MATH 1052H - S62 - Lecture 02

Confidence intervals

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like **fishing with a spear** in a murky lake, and using a confidence interval is like **fishing with a net**.
- We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

So the analogy: if we report a point estimate, we probably won’t hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$\text{point estimate} \pm 2 \times \text{SE}$$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

$$\begin{aligned} \bar{x} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \\ &= (2.7, 3.7) \end{aligned}$$

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

- the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.
- college students on average have been in between 2.7 and 3.7 exclusive relationships.
- a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.
- 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

- the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.
- *college students on average have been in between 2.7 and 3.7 exclusive relationships.*
- a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.
- 95% of college students have been in 2.7 to 3.7 exclusive relationships.

A more accurate interval

Confidence interval, a general formula

$$\text{point estimate} \pm z^{\star} \cdot SE$$

Conditions when the point estimate = $\bar{x}$:

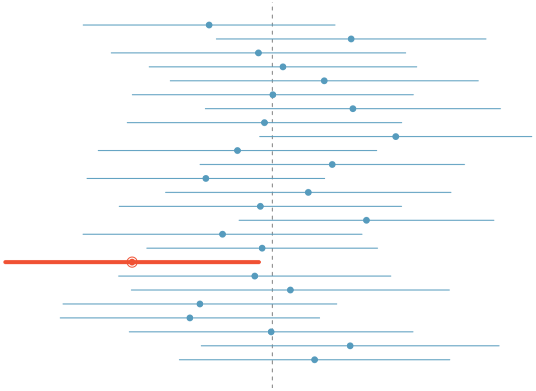
- **Independence:** Observations in the sample must be independent
  - random sample/assignment
  - if sampling without replacement, $n <$ 10% of population
- **Sample size / skew:** $n \geq 30$ and population distribution should not be extremely skewed

Note: We will discuss working with samples where $n < 30$ later.

Capturing the population parameter

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $\text{point estimate} \pm 2 \cdot SE$.
- Then about 95% of those intervals would contain the true population μ.
- The figure to the right shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



Width of an interval

If we want to be more certain that we capture the population parameter, *i.e.*, increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

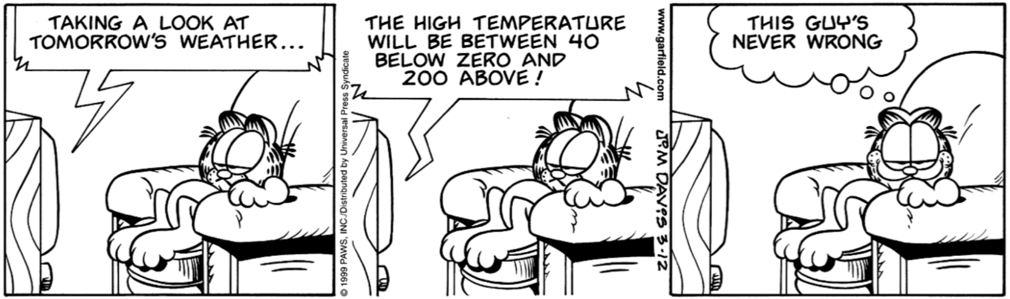
Can you see any drawbacks to using a wider interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVPS 3-12

If the interval is too wide it may not be very informative.

http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the confidence level

$$\text{point estimate} \pm z^{\star} \cdot SE$$

- In a confidence interval, $z^{\star} \cdot SE$ is called the **margin of error** (ME), and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^{\star}$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^{\star} = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^{\star}$ for any confidence level.

Practice

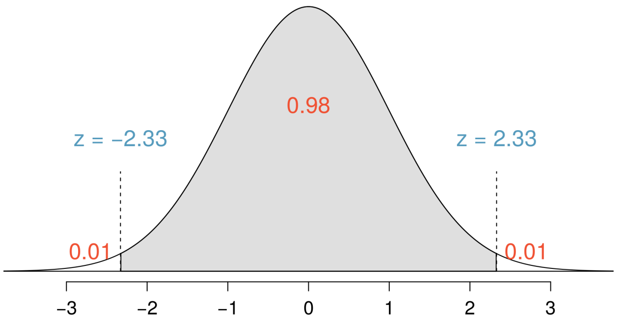
Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

- $Z = 2.05$
- $Z = 1.96$
- $Z = 2.33$
- $Z = -2.33$
- $Z = -1.65$

Practice

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

- $Z = 2.05$
- $Z = 1.96$
- $Z = 2.33$
- $Z = -2.33$
- $Z = -1.65$



Testing Hypotheses: CIs

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships?

- The associated hypotheses are:
  - H_0: $\mu = 3$: College students have been in 3 exclusive relationships, on average
  - H_A: $\mu > 3$: College students have been in more than 3 exclusive relationships, on average

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships?

- The associated hypotheses are:
  - H_0: $\mu = 3$: College students have been in 3 exclusive relationships, on average
  - H_A: $\mu > 3$: College students have been in more than 3 exclusive relationships, on average
- Since the null value is included in the interval, we do not reject the null hypothesis in favor of the alternative.
- This is a quick-and-dirty approach for hypothesis testing. However it doesn't tell us the likelihood of certain outcomes under the null hypothesis, i.e., the *p*-value, based on which we can make a decision on the hypotheses.

Summary

Confidence intervals for the population mean μ from large samples have the form

$$\bar{x} \pm \mathrm{ME} = \bar{x} \pm z^{\star} \cdot \mathrm{SE}$$

and explicitly, the Standard Error, SE, is

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}.$$

# Confidence Intervals versus Hypothesis Testing

We have slightly different **Success-Failure Conditions** (for number of samples required):

- **CI**: at least 10 *observed* successes and failures
- **HT**: at least 10 *expected* successes and failures, under the null assumption

Confidence Intervals versus Hypothesis Testing

Standard Error

- **CI**: calculated using the sample proportion, $\hat{p}$

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}.$$

- **HT**: calculated using the null hypothesis value, p_0

$$\mathrm{SE} = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Practice

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.80 \qquad\qquad H_A : p > 0.80$$

(we use a one-tailed hypothesis test because that is our question: “more than 80%”)

Practice

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.80 \qquad\qquad H_A : p > 0.80$$

$$\text{SE} = \sqrt{\frac{0.80(0.20)}{670}} = 0.0154$$

Practice

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.80 \qquad\qquad H_A : p > 0.80$$

$$\text{SE} = \sqrt{\frac{0.80(0.20)}{670}} = 0.0154$$

$$Z = \frac{0.85 - 0.80}{0.0154} = 3.25$$

Practice (*p*-value)

Compute the *p*-value for this value of z.

Our Method: Use R!

```
1 - pnorm(q = 3.25)
```

```
## [1] 0.000577025
```

```
pnorm(q = 3.25, lower.tail = FALSE)
```

```
## [1] 0.000577025
```

So the *p*-value is 0.0006.

Practice: Conclusion

Since the *p*-value is low, we reject H_0. The data provide convincing evidence that more than 80% of Americans have a good intuition on experimental design.

Practice Question 2

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: “More than 10% of all Americans have objections on religious grounds to celebrating Halloween.” At 95% confidence level, is this news piece's statement justified?

1. Yes
2. No
3. Can't tell

Practice Question 2

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is ± 3%. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

```
c(11-3, 11+3)
```

```
## [1]  8 14
```

1. Yes
2. *No* (10 is between 8 and 14)
3. Can't tell

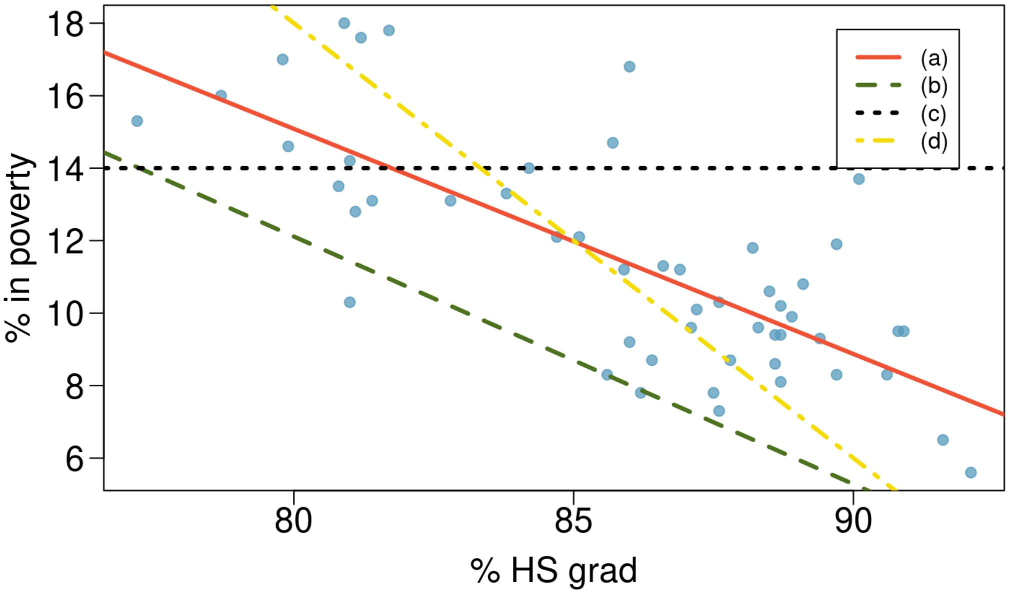
Linear Regression and Hypothesis

Objective

In science, we often obtain data from experiments or observational studies, and then are interested in **modeling** it. How we model it varies from field to field, but underlying many of the models used in science are **statistical models**. In this lecture, we will be examining one of the foundational models used all through every scientific field.

We will begin by quantifying the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable. The model is therefore “find the relationship between two variables”.

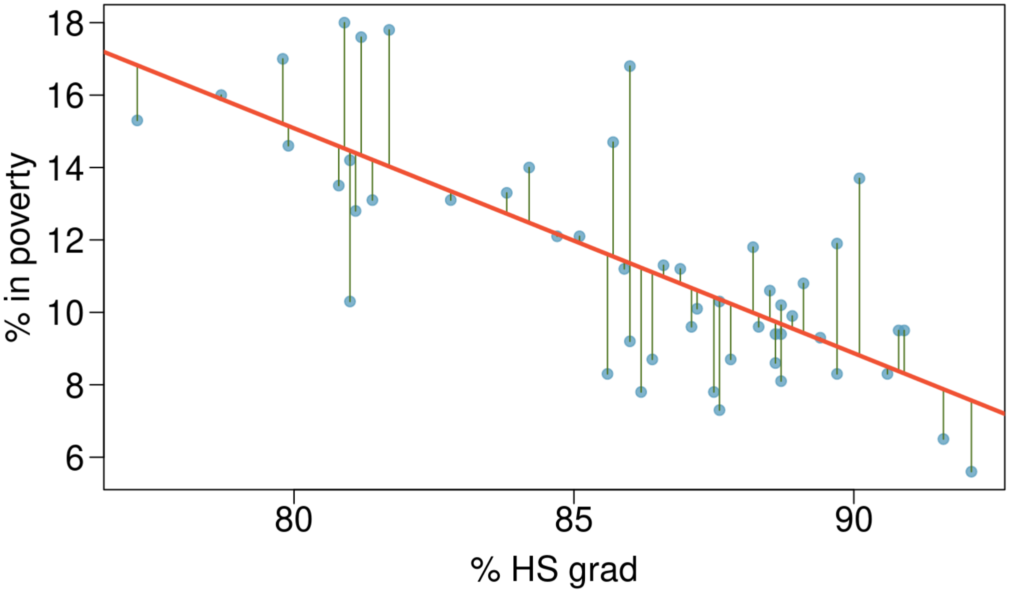
Eyeballing the line



Which of the lines on the figure appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

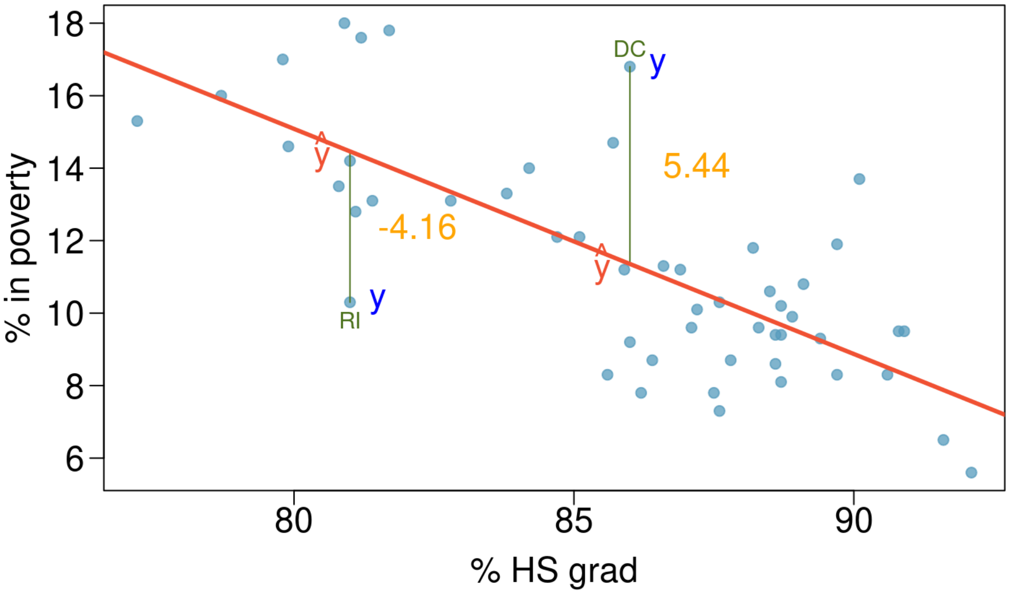


Residuals (cont.)

Formally, residuals are the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$

Specific Residuals



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

A measure for the best line

- We want a line that has small residuals:
  - Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squared residuals – **least squares**

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

  - Option 3 … 100: actual topics of research!
- Why least squares?
  - Most commonly used
  - Easier to compute by hand and using software
  - In many applications, a residual twice as large as another is usually more than twice as bad

The least squares line

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

Notation:

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

The least squares line

```r
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
```

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

$$H_0\colon \beta_1 = 0$$

is our null hypothesis.

Alternative Hypothesis for Slopes

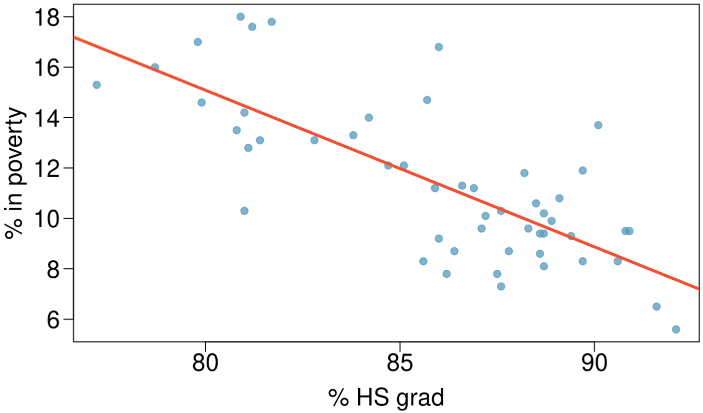
So what is the alternative hypothesis? We only ever care about one:

$H_A\colon \beta_1 \neq 0.$

So now we can perform hypothesis tests inside linear regressions!

Let's Consider an 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
```

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!

Summary

Where are we now?

As of the end of the second lecture, we have now done a whirlwind review of the key material of MATH 1051H that you are expected to remember, with references to the earlier probability material (such as the normal and t distributions, and probabilities under their curves).

What chapters have been covered?

If you want to review in the textbook, the following chapters are considered to be “done” and you are expected to be able to reference the material from them:

- Chapter 5
- Chapter 6.1
- Chapter 7.1
- Chapter 8.1-8.3

Our semester will cover the rest of Chapters 6, 7, and 8, and also the material of Chapter 9. In addition we will have a special unit toward the end of the semester where we cover some material not in the textbook aside from brief mentions in Chapter 1 (designing statistical experiments).