**Hypothesis tests – Two-tail tests**



These notes discuss a different approach to the inference portion of the diagram.

Hypothesis: A statement that something is true.

Below is an example to help introduce hypothesis testing:

Example: Light Bulbs (light\_bulbs.R)

Suppose the company wants to estimate the mean lifetime of its light bulbs. It hypothesizes that μ = 250 (this could be what is stated on the package). How can this be checked?

The company takes a random sample of 16 light bulbs and finds they last on average for 299.2 hours with a standard deviation of 80 hours. The 95% CI for μ is 264.14 < μ < 334.26.

Is μ = 250? Because 264.14 < μ < 334.26 with a 95% level of confidence, μ appears to be greater than 250. Therefore, reject the hypothesis of μ = 250.

Suppose before the sample was conducted, the company hypothesized that μ = 270. Is this correct?

Again, the sample was taken and the CI above was obtained. Because μ could be 264.15, 268, 270, 272, 300,…, μ = 270 may be correct. Therefore, do not reject the hypothesis of μ = 270.

There is not sufficient evidence from the sample to prove the hypothesized value of μ = 270 to be incorrect.

Finally, suppose before the sample was conducted, the company hypothesized that μ = 350. Is this correct?

Again, the sample was taken and the confidence interval above was obtained. Because 264.14 < μ < 334.26 with a 95% level of confidence, μ appears to be less than 350. Therefore, reject the hypothesis of μ = 350.

The above is an informal example of a hypothesis test. In many real life situations, there is a hypothesis about the population mean or other population parameters. A sample from the population is taken to investigate the hypothesis.

For the first hypothesis of μ = 250 in the light bulb example, two hypotheses were considered:

* Null Hypothesis, Ho: μ = 250
* Alternative Hypothesis, Ha: μ ≠ 250

One of two possible decisions was made:

* Reject Ho - This indicates μ is not 250
* Don't Reject Ho - This indicates there is not sufficient evidence from the sample to say μ is different from 250. You cannot say "Accept Ho"; i.e., cannot say Ho is true. See the reason in the following (and previous) example.

Notes:

* Most often when performing a hypothesis test, you will put your research hypothesis in Ha.
* Some people will use the terminology “Fail to reject Ho” instead of “Don’t reject Ho”. Both are fine to use. Some textbooks say “accept Ho” sometimes when the null hypothesis is not rejected. However, as we saw in the previous example, it would be difficult to say μ = 270 because you would need a confidence interval of (270, 270)!
* Rather than using “Ho”, some textbooks will use “H0”. Also, rather than using “Ha”, some textbooks will use “H1”.

Example: Jury Trials

Juries are asked to consider two hypotheses:

Ho: Defendant is innocent

Ha: Defendant is guilty

The defendant is assumed innocent until proven guilty. In hypothesis testing, we assume Ho is true until there is enough evidence to prove otherwise.

The jury listens to the prosecution and the defense to make a judgment. This is like taking a SAMPLE.

* If there is ENOUGH evidence (beyond a reasonable doubt) to convict: Reject Ho, the defendant is “guilty”.
* If there is NOT ENOUGH (reasonable doubt) evidence to convict: Don't Reject Ho, the defendant is “not guilty”. Notice, this does not mean the defendant is innocent.

Types of errors in hypothesis test decisions

* Type I ‑ Reject Ho, but in reality Ho is true (reality = population)
* Type II ‑ Don't reject Ho, but in reality Ha is true

These errors indicate that the sample led us to believe something about the population that is incorrect.

Example: Jury Trials

* Type I: Reject Ho = jury says the defendant is guilty, but Ho is really true = defendant is innocent.

Send an innocent person to jail

* Type II: Don't reject Ho = jury says the defendant is not guilty, but Ha is really true = defendant is guilty.

Let a criminal go free

## Probability of making errors

A type I error is the more serious error in the jury trial example and in statistics. Thus, the P(Type I error) is controlled in a hypothesis test at a pre-specified level denoted by α. Therefore,

P(Type I error) = P(Reject Ho | Ho is TRUE) = α.

This is also called the “level of significance”. Reasons for this name are given later in these notes.

A type II error is generally not as serious, so it is usually not controlled at a fixed level. We can still define the probability of committing this error:

P(Type II error) = P(Don’t reject Ho | Ha is TRUE) = β

Table describing the two errors:

|  |  |  |
| --- | --- | --- |
|  |  | Based on Sample |
|  |  | Reject Ho | Don’t Reject Ho |
| Population | Ho is TRUE | Type I Error | Correct Conclusion |
| Ha is TRUE | Correct Conclusion | Type II Error |

Same table, but with the conditional probabilities:

|  |  |  |
| --- | --- | --- |
|  |  | Based on Sample |
|  |  | Reject Ho | Don’t Reject Ho |
| Population | Ho is TRUE | P(Reject Ho | Ho is TRUE) = α | P(Don’t reject Ho | Ho is true) = 1-α |
| Ha is TRUE | P(Reject Ho | Ha is TRUE) = 1-β | P(Don’t reject Ho | Ha is TRUE) = β |

Use the basic definitions of conditional probabilities from Chapter 4 to help interpret the table! Remember that P(A|B) + P(|B) = 1.

Power

In hypothesis testing, we will make the assumption that Ho is true and then try to prove it to be incorrect using evidence gathered in the sample. Thus, it is important to define

P(Reject Ho | Ha is TRUE).

This is called the power of the test. Notice where this result falls in the above table and it has a probability of 1 – β.

Question: Do you want this probability to be small or large?

We will discuss power in more detail later in the course.

Three Methods for performing a hypothesis test

1. Confidence interval
2. Test statistic
3. P-value

All three provide the same answer when testing the population mean! Note that there may be slightly different conclusions when testing other parameters, like a population proportion (to be discussed later).

1) The confidence interval method - 4 Steps

1. State Ho:μ=μ0
 Ha:μ≠μ0 where μ0 is some number
2. Find the CI for μ
3. Reject or do not reject Ho – Check if the hypothesized value of μ is inside the interval.
4. Conclusion – Describe what 3. means in terms of the original problem

Example: GPA Example (gpa\_HypTest.R)

Test the hypothesis that the mean GPA of students is 3.0. Suppose P(Type I error) = α = 0.05,  = 2.9, n = 16, and s = 0.1.

1. Ho:μ=3.0
Ha:μ≠3.0
2. 

⇒ 

⇒ 2.847 < μ < 2.953

1. Reject Ho because μ = 3.0 is not in the interval.
2. The mean GPA of students is not 3.0.

Notes:

* The phrasing for step #4 provides a template for you to use:
	+ The mean value is different from the value stated in the null hypothesis.
	+ There is sufficient evidence to prove μ is different from the value stated in the null hypothesis.
* The probability of incorrectly rejecting μ = 3.0 is 5% (probability of making a type I error). Thus, if the whole process of taking a sample and doing the hypothesis is repeated 1,000 times WITH μ = 3.0, we would expect 0.05×1,000 = 50 times to incorrectly reject Ho:μ=3.0.

Example: Volleyball quality control (volleyball.R, volleyball.csv)

Suppose Mikasa, a volleyball manufacturer, is concerned about whether their volleyballs are being produced with the correct radius of 11.6cm. A sample of 36 volleyballs is taken with  = 11.5 and s = 1. Part of the data set is below.

|  |
| --- |
| **Volleyball Radius** |
| 11.38 |
| 12.78 |
| 10.61 |
|  |
| 10.10 |

Is there evidence to show the volleyballs are being made incorrectly? Conduct a hypothesis test with α = 0.05.

Below is the R code and output:

> volleyball <- read.csv(file = “volleyball.csv")

> head(volleyball)

 radius

1 11.37772

2 12.78000

3 10.60970

4 10.47060

5 10.28741

6 11.99000

> #Long way

> ybar <- mean(volleyball$radius)

> s <- sd(volleyball$radius)

> alpha <- 0.05

> n <- length(volleyball$radius)

> lower <- ybar - qt(p = 1 - alpha/2, df = n-1) \* s /

 sqrt(n)

> upper <- ybar + qt(p = 1 - alpha/2, df = n-1) \* s /

 sqrt(n)

> data.frame(lower, upper)

 lower upper

1 11.16164 11.83834

> #Short way

> t.test(x = volleyball$radius, alternative = "two.sided",

 mu = 11.6, conf.level = 0.95)

 One Sample t-test

data: volleyball$radius

t = -0.6001, df = 35, p-value = 0.5523

alternative hypothesis: true mean is not equal to 11.6

95 percent confidence interval:

 11.16164 11.83834

sample estimates:

mean of x

 11.49999

1. Ho:μ=11.6
Ha:μ≠11.6
2. 

⇒ 

⇒ 11.16 < μ < 11.84

1. Do not reject Ho because μ = 11.6 is in the interval.
2. There is not sufficient evidence to prove the volleyballs are being made incorrectly.

OR

There is not sufficient evidence to conclude the population mean radius is different from 11.6.

Notes:

* The phrasing for step #4 provides a template for you to use: There is not sufficient evidence to prove μ is different from the value stated in the null hypothesis.
* What should Mikasa do? Continue with production of the volleyballs.
* I did not say, "The volleyballs are being produced correctly." THIS IS WRONG because of the probability of committing a Type II Error is NOT controlled (β was not stated). Compare this to the GPA example!

Here’s how you can save the results from t.test() in an object. This can be useful if you need some of the results for later computations.

> save.results <- t.test(x = volleyball$radius, alternative

 = "two.sided", mu = 11.6, conf.level = 0.95)

> names(save.results)

 [1] "statistic" "parameter" "p.value" "conf.int"

 [5] "estimate" "null.value" "stderr" "alternative"

 [9] "method" "data.name"

> save.results$conf.int

[1] 11.16164 11.83834

attr(,"conf.level")

[1] 0.95

> save.results$conf.int[1]

[1] 11.16164

> save.results$conf.int[2]

[1] 11.83834

2) The test statistic method - 5 Steps

1. State Ho and Ha
2. Find the test statistic: 

The test statistic examines how far the sample mean is from the hypothesized mean. The numerator of t,  – μ0, is divided by  to account for the variation of .

To help understand why we can use this statistic for a hypothesis test, consider the random variable version of the test statistic:



We can examine the probability statement



The yellow highlighted area corrects a small typo that appeared in the video

provided that **Ho is true** and Y1, Y2, …, Yn is a random sample from a population characterized by a normal probability distribution with E(Yi) = μ0 and Var(Yi) = σ2 for i = 1, …, n. Thus, we have a very likely numerical range for the observed values of



to fall within. Graphically, we expect

observed values to fall  and . If an observed value falls outside of this range, this gives us evidence that our initial assumption of “Ho is true” is incorrect! This is why the title of the section is “two- tail tests”. Some people will call these “two- sided tests” as well.

1. Find the critical values: ±tα/2, n-1

There are two critical values which define the range of probable values for t.

1. Reject or do not reject Ho
	1. Draw the t distribution
	2. Plot the critical value
	3. Label the graph with reject and don't reject regions
	4. Plot the test statistic
	5. Write reject or don’t reject Ho and provide a reason
2. Conclusion – Describe what 4. means in terms of the original problem.

Example: Volleyball quality control (volleyball.R, volleyball.csv, and hyp\_1sample\_pic.xlsx)

> mu0 <- 11.6

> t <- (ybar - mu0)/(s/sqrt(n))

> crit.val <- qt(p = 1 - alpha/2, df = n-1)

> data.frame(t, crit.val)

 t crit.val

1 -0.600065 2.030108

> t.test(x = volleyball$radius, alternative = "two.sided",

 mu = 11.6, conf.level = 0.95)

 One Sample t-test

data: volleyball$radius

t = -0.6001, df = 35, p-value = 0.5523

alternative hypothesis: true mean is not equal to 11.6

95 percent confidence interval:

 11.16164 11.83834

sample estimates:

mean of x

 11.49999

1. Ho:μ=11.6
Ha:μ≠11.6
2. = (11.5 - 11.6)/(1/6) = -0.6
3. ±tα/2, n-1 = ±2.03
4.



Because –2.03 < -0.6 < 2.03, do not reject Ho.

1. There is not sufficient evidence to prove the volleyballs are being made incorrectly.

OR

There is not sufficient evidence to conclude the population mean radius is different from 11.6.

Extracting results from an object created by t.test():

> save.results <- t.test(x = volleyball$radius, alternative

 = "two.sided", mu = 11.6, conf.level = 0.95)

> names(save.results)

 [1] "statistic" "parameter" "p.value" "conf.int"

 [5] "estimate" "null.value" "stderr" "alternative"

 [9] "method" "data.name"

> save.results$statistic

 t

-0.6000647

> save.results$null.value

mean

11.6

To help better understand the test statistic method, suppose we had a different  and everything else remained the same. Below is a table showing what would happen with the hypothesis test.

| **Case**  |  | **t** | **Decision** |
| --- | --- | --- | --- |
| 1 | 11.2 | -2.4 | Reject Ho |
| 2 | 11.4 | -1.2 | Don't Reject Ho |
| 3 | 11.6 | 0 | Don't Reject Ho |
| 4 | 11.8 | 1.2 | Don't Reject Ho |
| 5 | 12.0 | 2.4 | Reject Ho |

See hyp\_1sample\_pic.xlsx (replace “x\_bar” with “ybar” in the picture):



All values in red can be changed by the user to see the effect on the test statistic, critical values, and the hypothesis test decision. Make changes on your own so that you familiarize yourself with what happens if the sample size increases, standard deviation changes, …

P-values will be discussed later in this chapter.

Notes:

* Hypothesis testing is set up is to try to find evidence (through a sample) against the null hypothesis (Ho). If enough evidence is found, we can conclude that the alternative hypothesis (Ha) to be true the where type I error probability is α. Because β (probability of type II error) is not controlled, we cannot set up hypothesis testing to go the other way.
* The type I error probability needs to be interpreted in a similar manner as the confidence level for a confidence interval. Thus, if the sampling and testing process was repeated many times and Ho was true, we would expect α of them to incorrect reject the null hypothesis.
* Notice that in the formula for t we put in the hypothesized value of μ, μ0. We assume the null hypothesis to be true by doing this (remember the jury trial example). We put values from the sample (, s, n) into the test statistic to see if the sample mean is far enough from the hypothesized mean to conclude that the null hypothesis is incorrect.
* Why was Ho:μ=11.6 vs. Ha:μ≠11.6?
	+ In order for the theory behind all of this to work, we need the equal sign in Ho.
	+ If some kind of new "action" is to be taken when a hypothesis is proved to be true, this hypothesis typically should be in Ha. This is because we can control the probability of making an error in our decision (i.e., α is specified).
		- What would happen if Mikasa's volleyballs did not have an average radius of 11.6? Production of volleyballs would be stopped and the manufacturing process would be investigated to find the problem. This implies we should use Ha:μ≠11.6.
		- What would happen if Mikasa's volleyballs have an average radius of 11.6? The production of volleyballs would continue.
* Some textbooks use the normal distribution instead of the t distribution when n ≥ 30 and σ is known.
	+ The same problems with using the normal distribution version of the CI occur here; σ is unknown in real-life applications.
	+ Remember that for large samples (n ≥ 30) the t distribution is approximately a standard normal distribution.
	+ **In this course, we will only use the t distribution!**

3) The p-value method – 5 steps

The test statistic method compared t and the critical values from the t distribution. The p-value method compares probabilities. Thus, we have “p”-value method.

1. State Ho and Ha
2. Find the p-value

The p-value is 2×P(T > |t|). This can be calculated in R using

2\*(1 - pt(q = abs(t), df = n - 1))

where t is the observed test statistic value and n is the sample size.

Below is a graph showing part of the probability.

Be careful with the notation in some textbooks. If a book does not use an uppercase T to represent a random variable, they have a notational problem with representing a p-value as “2×P(t > |t|)”. To get around this problem, they will often use something like “tc” or “computed t” to represent the observed value of the test statistic.

What is the p-value giving us?

The p-value gives the probability of finding a value of |t| at least this great assuming the null hypothesis is true. Thus, it is a measure of how extreme the test statistic is relative to the t distribution. This is the same idea as comparing t to critical values.

The probability, P(T > |t|), is multiplied by two because the disagreement between the data and Ho can be in two directions; i.e., on **two** **sides** (**tails)** of the probability distribution.

The p-value is just a probability found through integration! For this type of hypothesis test, the p-value is



where I use “u” as the variable we are integrating over to differentiate from the observed value t of the test statistic.

1. State α
2. Reject or do not reject Ho
* Reject Ho if p-value < α
* Don’t reject Ho if p-value ≥ α

As mentioned earlier, α is sometimes called "the level of significance" due to it being the threshold here.

5. Conclusion – Describe what 4. means in terms of the original problem.

Example: Volleyball quality control (volleyball.R, volleyball.csv, and hyp\_1sample\_pic.xlsx)

> mu0 <- 11.6

> t <- (ybar - mu0)/(s/sqrt(n))

> 2\*(1 - pt(q = abs(t), df = n - 1))

[1] 0.5523286

> t.test(x = volleyball$radius, alternative = "two.sided",

 mu = 11.6, conf.level = 0.95)

 One Sample t-test

data: volleyball$radius

t = -0.6001, df = 35, p-value = 0.5523

alternative hypothesis: true mean is not equal to 11.6

95 percent confidence interval:

 11.16164 11.83834

sample estimates:

mean of x

 11.49999

1. Ho: μ=11.6
Ha: μ≠11.6
2.  = (11.5 - 11.6)/(1/6) = -0.6
2×P(T > |-0.6|) = 2×P(T > 0.6) = 2×0.2762 = 0.5523

where ν = n – 1 = 35. The probability of observing a test statistic value this great in magnitude, |-0.6|, is 0.5524 if μ=11.6 was true. Therefore, this is a likely event to happen if μ=11.6.

Another way to think about the p-value is the following: If μ really was 11.6, then a test statistic value, t, at least this large in absolute value (0.6) would occur about 55% of the time if the hypothesis test process (take a new sample and perform a new hypothesis test) is repeated a very large number of times. In other words, this is likely to occur if μ = 11.6. Thus, μ could be 11.6 because this is a likely event.

1. α = 0.05
2. Because 0.5524 > 0.05, do not reject Ho



α/2=
0.025

1. The sample does not provide enough evidence to suggest that the volleyballs are being made with the wrong radius.

OR

There is not sufficient evidence to conclude the population mean radius is different from 11.6.

Fill in the p-values for the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Case**  |  | **t** | **Decision** | **p-value** |
| 1 | 11.2 | ‑2.4 | Reject Ho |  |
| 2 | 11.4 | ‑1.2 | Don't Reject Ho |  |
| 3 | 11.6 | 0 | Don't Reject Ho |  |
| 4 | 11.8 | 1.2 | Don't Reject Ho |  |
| 5 | 12.0 | 2.4 | Reject Ho |  |

Examine hyp\_1sample\_pic.xlsx again to see what happens if the sample size increases, standard deviation changes,…

Make sure you can do the hypothesis test problems with EACH method. Remember all three hypothesis test methods give the same answers for tests involving μ and the t distribution.

Understanding the type I error probability (repeat)

Suppose Ho is true and the type I error rate is denoted by α. If the hypothesis testing procedure is repeated R times (take a new sample and perform a new hypothesis test), we would expect R×α of the hypothesis tests to incorrectly reject Ho.