test statistic is greater than the critical value, then we provisionally reject the null hypothesis. This critical value comes from a probability distribution appropriate for the kind of sampling and properties of the measurement we are using. In other words, the rejection criterion for the null hypothesis is set to a critical value, which corresponds to a known probability, the Type I error rate.

Before proceeding with yet another interpretation, and hopefully a less technical discussion about test statistics and critical values, we need to discuss the two types of statistical errors. The Type I error rate is the statistical error assigned to the probability that we may reject a null hypothesis as a result of our evaluation of our data when in fact in the reference population, the null hypothesis is, in fact, true. In Biology we generally use Type I error $\alpha = 0.05$ level of significance. We say that the probability of obtaining the observed value AND H_0 is true is 1 in 20 (5%) if $\alpha = 0.05$. Put another way, we are willing to reject the Null Hypothesis when there is only a 5% chance that the observations could occur and the Null hypothesis is still true. Our test statistic is associated with the p-value; the critical value is associated with the Type I error rate. If and only if the test statistic value equals the critical value will the p-value equal the Type I error rate.

The second error type associated with hypothesis testing is β , the **Type II statistical error rate**. This is the case where we accept or fail to reject a null hypothesis based on our data, but in the reference population, the situation is that indeed, the null hypothesis is actually false.

Thus, we end with a concept that may take you a while to come to terms with — there are four, not two possible outcomes of an experiment.

Outcomes of an experiment

What are the possible outcomes of a comparative experiment? We have two treatments: one in which subjects are given a treatment and the other, in which subjects receive a placebo. Subjects are followed and an outcome is measured. We calculate the descriptive statistics aka summary statistics, means, standard deviations, and perhaps other statistics, and then ask whether there is a difference between the statistics for the groups. So, two possible outcomes of the experiment, correct? If the treatment has no effect, then we would expect the two groups to have roughly the same values for means, etc., in other words, any difference between the groups is due to chance fluctuations in the measurements and not because of any systematic effect due to the treatment received. Conversely, then if there is a difference due to the treatment, we expect to see a large enough difference in the statistics so that we would notice the systematic effect due to the treatment.

Actually, there are four, not two, possible outcomes of an experiment, just as there were four and not two conclusions about the results of a clinical assay. The four possible outcomes of a test of a statistical null hypothesis are illustrated in Table 8.1.1.

Table 8.1.1. When conducting hypothesis testing, four outcomes are possible.

		H_0 in the population	
		True	False
Result of statistical test	Reject H_0	Type I error with probability equal to α (alpha)	Correct decision, with probability equal to $1 - \beta$ (1 - beta)
	Fail to reject the H_0	Correct decision with probability equal to $1 - \alpha$ (1 - alpha)	Type II error with probability equal to β (beta)

In the actual population, a thing happens or it doesn't. The null hypothesis is either true or it is not. But we don't have access to the reference population, we don't have a census. In other words, there is truth, but we don't have access to the truth. We can weight, assigned as a probability or p-value, our decisions by how likely our results are given the assumption that the truth is indeed "no difference."

If you recall, we've seen a table like Table 8.1.1 before in our discussion of conditional probability and risk analysis (Chapter 7.3). We made the point that statistical inference and the interpretation of clinical tests are similar (Browner and Newman 1987). From the perspective of ordering a **diagnostic test**, the proper null hypothesis would be that the patient does not have the disease. For your review, here's that table (Table 8.1.2).

Table 8.1.2. Interpretations of results of a diagnostic or clinical test.

		Does the person have the disease?	
		Yes	No
Result of the diagnostic test	Positive	Sensitivity of the test (a)	False positive (b)
	Negative	False negative (c)	Specificity of the test (d)

Thus, a positive diagnostic test result is interpreted as rejecting the null hypothesis. If the person actually does not have the disease, then the positive diagnostic test is a false positive.

Questions

1. Match the corresponding entries in the two tables. For example, which outcome from the inference/hypothesis table matches *specificity of the test*?
2. Find three sources on the web for definitions of the p-value. Write out these definitions in your notes and compare them.
3. In your own words distinguish between the test statistic and the critical value.
4. Can the p-value associated with the test statistic ever be zero? Explain.
5. Since the p-value is associated with the test statistic and the null hypothesis is true, what value must the p-value be for us to provisionally reject the null hypothesis?
6. All of our discussions have been about testing the null hypothesis, about accepting or rejecting, provisionally, the null hypothesis. If we reject the null hypothesis, can we say that we have evidence for the alternate hypothesis?
7. What are the p-values for $-\log P$ of 5, 6, 7, 8, 9, and 10? Complete the p-value $-\log P$ transform table.
8. Instead of \log_{10} transform, create a similar table but for negative natural log transform. Which is more convenient? Hint: `log(x, base=exp(1))`

This page titled [8.1: The null and alternative hypotheses](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by [Michael R Dohm](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.