MATH 1051H 2020FA - Lecture 15

Inference for other estimators

Non-normal point estimates

- We may apply the ideas of hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:
  - the sample size is too small for the normal approximation to be valid;
  - the standard error estimate may be poor; or
  - the point estimate tends towards some distribution that is not the normal distribution.

Non-normal point estimates

- We may apply the ideas of hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:
  - the sample size is too small for the normal approximation to be valid;
  - the standard error estimate may be poor; or
  - the point estimate tends towards some distribution that is not the normal distribution.
- For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for hypothesis testing to these alternative distributions.

When to retreat

- Statistical tools rely on the following two main conditions:
  - **Independence**: A random sample from less than 10% of the population ensures independence of observations. In experiments, this is ensured by random assignment. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.
  - **Sample size and skew**: For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.
- Whenever conditions are not satisfied for a statistical technique:
  - Learn new methods that are appropriate for the data.
  - **Consult a statistician.**
  - **Ignore the failure of conditions.** This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Practice

All else held equal, will the p-value be lower if $n = 100$ or $n = 10000$?

1. $n = 100$
2. $n = 10000$

Practice

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

1. $n = 100$
2. $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu > 49.5$.

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.0062$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

As n increases - $SE \downarrow$, $Z \uparrow$, p-value $\downarrow$

Demonstration

Imagine we test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.10	10.2
$n = 30$	p = 0.45	p = 0.39	p = 0.29
$n = 5000$			

Demonstration

Imagine we test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.10	10.2
$n = 30$	p = 0.45	p = 0.39	p = 0.29
$n = 5000$	p = 0.04	p = 0.0002	$p \approx 0$

Demonstration

Imagine we test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.10	10.2
$n = 30$	p = 0.45	p = 0.39	p = 0.29
$n = 5000$	p = 0.04	p = 0.0002	$p \approx 0$

When n is large, even small deviations from the null (small **effect sizes**), which may be considered practically insignificant, can yield statistically significant results.

Statistical vs. practical significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (**effect size**), even when the difference is not practically significant.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).
- The role of a statistician is not just in the analysis of data, but also in planning and design of a study.

Aphorism

“To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.” - Sir Ronald Aylmer Fisher, Address to Indian Statistical Congress (1938)

It’s ok to need help: you’re not statisticians, and you won’t be at the end of this class!

The *t* Distribution

A Specific Example of Non-Normal

In the last discussion, we talked about how we might approach certain problems where the **normal assumption** does not hold. We're now going to start looking at a specific, famous example of this kind of problem - the *t* distribution.

Example

Mercury in seafood due to pollution is a known problem, especially in heavy industrial areas, although mercury has spread a long way from explicit polluters. Japan as a country consumes a large amount of seafood, and researches were interested in the average mercury content in Rossi's dolphins from the Taiji area. They analyzed 19 dolphins' muscles for mercury content.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Measurements are in micrograms of mercury per wet gram of muscle (μg/wet g).

So, a "begged question": could we do a hypothesis test on this data using what we know so far (e.g., a Z distribution)?

Review: Purpose of Large Sample

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that:

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error (SE), as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size as long as the population distribution is nearly normal.

While this is a helpful special case, it's inherently difficult to verify normality in small data sets.

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.

For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

In Context (Dolphins)

How big is our sample? And, given the summary, how symmetric is our data?

- only 19 samples
- no population σ
- data seems mostly symmetric

The t Distribution

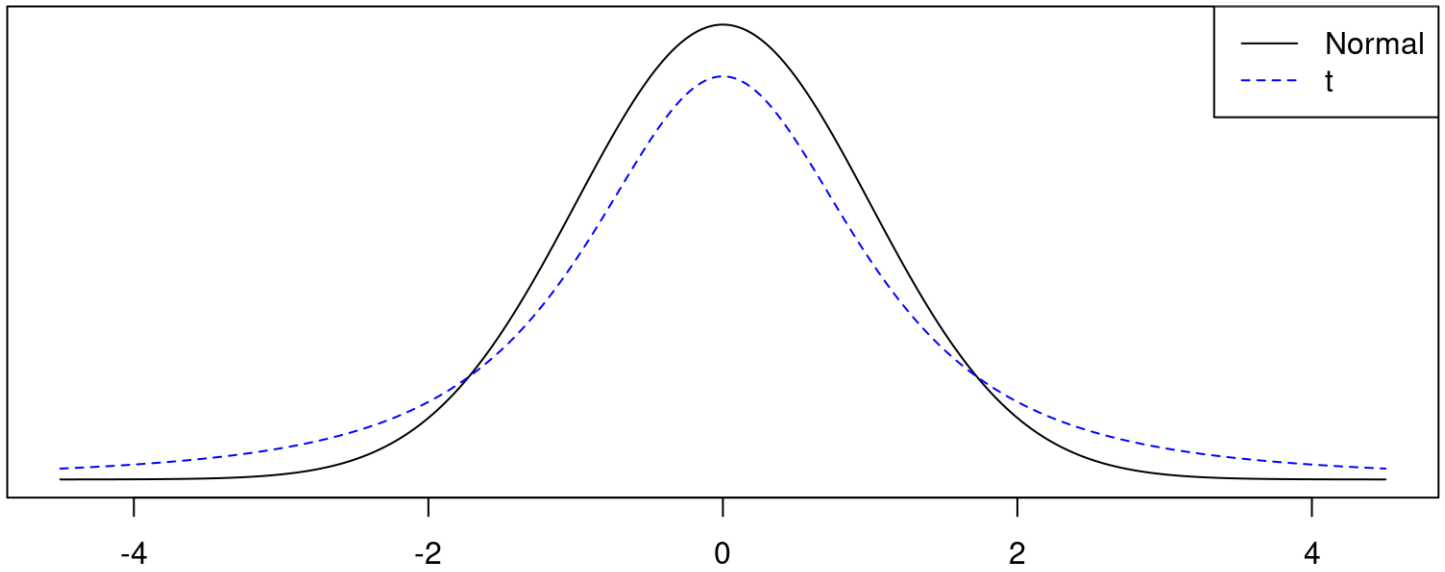
When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.

This distribution also has a bell shape, but its tails are thicker than the normal model's.

Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)

A plot of t versus $\mathcal{N}$



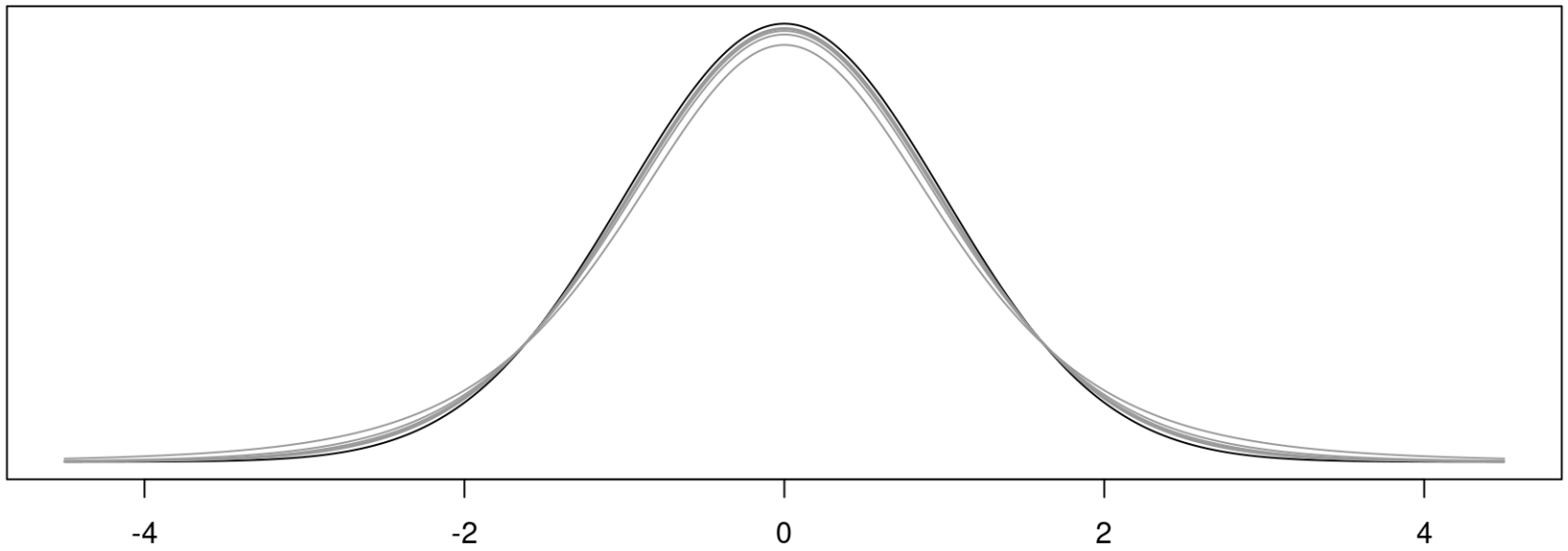
The t Distribution (ctd.)

Always centered at zero, like the standard normal (z) distribution.

Has a single parameter: degrees of freedom (df) – like χ^2.

What happens to the shape of the t distribution as df increases?

The t Distribution (ctd.)



As $df \longrightarrow \infty$, the t distribution approaches the normal!

Asymptotic

What df is required to give arbitrary decimal agreement between the t and z curves? (on a restricted domain)

- 2 decimals: $df = 14$
- 3 decimals: $df = 136$
- 4 decimals: $df = 1370$

What do we usually ask for? 30 df corresponds to 3 decimals for the $[-3, 3]$ domain, which is good enough. So once $df > 30$, people often just use a z instead.

Recap: Inference using a small sample mean

If $n < 30$, sample means follow a t distribution with $\text{SE} = \frac{s}{\sqrt{n}}$, **unless** you are positive you know the population standard deviation σ.

1. **Conditions**:
   - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
   - $n < 30$ and no extreme skew
2. **Hypothesis Testing**:

$$t_{\text{df}} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1.$$

Back to the Dolphins

Researchers want to know if the average mercury content in these dolphins exceeds 4 μg/wet g. Perform a hypothesis test to answer this question.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Hypothesis Test

$H_0 : \mu \leq 4 \qquad \text{versus} \qquad H_A : \mu > 4$

Conditions:

- we assume independence of observations
- $n < 30$
- don't know σ

So we need to use the t!

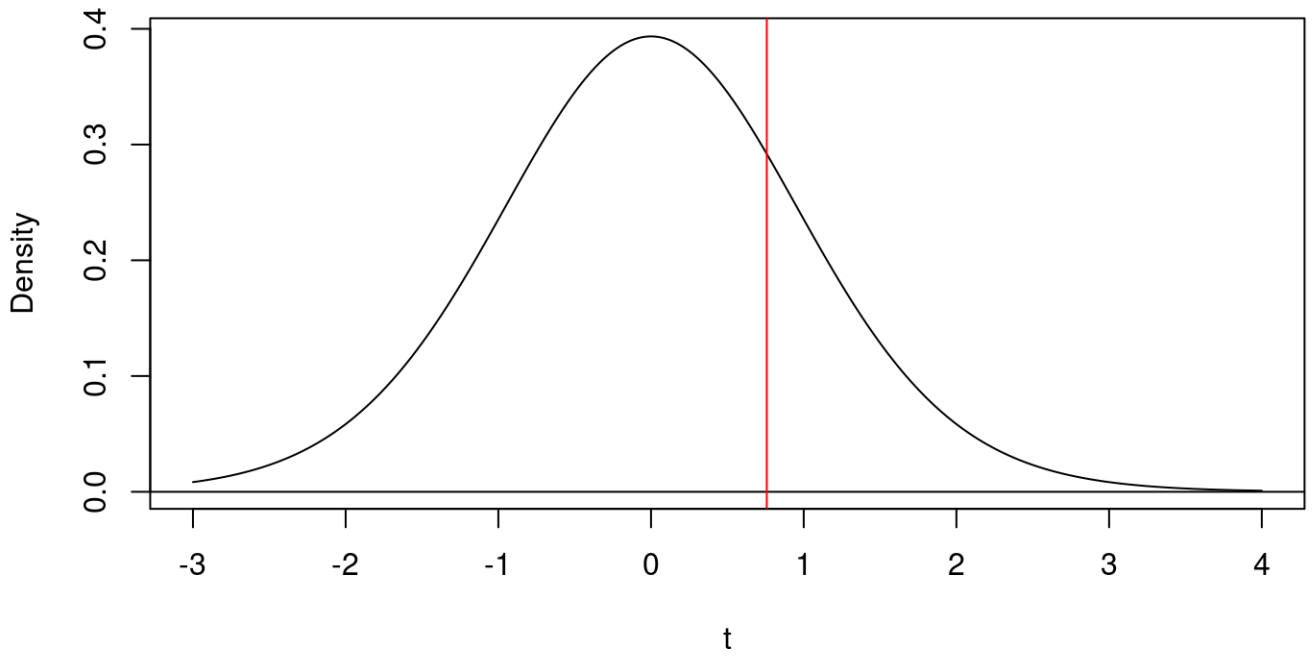
Test Statistic

The test statistic doesn't change: it's a test statistic on the mean, so the **statistic** stays the same! The only difference is that the random variable we get is no longer assumed to be a Z ... instead, it's a t.

$$t = \frac{\bar{x} - \mu_0}{\text{SE}_{\bar{x}}} = \frac{4.4 - 4}{\frac{2.3}{\sqrt{19}}} = 0.7581.$$

Computing the p-value

Our alternative is $\mu > 4$, so our p-value goes **up** ...



p-value

But there's a trick … when you specify a *t* distribution, you don't specify mean/SD … you have to specify the **degrees-of-freedom** (df). Our rule for the mean is that df is **n-1**: one less than the number of samples you have.

```
pt(q = 0.7581, df = 19 - 1, lower.tail = FALSE)
```

```
## [1] 0.2291013
```

Thus our p-value is 0.229, which is "big", so we would **fail to reject the null** hypothesis, meaning we cannot conclude that the mean is greater than 4 μg/wet g.

Summary of Tests on Means for t Distributions

- Similar assumptions to the Z
- Same test statistic
- Notice the df argument (n-1)
- Same way of computing p-value, except use **pt()** not **pnorm()**
- Same interpretation

Linear Regression and Hypothesis

Linear Regression

$$\hat{y} = \beta_0 + \beta_1 x$$

Notation:

- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

Note

Both parameters have **point estimates** … these are statistics! That means we can do hypothesis tests on them!

Hypothesis Tests for Linear Regression

Most of the time, we only do hypothesis tests for linear regression on the **slope** - specifically, on β_1 as the parameter.

So what is the hypothesis?

What do we say about the null … default, base, nothing going on, nothing to see, do not pass go …

Null Hypothesis for Slopes

Remember what a slope is: it's actually a representation of the relationship between two variables (like correlation). So our default is that there **is no relationship**. What does this correspond to? If there's no relationship, that corresponds to correlation 0 … which is also slope 0!

Thus

$$H_0: \beta_1 = 0$$

is our null hypothesis.

Alternative Hypothesis for Slopes

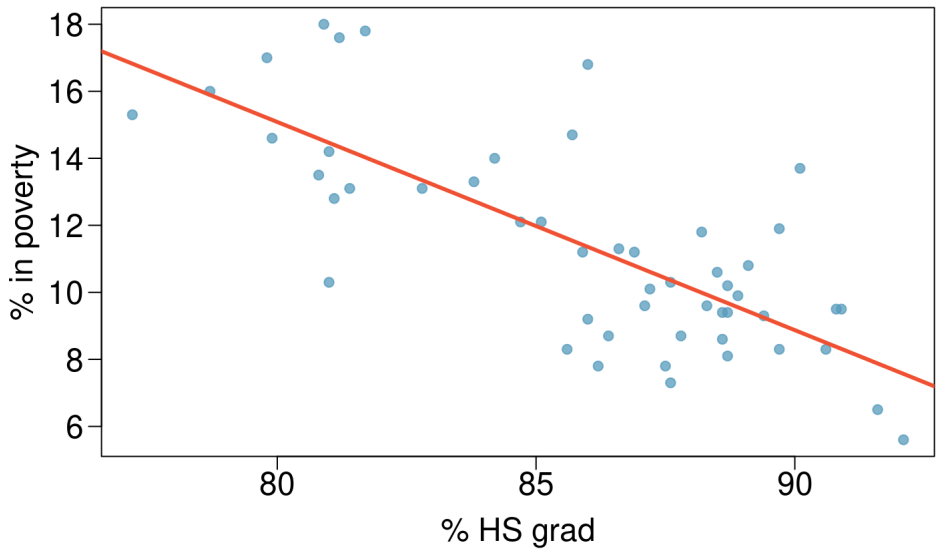
So what is the alternative hypothesis? We only ever care about one:

$H_A: \beta_1 \neq 0.$

So now we can perform hypothesis tests inside linear regressions!

Let's Return to our Previous Example

USA states (and DC), with percent of population in poverty, and percent of population that graduated from high school.



Poverty - HS Grads

So what is our hypothesis in this problem? Let's start with words:

Null Hypothesis: there is no relationship between the percentage of the population that graduate from high school, and the percentage of the population living in poverty.

Alternative Hypothesis: there is a relationship between these two variables.

Poverty - HS Grads

Now, translate this to symbols:

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_A : \beta_1 \neq 0.$$

Now, how do we **do** this? We can't do the test statistic like we normally do ... but we don't have to!

Poverty - HS Grads - Doing the Test

Take a close look at the row starting with **Graduates**: what do you see?

```
mod <- lm(Poverty ~ Graduates, data = poverty)
summary(mod)
```

```
##
## Call:
## lm(formula = Poverty ~ Graduates, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1624 -1.2593 -0.2184  0.9611  5.4437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
## Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.082 on 49 degrees of freedom
## Multiple R-squared:  0.5578, Adjusted R-squared:  0.5488
## F-statistic: 61.81 on 1 and 49 DF,  p-value: 3.109e-10
```

Poverty - HS Grads - Doing the Test

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Graduates	-0.62122	0.07902	-7.862	3.11e-10 ***

So we have:

- point estimate: -0.62122
- SE: 0.07902
- test statistic: $t = -7.862$
- p-value: $p = 3.11e-10$ (very, very small!)

So what's our conclusion?

Poverty - Conclusions

Since we find an extremely small p-value (smaller than $\alpha = 0.05$ for sure), we **reject the null hypothesis**, and conclude that there **is** a relationship between the percentage of the population graduating from high school, and the percentage of the population living in poverty. We estimate $b_1 = -0.621$.

So … Hypotheses on Linear Models

So we can estimate slopes, then do hypothesis tests on them, which lets us determine if we believe there are associations (or “relationships”) between them. And we don’t have to do much … just fit a model in R, and then read the answer.

Connections between t-tests and Linear Regression

Why are they the same thing?

So why are t-tests and linear regression the same thing?

- t-test: we're comparing a set of data to some value μ_0
- regression: we have a row we haven't considered, labeled (Intercept)

So, this actually means that

$$H_0 : \mu = \mu_0 \qquad (\text{t-test on mean})$$

is equivalent to

$$H_0 : \beta_0 = \mu_0 \qquad (\text{t-test on intercept - mean!})$$

In other words, a t-test it's our linear model $y = \beta_0 + \beta_1 x$ where the slope term is gone since there is no x.

Let's Try an Example

Consider the dolphin example from before, with some specific data (not exactly the same). We are interested in $H_0 : \mu = 4.0$.

```
dat <- c(2.56, 3.86, 5.66, 5.95, 7.67, 1.92, 10.01, 6.00,
         4.47, 6.83, 1.47, 1.02, 8.36, 3.36, 5.34, 1.75,
         4.55, 4.78, 4.91)
mod_a <- t.test(x = dat, alternative = "two.sided", mu = 4)
mod_b <- lm((dat - 4) ~ 1)
```

Fitting using the t.test() function

```
mod_a
```

```
##
##  One Sample t-test
##
## data:  dat
## t = 1.3611, df = 18, p-value = 0.1903
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.586035 5.937123
## sample estimates:
## mean of x
##  4.761579
```

Fitting using the Linear Model

```
summary(mod_b)
```

```
##
## Call:
## lm(formula = (dat - 4) ~ 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7416 -1.8016  0.0184  1.2134  5.2484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7616     0.5595   1.361     0.19
##
## Residual standard error: 2.439 on 18 degrees of freedom
```

So ...

That (Intercept) piece we ignored from our previous section turns out to be a t-test on the mean, when considered without a slope! Neat!

Coming Up Next

In our coming lectures, we will explore additional hypothesis testing ideas, and then discuss a very important idea called “confidence intervals” that are related to hypothesis tests, and used for describing estimates!