

Lab 6: More Hypothesis Testing - Classical Approach

Objectives:

1. Understand how to perform hypothesis tests for means (one population and two populations) using the classical approach.
2. Understand how to use the `t.test()` function in R to calculate P -values

Definitions:

- hypothesis testing
- null vs. alternative hypothesis
- left- vs. right- vs. two-tail test
- test statistic
- P -value
- statistical significance
- t -test
- matched pairs vs. independent samples

Introduction:

In Lab 2, we introduced hypothesis testing, a formal procedure for testing the validity of a claim about a population or populations. Specifically, we developed a procedure called *permutation testing* in the context of testing hypotheses about two population means. Permutation testing does not make any assumptions about the distributions of the populations involved in the hypotheses. In this lab, we consider the classical approach to hypothesis testing, where now we will make assumptions about the distribution of the population or at least use a probability distribution to compute (approximate) P -values. This builds on the work we did in Labs 4 & 5, where we used either the standard normal or t distributions.

We will still use the same framework for performing a hypothesis test that we established in Lab 2. Namely, we compute a test statistic from the data and then a corresponding P -value that tells us the probability of getting a value as extreme as or more extreme than the observed test statistic *assuming the null hypothesis is true*. The smaller the P -value, the more evidence we have against the null hypothesis, because the observed result cannot be easily explained by chance alone.

Activities:

Getting Organized: *If you are already organized, and remember the basic protocol from previous labs, you can skip this section.*

Navigate to your class folder structure. Within your "Labs" folder make a subfolder called "Lab6". Next, download the lab notebook .Rmd file for this lab from Blackboard and save it in your "Lab6" folder. You will be working with the SAT and NCBI rhts2004 data sets on this lab. You should download the data files into your "Lab6" folder from Blackboard.

Within RStudio, navigate to your "Lab6" folder via the file browser in the lower right pane and then click "More > Set as working directory". Get set to write your observations and R commands in an R Markdown file by opening the "lab6_notebook.Rmd" file in RStudio. Remember to add your name as the author in line 3 of the document. For this lab, enter all of your commands into code chunks in the lab notebook. You can still experiment with code in an R script, if you want. To set up an R Script in RStudio, in the upper left corner click "File > New File > R script". A new tab should open up in the upper left pane of RStudio.

Pause for Reflection #1:

In Exercise #1 from Tuesday's class, we tested the claim that the mean dissolved oxygen content μ in a certain stream is less than 5 mg per liter based on a sample of 45 specimens with a mean of 4.62 mg/l. We assumed that dissolved oxygen content varies among locations in the stream according to a normal distribution with standard deviation $\sigma = 0.92$ mg, and so calculated the P -value as follows:

$$P\text{-value} = P(\bar{X} \leq 4.62 \mid \mu = 5) = P\left(\frac{\bar{X} - 5}{0.92/\sqrt{45}} \leq \frac{4.62 - 5}{0.92/\sqrt{45}}\right) = P(Z \leq -2.77) = 0.0028$$

Comment on why we calculated the probability that a sample mean \bar{X} would be "less than or equal to" the observed sample mean 4.62, instead of simply "equal to" or "greater than or equal to". Next, explain why we subtract 5 and divide by $0.92/\sqrt{45}$ in the middle probability expression.

One-Sided Tests: The claim we tested in Exercise #1 on Tuesday claimed that the actual population mean was *less* than a specific number. In particular, the alternative hypothesis for Exercise #1 was $H_A : \mu < 5$ mg/l. Testing a claim of "less than" is a *one-sided test*, specifically referred to as a **left-tail test**. We now consider an example of a claim of "greater than", i.e., a **right-tail test**.

SAT Example: We suspect that on average students will score higher on their second attempt at the SAT math exam than on their first attempt. The data set SAT gives the changes in score (second try minus first try) results for 46 randomly chosen high school students. We will perform a hypothesis test to see if these data provide good evidence that the mean change in the population is greater than zero.

Pause for Reflection #2:

For the SAT example just introduced, state the null and alternative hypotheses being tested using appropriate parameter notation.

We now load the SAT data and compute the sample mean and standard deviation:

```
SAT = read.csv("SAT.csv")
xbar = mean(SAT$SAT.change)
s = sd(SAT$SAT.change)
```

If we assume that changes in SAT scores are normally distributed, then the test statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{46}}, \quad (1)$$

has a t distribution with 45 degrees of freedom, since the population standard deviation is unknown.

Pause for Reflection #3:

Calculate the P -value associated with the observed SAT results and clearly state your conclusion for testing the hypotheses you stated in Reflection #2. Remember that the P -value is the probability of getting a value as extreme as or more extreme than the observed test statistic assuming the null hypothesis is true.

Pause for Reflection #4:

Does the sample of changes in SAT scores support the assumption that the population is normally distributed?

Pause for Reflection #5:

Comment on why you think the terminology *left-tail test* and *right-tail test* are used to describe one-sided tests. Specifically, why is testing a claim of "less than" done with a left-tail test, and a claim of "greater than" with a right-tail test?

Two-Sided Tests: Suppose that we were not sure whether on average students score higher or lower on their second attempt, and instead we just want to test the claim that on average the scores are not the same. In this case, we are testing a claim of either "less than" or "greater than". Simply put, we are testing a claim of "difference", which includes both cases. This type of test is referred to as a *two-sided test* or a *two-tail test*.

SAT Example: For the change in SAT scores, the null hypothesis remains the same, but the alternative is now stated simply as the mean change in SAT scores for all students μ is not equal to zero:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_A : \mu \neq 0.$$

We use the same test statistic given in equation (1), but to calculate the P -value in a two-sided test we consider both less than and greater than values as extreme, which is accomplished using absolute values:

$$P\text{-value} = P(|T| \geq |t|) = P(T \leq -|t|) + P(T \geq |t|),$$

where t denotes the observed value of the test statistic.

Pause for Reflection #6:

Calculate the P -value for the two-sided test. How does it compare to the P -value you calculated in Reflection #3 for the one-sided, right-tail test?

The T -Test in General: Notice the general procedure we have followed in the water quality example and the change in SAT scores example:

1. state the null and alternative hypotheses,
2. calculate a test statistic from observed data,
3. find or estimate a sampling distribution for the test statistic, assuming the null hypothesis is true,
4. calculate a P -value using that distribution,
5. and finally state a conclusion of the test, rejecting H_0 if P -value is small.

In permutation testing, the sampling distribution found in step 3 is given by the permutation distribution obtained by permuting the data. In the classical approach taken in the above examples, the sampling distributions are *parametric*, normal or t distributions.

We note that the second example is more likely, i.e., it is more likely that the population standard deviation is unknown and so the P -value will be calculated using a t distribution. The following summarizes the classical approach for testing a claim about a

population mean when the population standard deviation is unknown, known as a *t*-test.

T-Test for a Normal Mean

Let X_1, \dots, X_n be a random sample from a normal population with unknown μ and σ . Let \bar{X} and S denote the sample mean and standard deviation. For a null hypothesis given by

$$H_0 : \mu = \mu_0,$$

where μ_0 is a constant, we form the *t*-test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

If the null hypothesis is true, then T has a *t* distribution with $(n - 1)$ degrees of freedom. The *P*-value is the probability that chance alone would produce a test statistic as extreme as or more extreme than the observed value of the test statistic $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, if the null hypothesis is true. What is considered "extreme" depends on the alternative hypothesis: left-tail, right-tail, or two-tail.

Note that the *t*-test is *exact* when the sample comes from a normally distributed population. If the population is not normal, then the *t*-test provides an approximate *P*-value, since by the Central Limit Theorem the sample mean \bar{X} will be approximately normal even though the population is not. So, in practice, the *t*-test is used to compute approximate *P*-values when the sample size is large and the sample is not too skewed. If the sample size is not large and/or the sample is severely skewed, then a permutation test should be used.

Using R to do all the calculations: The `t.test()` function in R performs the *t*-test for a normal mean, provided that we have all sample data and not just sample statistics (i.e., sample mean and standard deviation values). For the change in SAT scores example, the following performs the two-sided test:

```
t.test(SAT$SAT.change)
```

The default settings of `t.test()` are to perform a two-sided test that the population mean is zero. But the type of test can be changed by adding the argument `alternative = "less"` for a left-tail test or `alternative = "greater"` for a right-tail test. So, for the one-sided test that the change in SAT scores is greater than zero, try the following and compare to what you found in Reflection #3:

```
t.test(SAT$SAT.change, alternative = "greater")
```

We can also change the value of μ claimed under the null hypothesis. For example, suppose that in the past it was determined that on average students score 10 points higher on their second attempt at the SAT math test, and we want to test that the average change is now more than 10 points. In this case, the null hypothesis is now $H_0 : \mu = 10$, and we can test this with the data as follows:

```
t.test(SAT$SAT.change, alternative = "greater", mu = 10)
```

Pause for Reflection #7:

Comment on the effect that changing the null hypothesis to $H_0 : \mu = 10$ had on the resulting test statistic and P -value compared to the corresponding values you found in Reflection #3 for the test that $\mu = 0$. Explain why the test statistic and P -value changed in this way.

Tests Comparing Two Population Means: In Lab 2, we actually performed a permutation test for hypotheses concerning two populations, women and men. There is also a t -test for comparing two populations.

If μ_1 and μ_2 denote the two population means, \bar{X}_1 and \bar{X}_2 denote sample means for samples of size n_1 and n_2 , respectively, taken from the two populations, and S_1 and S_2 denote the sample standard deviations, then the test statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}$$

has approximately a t distribution. Notice that the test statistic involves *differences* in the means, and because of this hypotheses for comparing two population means are typically stated in terms of the difference in the means. Furthermore, the null hypothesis is almost always the claim that there is no difference, i.e., that $\mu_1 - \mu_2 = 0$. Let's look at an example.

Births Example: Consider again the `NCBirths2004` data set from Lab 5, which contains information on a random sample of babies born in North Carolina during 2004. In addition to the birth weights of the babies, the variable `Tobacco` in the data set indicates whether or not the mothers of the babies smoked during the pregnancy. Using this data, we can test the claim that smoking during pregnancy results in a lower birth weight than not smoking, on average.

Pause for Reflection #8:

Let μ_1 and μ_2 denote the true mean weight of babies born to smoking and nonsmoking mothers, respectively. Using the difference $\mu_1 - \mu_2$, state the null and alternative hypotheses to test the claim that smoking during pregnancy results in a lower birth weight than not smoking, on average.

We can use R/RStudio to do all of the necessary calculations for us, so that we do not have to work with equation (2) directly. The same `t.test()` function can be used to perform a two-sample test:

```
births = read.csv("NCBirths2004.csv")
smoker = subset(births, select = weight, subset = Tobacco == "Yes", drop = TRUE)
nonsmoker = subset(births, select = weight, subset = Tobacco == "No", drop = TRUE)
t.test(smoker, nonsmoker, alternative = "less")
```

Pause for Reflection #9:

Based on the P -value for the above two-sample t -test, is the result statistically significant? Write a conclusion to the test of the hypotheses stated in Reflection #8.

Look at the output from the `t.test()` and locate the sample estimates. What is the mean birth weight for babies born to smoking mothers and what is the mean for nonsmoking mothers?

Assumptions of the Two-Sample T -Test: As with the one-sample t -test, we need to check that the two populations being compared in a two-sample t -test are normally distributed. If this assumption is violated, in particular, if the distributions are skewed and the sample sizes of the two samples are different, then the actual distribution of the test statistic given in equation (2) may differ substantially from the t distribution. In that case, a permutation test would be more reliable.

Pause for Reflection #10:

Check that the populations of birth weights for babies born to smokers and babies born to nonsmokers both appear to be normally distributed.

Matched Pairs: The two-sample t -test also requires that the two samples be from two *independent* populations. If the two samples are *paired* (not independent), then we need to perform a *paired t -test*, which is done with the same `t.test()` function by adding the argument `paired = TRUE`. The change in SAT scores example can be performed as a paired t -test, instead of a one-sample t -test as we did previously.

SAT Example: Take a closer look at the SAT data set:

```
View(SAT)
summary(SAT)
```

Notice that in addition to the variable `SAT.change`, which contains the differences between the second attempt and first attempt for each student, we also have the actual scores on the second attempt in variable `SAT.2` and a list of scores on the first attempt in variable `SAT.1` for each student.

In this case, the two samples of scores for the first and second attempts are paired, consisting of pairs of SAT scores for the random sample of students. In other words, the rows in the SAT data set correspond to a single student and so the values are *related*, not independent. Thus, if we set up the test of the claim that on average students will score higher on their second attempt at the SAT math exam than on their first attempt by comparing the two samples, we need to perform a paired t -test.

Let μ_1 denote the true mean of SAT scores for the first attempt and let μ_2 denote the true mean for the second attempt. Following the order we took the difference in earlier - second attempt minus first - we state the hypotheses being tested as follows:

$$H_0 : \mu_2 - \mu_1 = 0 \quad \text{vs.} \quad H_A : \mu_2 - \mu_1 > 0.$$

We can use the `t.test()` function to calculate the test statistic and associated P -value needed to perform the test:

```
t.test(SAT$SAT.2, SAT$SAT.1, alternative = "greater", paired = TRUE)
```

Notice the order that the samples of SAT scores are listed in the argument of `t.test()`, it matches the order we that we took the difference in population means when stating the hypotheses.

Pause for Reflection #11:

Compare the results of the paired t -test to the results you found in Reflection #3. How do the test statistics and corresponding P -values compare? Can you provide an explanation for why?

It is important to perform the paired t -test when samples are paired, because the results of the test are often very different from the results of the two-sample t -test. For example, omit the argument `paired = TRUE` from the `t.test()` we ran on the SAT scores:

```
t.test(SAT$SAT.2, SAT$SAT.1, alternative = "greater")
```

Pause for Reflection #12:

Now compare the results of the two-sample t -test you just ran to the results of the paired t -test. Again focus on the test statistics and corresponding P -values. What do you notice?

Lab 6: More Hypothesis Testing - Classical Approach is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.