# Null Hypothesis Significance Testing II
# Class 18, 18.05
# Jeremy Orloff and Jonathan Bloom

## 1 Learning Goals

1. Be able to list the steps common to all null hypothesis significance tests.

2. Be able to define and compute the probability of Type I and Type II errors.

3. Be able to look up and apply one- and two-sample $t$-tests.

## 2 Introduction

We continue our study of significance tests. In these notes we will introduce two new tests: one-sample $t$-tests and two-sample $t$-tests. You should pay careful attention to the fact that every test makes some assumptions about the data – often that is drawn from a normal distribution. You should also notice that all the tests follow the same pattern. It is just the computation of the test statistic and the type of the null distribution that changes.

## 3 Review

### 3.1 Setting up and running a significance test

There is a fairly standard set of steps one takes to set up and run a null hypothesis significance test.

1. Design an experiment to collect data and choose a test statistic $x$ to be computed from the data. The key requirement here is to know the null distribution $\phi(x|H_0)$. To compute power, one must also know the alternative distribution $\phi(x|H_A)$.

2. Decide if the test is one or two-sided based on $H_A$ and the form of the null distribution.

3. Choose a significance level $\alpha$ for rejecting the null hypothesis.

4. Decide how much data you need to collect to achieve the desired power for the test.

5. Run the experiment to collect data $x_1, x_2, \ldots, x_n$.

6. Compute the test statistic $x$.

7. Compute the $p$-value corresponding to $x$ using the null distribution.

8. If $p < \alpha$, reject the null hypothesis in favor of the alternative hypothesis.

**Notes.**

1. Rather than choosing a significance level, you could instead choose a rejection region and reject $H_0$ if $x$ falls in this region. The corresponding significance level is then the probability, assuming $H_0$, that $x$ falls in the rejection region.

2. The null hypothesis is often the 'cautious hypothesis'. The lower we set the significance level, the more "evidence" we will require before rejecting our cautious hypothesis in favor of a more sensational alternative. It is standard practice to publish the $p$ value itself so that others may draw their own conclusions.

3. **A key point of confusion:** A significance level of 0.05 does not mean the test only makes mistakes 5% of the time. It means that if the null hypothesis is true, then the probability the test will mistakenly reject it is 5%. The power of the test measures the accuracy of the test when the alternative hypothesis is true. Namely, the power of the test is the probability of rejecting the null hypothesis if the alternative hypothesis is true. Therefore the probability of falsely failing to reject the null hypothesis is 1 minus the power.

4. **Another key point of confusion:** We use $p$-values, but conceptually the $p$-value is just a computational trick. After choosing a test statistic, the conceptual order is: first pick a significance level, then use this to define the rejection region. We reject the null hypothesis if the test statistic is in the rejection region. All the $p$-value does is tell us in one computation whether or not the test stastic is in the rejection region.

**Errors**. We can summarize these two types of errors and their probabilities as follows:

$$\begin{aligned}
\text{Type I error} &= \text{rejecting } H_0 \text{ when } H_0 \text{ is true.}\\
\text{Type II error} &= \text{failing to reject } H_0 \text{ when } H_A \text{ is true.}
\end{aligned}$$

$$\begin{aligned}
\text{P(type I error)} &= \text{probability of falsely rejecting } H_0\\
&= \text{P(test statistic is in the rejection region} \mid H_0)\\
&= \text{significance level of the test}\\
\text{P(type II error)} &= \text{probability of falsely not rejecting } H_0\\
&= \text{P(test statistic is in the acceptance region} \mid H_A)\\
&= \text{1 - power.}
\end{aligned}$$

**Helpful analogies**.

In terms of medical testing for a disease: a Type I error is a false positive and a Type II error is a false negative.

In a jury trial, a Type I error is convicting an innocent defendant and a Type II error is acquitting a guilty defendant.

## 3.2 Power

We discussed power in the Class 17 notes. Power is the probabilitiy of correctly rejecting the null hypothesis. It depends on the alternative hypothesis $H_A$ being considered.

The ideal test has power equal to 1.0 and significance equal to 0.0. Of course, in general, this is impossible. And we want to find some compromise where power is high and signficance is low.

In symbols: power $= P(\text{data is in the rejection region} \mid H_A)$.

Compare this with: signficance $= P(\text{data is in the rejection region} \mid H_0)$.

## 4 Understanding a significance test

Questions to ask:

1. How did they collect data? What is the experimental setup?
2. What are the null and alternative hypotheses?
3. What type of significance test was used?
   Does the data match the criteria needed to use this type of test?
   How robust is the test to deviations from these criteria?
4. For example, some tests comparing two groups of data assume that the groups are drawn from distributions that have the same variance. This needs to be verified before applying the test. Often the check is done using another significance test designed to compare the variances of two groups of data.
5. How is the $p$-value computed?
   A significance test comes with a test statistic and a null distribution. In most tests the $p$-value is

   $$p = P(\text{data at least as extreme as what we got} \mid H_0)$$

   What does 'data at least as extreme as the data we saw' mean? For example, is the test one or two-sided?
6. What is the significance level $\alpha$ for this test? If $p < \alpha$ then the experimenter will reject $H_0$ in favor of $H_A$.
7. What is the power of the test?

## 5 $t$ tests

Many significance tests assume that the data are drawn from a normal distribution, so before using such a test you should examine the data to see if the normality assumption is reasonable. We will describe how to do this in more detail later, but plotting a histogram is a good start. Like the $z$-test, the one-sample and two-sample $t$-tests that we consider below start from this normality assumption.

We don't expect you to memorize all the computational details of these tests and those to follow. In real life, you have access to textbooks, google, and wikipedia; on the exam, you'll have your notecard. Instead, you should be able to identify when a $t$-test is appropriate and apply this test after looking up the details and using a table or software like R.

### 5.1 $z$-test

Let's first review the $z$-test.

- Data: we assume $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$, where $\mu$ is unknown and $\sigma$ is known.
- Null hypothesis: $\mu = \mu_0$ for some specific value $\mu_0$
- Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ = standardized mean
- Null distribution: $\phi(z \,|\, H_0)$ is the pdf of $Z \sim N(0,1)$
- One-sided $p$-value (right side): $p = P(Z \geq z \,|\, H_0)$
  One-sided $p$-value (left side): $p = P(Z \leq z \,|\, H_0)$
  Two-sided $p$-value: $p = \begin{cases} 2P(Z \geq z) & \text{if } z > 0 \\ 2P(Z \leq z) & \text{if } z < 0. \end{cases}$.

  Because of the symmetry of the distribution around 0, we can also write this as $p = P(|Z| \geq |z|)$.

  See Example 1b for the rationale for this.

**Example 1.** Suppose that we have data that follows a normal distribution of unknown mean $\mu$ and known variance 4. Let the null hypothesis $H_0$ be that $\mu = 2$. Let the alternative hypothesis $H_A$ be that $\mu > 2$. Suppose we collect the following data:

$$3,\ 2,\ 5,\ 7,\ 1$$

At a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

**Solution:** There are 5 data points with average $\overline{x} = 3.6$. Because we have normal data with a known variance we should use a $z$ test. Our $z$ statistic is

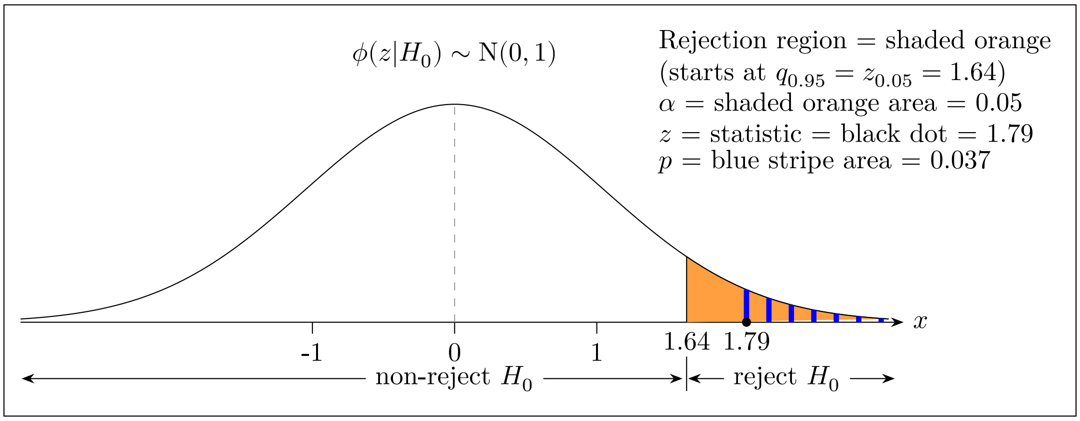
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.6 - 2}{2/\sqrt{5}} = 1.79$$

Our test is one-sided because the alternative hypothesis is one-sided. So (using R) our $p$-value is

$$p = P(Z > z) = P(Z > 1.79) = 0.037$$

Since $p < \alpha = 0.05$, we reject the null hypothesis in favor of the alternative hypothesis that $\mu > 2$.

We can visualize the test as follows:



**Example 1b.** Repeat Example 1 as a two-sided test, i.e. with $H_A$ being $\mu \neq 2$.

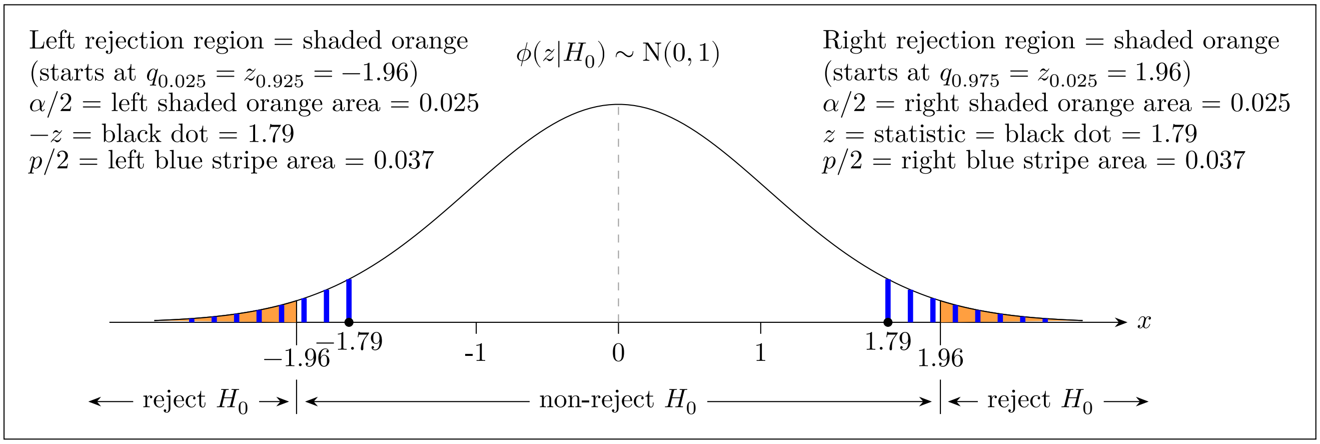
**Solution:** Let's do the test and then we'll explain the rationale behind the computation of the $p$-value.

Since $z > 0$, $p = 2P(Z > z) = 0.074$. Since, $p > \alpha = 0.05$, the data does not support rejecting the null hypothesis in favor of $H_A$.

**Reason for the factor of 2 in the computation of $p$**

The reason is essentially arithmetic. Remember, the purpose of the $p$-value is that $p \leq \alpha$ indicates that the test statistic is in the rejection region.

The picture below illustrates the following. For a two-sided test, each side of the rejection region has probability $\alpha/2$. So, if the test statistic is on the right, then it is in the rejection region if $P(Z > z) \leq \alpha/2$, i.e. if $p = 2P(Z > z) \leq \alpha$

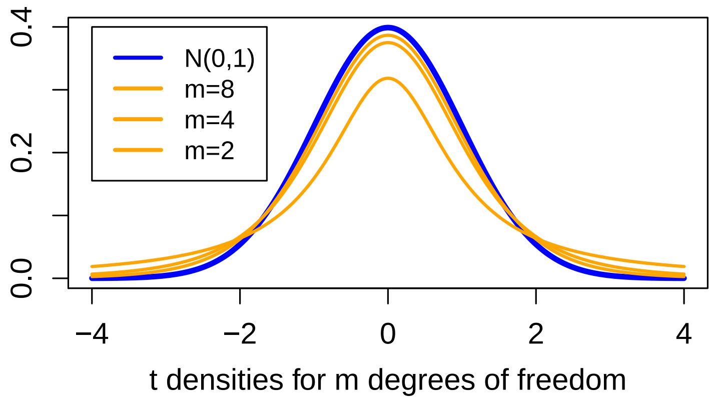


## 5.2 The Student $t$ distribution

'Student' is the pseudonym used by the William Gosset who first described this test and distribution. See https://en.wikipedia.org/wiki/Student's_t-test

The $t$-distribution is symmetric and bell-shaped like the normal distribution. It has a parameter $df$ which stands for degrees of freedom. For $df$ small the $t$-distribution has more probability in its tails than the standard normal distribution. As $df$ increases $t(df)$ becomes more and more like the standard normal distribution.

Here is a simple applet that shows $t(df)$ and compares it to the standard normal distribution: https://mathlets.org/mathlets/t-distribution/



As degrees of freedom increases the t-distribution becomes normal

### 5.3 R

As usual in R, the functions `pt, dt, qt, rt` correspond to cdf, pdf, quantiles, and random sampling for a $t$ distribution. Remember that you can type `?dt` in RStudio to view the help file specifying the parameters of `dt`. For example, `pt(1.65,3)` computes the probability that $x$ is less than or equal 1.65 given that $x$ is sampled from the $t$ distribution with 3 degrees of freedom, i.e. $P(x \leq 1.65)$ given that $x \sim t(3)$).

### 5.4 One sample $t$-test

For the $z$-test, we assumed that the variance of the underlying distribution of the data was known. However, it is often the case that we don't know $\sigma$ and therefore we must estimate it from the data. In these cases, we use a one sample $t$-test instead of a $z$-test and the studentized mean in place of the standardized mean

- Data: we assume $x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$, where both $\mu$ and $\sigma$ are unknown.
- Null hypothesis: $\mu = \mu_0$ for some specific value $\mu_0$
- Test statistic:
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$
where
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$
Here $t$ is called the Studentized mean and $s^2$ is called the sample variance. The latter is an estimate of the true variance $\sigma^2$.
- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(n-1)$, the $t$ distribution with $n-1$ degrees of freedom.*

- One-sided $p$-value (right side): $p = P(T \geq t \,|\, H_0)$
  One-sided $p$-value (left side): $p = P(T \leq t \,|\, H_0)$
  Two-sided $p$-value: $p = \begin{cases} 2P(T \geq t) & \text{if } t > 0 \\ 2P(T \leq t) & \text{if } t < 0. \end{cases}$

  Because of the symmetry of the distribution around 0, we can also write this as $p = P(|T| \geq |t|)$.

***Important note.** This is a good example of how we will work with significance tests. Once we know the distribution of the test statistic, all the tests have the same basic form. In this case, we make use of a theorem that says, for normal data the Studentized mean follows a $t$-distribution. We will not prove this in 18.05, but you can look up the proof if you want: https://en.wikipedia.org/wiki/Student's_t-distribution#Derivation

**Example 2.** Now suppose that in Example 1 the variance is unknown. That is, we have data that follows a normal distribution of unknown mean $\mu$ and and unknown variance $\sigma$. Suppose we collect the same data as before:

$$1,\ 2,\ 3,\ 6,\ -1$$

As above, let the null hypothesis $H_0$ be that $\mu = 0$ and the alternative hypothesis $H_A$ be that $\mu > 0$. At a significance level of $\alpha = 0.05$, should we reject the null hypothesis?

**Solution:** There are 5 data points with average $\overline{x} = 2.2$. Because we have normal data with unknown mean and unknown variance we should use a one-sample $t$ test. Computing the sample variance we get

$$s^2 = \frac{1}{4}\left((1-2.2)^2 + (2-2.2)^2 + (3-2.2)^2 + (6-2.2)^2 + (-1-2.2)^2\right) = 6.7$$

Our $t$-statistic is the Studentized mean:

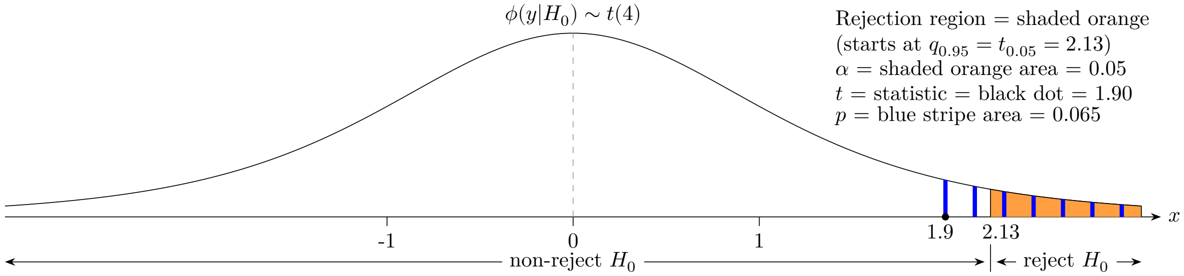
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{2.2 - 0}{\sqrt{6.7}/\sqrt{5}} = 1.901$$

Our test is one-sided because the alternative hypothesis is one-sided. So (using R) the $p$-value is

$$p = P(T > t) = P(T > 1.901) = \texttt{1-pt(1.901,4)} = 0.065$$

Since $p > 0.05$, we do not reject the null hypothesis.

We can visualize the test as follows:



### 5.5 Two-sample $t$-test with equal variances

We next consider the case of comparing the means of two samples. For example, we might be interested in comparing the mean efficacies of two medical treatments.

- Data: We assume we have two sets of data drawn from normal distributions

  $$\begin{aligned} x_1, x_2, \ldots, x_n &\sim N(\mu_1, \sigma^2) \\ y_1, y_2, \ldots, y_m &\sim N(\mu_2, \sigma^2) \end{aligned}$$

  where the means $\mu_1$ and $\mu_2$ and the variance $\sigma^2$ are all unknown. Notice the assumption that the two distributions have the same variance. Also notice that there are $n$ samples in the first group and $m$ samples in the second.

- Null hypothesis: $\mu_1 = \mu_2$ (the values of $\mu_1$ and $\mu_2$ are not specified)

- Test statistic:

  $$t = \frac{\overline{x} - \overline{y}}{s_p},$$

  where $s_p^2$ is the pooled variance

  $$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)$$

  Here $s_x^2$ and $s_y^2$ are the sample variances of the $x_i$ and $y_j$ respectively. The expression for $t$ is somewhat complicated, but the basic idea remains the same and it still results in a known null distribution.

- Null distribution: $\phi(t \,|\, H_0)$ is the pdf of $T \sim t(n+m-2)$.

- One-sided $p$-value (right side): $p = P(T > t \,|\, H_0)$
  One-sided $p$-value (left side): $p = P(T < t \,|\, H_0)$
  Two-sided $p$-value: $p = P(|T| > |t|)$.

**Note 1:** Some authors use a different notation. They define the pooled variance as

$$s^2_{p\text{-other-authors}} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

and what we called the pooled variance they point out is the estimated variance of $\overline{x} - \overline{y}$. That is,

$$s_p^2 \;=\; s_{p\text{-other-authors}} \times (1/n + 1/m) \;\approx\; s^2_{\overline{x}-\overline{y}}$$

**Note 2:** There is a version of the two-sample $t$-test that allows the two groups to have different variances. In this case the test statistic is a little more complicated but R will handle it with equal ease.

**Note 3:** We reiterate our 'important note' from above: It can be proved that under the assumptions on the data (independent samples, normal data, equal variances), the null distribution is a $t$-distribution. We won't prove this in 18.05. But knowing it, we can work with and understand the gist of the two-sample $t$-test in exactly the same way we can understand other significance tests.

**Example 3.** The following data comes from a real study in which 1408 women were admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period. We can summarize the data as follows:

Medical: 775 observations with $\bar{x}_M = 39.08$ and $s_M^2 = 7.77$.

Emergency: 633 observations with $\bar{x}_E = 39.60$ and $s_E^2 = 4.95$

Set up and run a two-sample $t$-test to investigate whether the mean duration differs for the two groups.

What assumptions did you make?

**Solution:** The pooled variance for this data is

$$s_p^2 = \frac{774(7.77) + 632(4.95)}{1406}\left(\frac{1}{775} + \frac{1}{633}\right) = 0.0187$$

The $t$ statistic for the null distribution is

$$\frac{\bar{x_M} - \bar{y_E}}{s_p} = -3.8064$$

We have 1406 degrees of freedom. Using R to compute the two-sided $p$-value we get

$$p = P(|T| > |t|) = \texttt{2*pt(-3.8064, 1406) = 0.00015}$$

$p$ is very small, much smaller than $\alpha = 0.05$ or $\alpha = 0.01$. Therefore we reject the null hypothesis in favor of the alternative that there is a difference in the mean durations.

Rather than compute the two-sided $p$-value exactly using a $t$-distribution we could have noted that with 1406 degrees of freedom the $t$ distribution is essentially standard normal and 3.8064 is almost 4 standard deviations. So

$$P(|t| \geq 3.8064) \approx P(|z| \geq 3.8064) < 0.001$$

We assumed the data was normal and that the two groups had equal variances. Given the large difference between the sample variances this assumption may not be warranted.

In fact, there are other significance tests that test whether the data is approximately normal and whether the two groups have the same variance. In practice one might apply these first to determine whether a $t$ test is appropriate in the first place. We don't have time to go into normality tests here, but we will see the $F$ distribution used for equality of variances next week.

https://en.wikipedia.org/wiki/Normality_test
https://en.wikipedia.org/wiki/F-test_of_equality_of_variances