# Psych 100A Spring 2019: Week 9 Slides

Amanda Montoya

May 21, 2019

```r
NLSdata <- read.csv("http://bit.ly/NLSdata", header = TRUE)
Cohab.model <- lm(HrsSleep2009~Cohab2009, data = NLSdata)
```

# Learning Outcomes Today

- Compare t-distribution approach for confidence intervals to other approaches
- Apply the ideas of Occam's Razor to model comparison in statistics
- Describe the use of simulation for model comparison

# Simulation: Pros and Cons

Pros:

- Have complete control over elements of the population (could change anything we want)
- Performs well (gets the right answer a lot of the time)

Cons:

- Creating a population can be pretty intense
- Rely on assumption about the shape of the distribution of the errors
- Interval not guaranteed to be symmetric
- Answer will be a little different every time when we rerun code

# Bootstrapping: Pros and Cons

Pros:

- Using the sample as representation of population makes simulation easier
- No assumptions about the shape of distribution of the errors
- Performs well (gets the right answer a lot of the time)

Cons:

- Interval not guaranteed to be symmetric
- Answer will be a little different every time when we rerun code

# Normal Distribution: Pros and Cons

Pros:

- VERY EASY! (thanks to R)
- Interval guaranteed to be symmetric
- Answer is always the same

Cons:

- Central Limit Theory doesn't always work for small samples (no exact definition of how small is too small)
- Doesn't take into account that the standard error is an estimate (not a population value)

# T-distribution: How

- Use a t-distribution with mean $= 0$ and $df = n - 2$ from sample to approximate sampling distribution (centered around 0)
- Calculate points which denote the top and bottom 2.5% (qt), to get $t_{critical}$
- Combine estimate and MOE to get a confidence interval

# T-statistic

t-distribution takes into account the fact that the standard deviation of the sampling distribution $\sigma_{b_1}$ is not known, but rather estimated.

t-distribution is not a distribution of means, but rather a distribution of t-values (this part is ignored in the book). Anything where we take an estimate and divide by it's estimated standard error is called a t-statistic

$$t = b_1/s_{b_1}$$

```r
summary(Cohab.model)
```

```
##
## Call:
## lm(formula = HrsSleep2009 ~ Cohab2009, data = NLSdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8119 -0.8119  0.1881  1.1881 11.1881
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.61368    0.03371  196.17  < 2e-16 ***
## Cohab2009TRUE  0.19825    0.04191    4.73 2.32e-06 ***
```

## T-distribution

Since the estimate of the standard error depends on the number of degrees of freedom, the t-distribution also depends on the degrees of freedom.

Fewer degrees of freedom (smaller sample size) means more variability in distribution

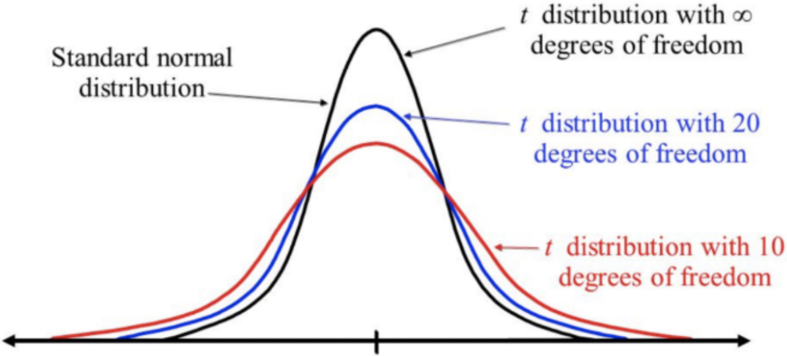
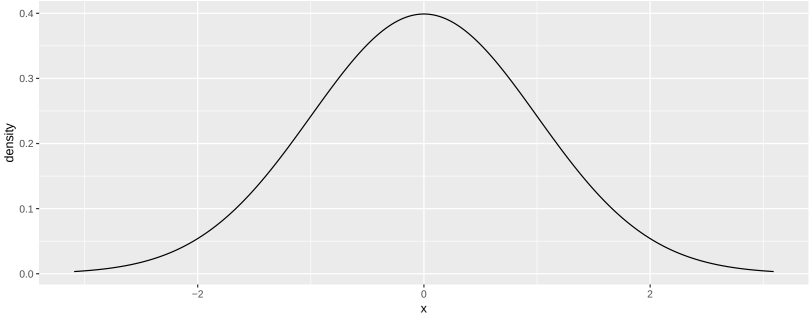


Figure 1: t-distribution

# T-distribution: How

- Use a t-distribution with mean $= 0$ and $df = n - 2$ from sample to approximate sampling distribution (centered around 0)

```
gf_dist("t", params = (4224-2))
```



# T-distribution: How

- Calculate points which denote the top and bottom 2.5% (qt), to get $t_{critical}$

$$b_1/s_{b_1} \pm t_{critical}$$

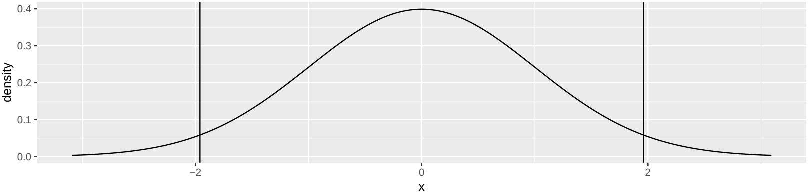
```
qt(0.025, 4224-2)
```

```
## [1] -1.960526
```

```
qt(0.975, 4224-2)
```

```
## [1] 1.960526
```

```
gf_dist("t", params = (4224-2))%>%
  gf_vline(xintercept = qt(0.025, 4224-2))%>%
  gf_vline(xintercept = qt(0.975, 4224-2))
```



# T-distribution: How

- Combine estimate and MOE to get a confidence interval

$$b_1/s_{b_1} \pm t_{critical}$$

$$b_1 \pm s_{b_1} \times t_{critical}$$

```
b1(Cohab.model) + 0.04191*qt(0.975, 4224-2)
```

```
## [1] 0.2804118
```

```
b1(Cohab.model) - 0.04191*qt(0.975, 4224-2)
```

```
## [1] 0.1160805
```

```
confint(Cohab.model)
```

```
##                    2.5 %    97.5 %
## (Intercept)    6.5475863 6.6797779
## Cohab2009TRUE 0.1160757 0.2804167
```

# T-distribution: Pros and Cons

Pros:

- VERY EASY! (thanks to R)
- Interval guaranteed to be symmetric
- Answer is always the same
- Takes into account that the standard error is an estimate (not a population value)

Cons:

- Central Limit Theory doesn't always work for small samples (no exact definition of how small is too small)

# All intervals together

```
#simulation
c(sim.ll, sim.ul)
```

```
## [1] 0.1148916 0.2791393
```

```
#bootstrap
c(boot.ll, boot.ul)
```

```
## [1] 0.1147754 0.2798488
```

```
#normal distr
confint.default(Cohab.model)
```

```
##                   2.5 %    97.5 %
## (Intercept)   6.5476053 6.6797589
## Cohab2009TRUE 0.1160992 0.2803931
```

```
#t-dist
confint(Cohab.model)
```

```
##                   2.5 %    97.5 %
## (Intercept)   6.5475863 6.6797779
## Cohab2009TRUE 0.1160757 0.2804167
```

# Making Conclusions

Regardless of the method, we would conclude that $\beta_1 = 0$ is not among the population means for which our data are likely.

INFERENCE: This means, that we can safely conclude our data came from a positive $\beta_1$ (i.e., We believe that there is a difference between sleep for cohabitating people (cohab $>$ no cohab))

# Model Comparison

Goal: Decide between two models (simple and complex) to represent our *current understanding* of the population/DGP.

Research Question: Does cohabitating with your romantic partner predict hours of sleep? (Notice we don't use causal language because the data is correlational.)

Statistical Question: Which represents the data better: simple or complex model?

Complex Model: $Y_i = b_0 + b_1 X_i + e_i$

Simple Model: $Y_i = b_0 + 0 \times X_i + e_i = b_0 + e_i$

# Model Comparison

https://www.youtube.com/watch?v=M5WDdvkFaDg

# Confidence Interval Approach

Research Question: Does cohabitating with your romantic partner predict hours of sleep?

- Translating to statistics: $\beta_1 \neq 0$
- Create a sampling distribution of $b_1$ using simulation, bootstrapping, or t-distribution (`confint()`)
- Cut off tails of the sampling distribution of $b_1$ to create 95% confidence interval
- Evaluate if 0 is contained in interval

# Model Comparison

When we're curious about comparing two specific models, we can compare them statistically using the F-value (or PRE).

Model comparison is a very general approach which can be applied outside of situations that confidence intervals can be applied in.

- Confidence intervals and model comparison result in the same information and decision when comparing models which differ by only 1 degree of freedom (compare one parameter against 0)
- Model comparison can be used more generally to compare models of varying complexity (e.g., 3 group model) where confidence intervals can't always be used
- Model comparison also provides an exact probability of our observed statistics under the simple model (not provided in CIs)

# Cohab Model

```r
supernova(Cohab.model)
```

```
##  Analysis of Variance Table (Type III SS)
##  Model: HrsSleep2009 ~ Cohab2009
##
##                                SS   df     MS      F    PRE     p
##  ----- ----------------- -------- ---- ------ ------ ------ -----
##  Model (error reduced) |   37.914    1 37.914 22.373 0.0053 .0000
##  Error (from model)    | 7154.812 4222  1.695
##  ----- ----------------- -------- ---- ------ ------ ------ -----
##  Total (empty model)   | 7192.726 4223  1.703
```

Based on the supernova output, cohabitation explains 0.5% of the variability in Hours of Sleep (PRE), or 22.373 sums of squares per degree of freedom.

Are these values big? Are they small?

Are these values likely to occur even if there is no relationship between Cohabitation and Sleep? (i.e., $\beta_1 = 0$)

# Sampling Distributions of F and PRE

F and PRE are both sample statistics, which means if we collected another random sample they would have a slightly different value.

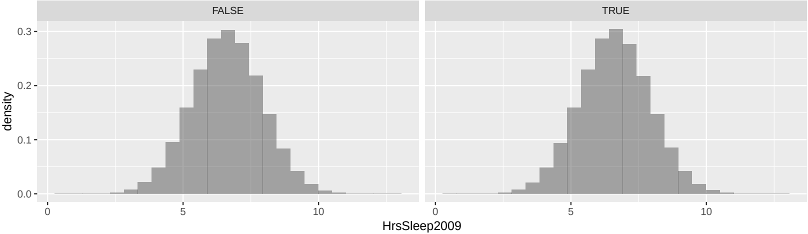
How much variability can we expect in F and PRE?

Let's imagine a world where $\beta_1 = 0$ and examine the sampling variability in these two statistics.

# A simulation

(Note: Your book uses shuffling which we will discuss on Thursday)

```
popsize <- 1000000
nullpop<-data.frame(Cohab2009=c(rep(TRUE,times=0.647017*popsize),
                                   rep(FALSE,times=0.352983*popsize)))
#the only difference between this and the previous simulation is the 0
nullpop$HrsSleep2009<-6.6137+0*nullpop$Cohab2009+rnorm(popsize,0,sqrt(1
gf_dhistogram(~HrsSleep2009, data = nullpop) %>%
  gf_facet_wrap(Cohab2009~.)
```



## Drawing a random sample

Let's take a random sample of size 4224 from our population and examine the Cohab model.

Note that in the population there is NO RELATIONSHIP between Sleep and Cohab, but we may still see one in the sample.

```r
sample1 <- sample(nullpop, 4224)
Cohab.model1 <- lm(HrsSleep2009~Cohab2009, data = sample1)
supernova(Cohab.model1)
```

```
## Analysis of Variance Table (Type III SS)
## Model: HrsSleep2009 ~ Cohab2009
##
##                               SS   df    MS     F    PRE     p
## ----- ----------------- -------- ---- ----- ----- ------ -----
## Model (error reduced) |    0.079    1 0.079 0.049 0.0000 .8250
## Error (from model)    | 6859.981 4222 1.625
## ----- ----------------- -------- ---- ----- ----- ------ -----
## Total (empty model)   | 6860.060 4223 1.624
```

# Many random samples

We can repeat this process many many times, saving for F and PRE to get a sampling distribution of these two statistics.

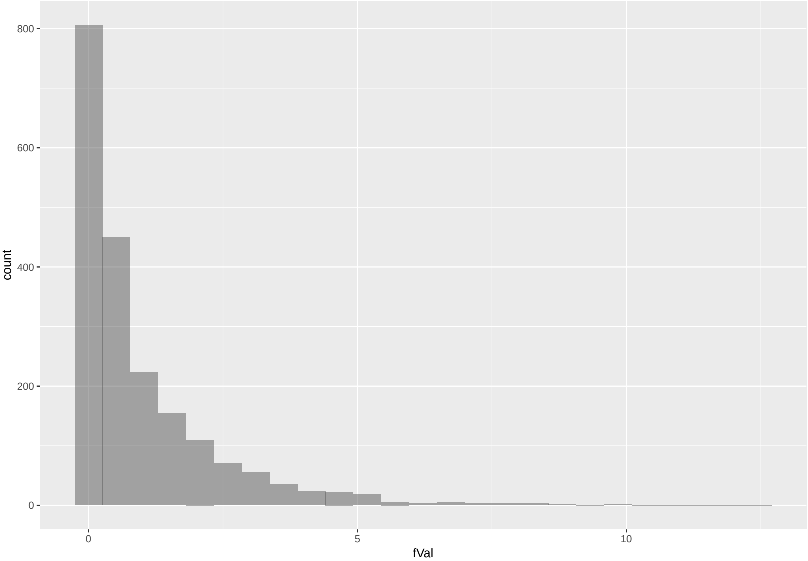
```
SDoF<-do(2000)*fVal(HrsSleep2009~Cohab2009,data=sample(nullpop,4224))
SDoPRE<-do(2000)*PRE(HrsSleep2009~Cohab2009,data=sample(nullpop,4224))

cbind(head(SDoF), head(SDoPRE))
```

```
##         fVal          PRE
## 1 0.03205002 7.346754e-05
## 2 0.21883291 4.354970e-04
## 3 0.20912164 3.253017e-04
## 4 0.43093607 1.251034e-04
## 5 1.96742479 1.022594e-03
## 6 0.17803158 2.448804e-04
```

# Visualizing the sampling distributions

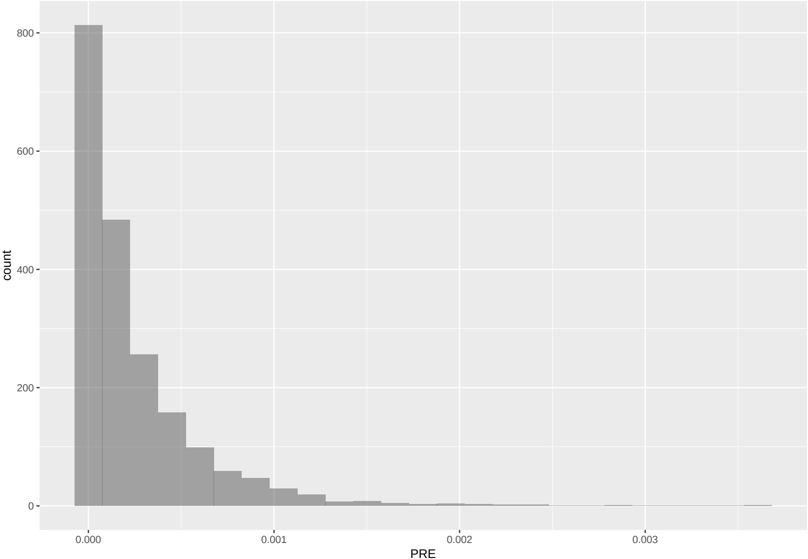
```
gf_histogram(~fVal, data = SDoF)
```



# Visualizing the sampling distributions

F and PRE have distributions of similar shape (non-normal, skewed)

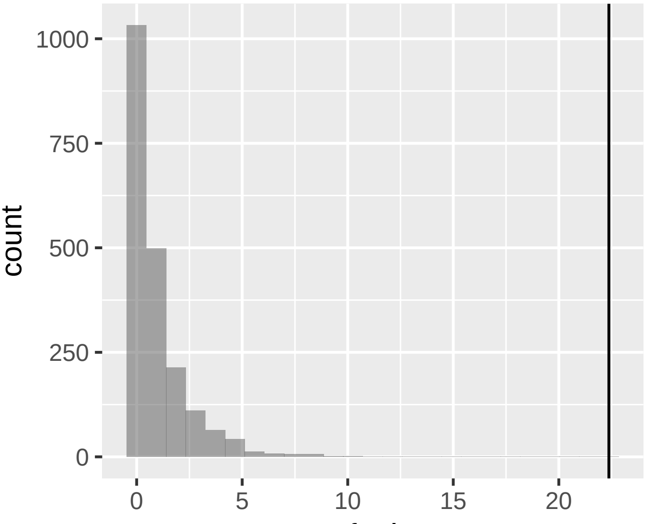
```
gf_histogram(~PRE, data = SDoPRE)
```



# F-distribution and PRE distribution

We can find our observed F-value in our sampling distribution.

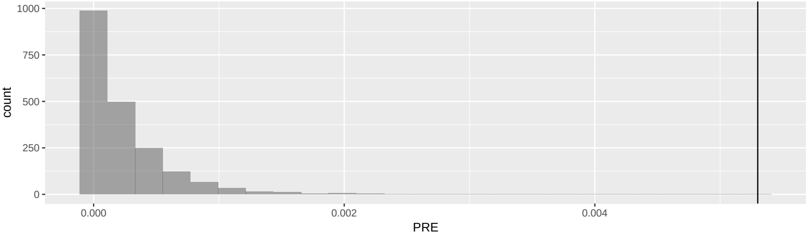
```
gf_histogram(~fVal, data = SDoF)%>%
  gf_vline(xintercept = 22.373)
```



# PRE distribution

Even though the PRE seems small, it's very unlikely given the simple model is true (i.e., no relationship between Sleep and Cohab)

```
gf_histogram(~PRE, data = SDoPRE)%>%
  gf_vline(xintercept = 0.0053)
```



# Calculating Exact Probabilities

What's the probability that we observe and F or a PRE this large under the no relationship hypothesis (often called Null Hypothesis [simple model])?

```
tally(~(fVal > fVal(Cohab.model)), data = SDoF, format = "proportion")
```

```
## (fVal > fVal(Cohab.model))
##  TRUE FALSE
##     0     1
```

```
tally(~(PRE > PRE(Cohab.model)), data = SDoPRE, format = "proportion")
```

```
## (PRE > PRE(Cohab.model))
##  TRUE FALSE
##     0     1
```

# Using an F distribution

There is a formal mathematical distribution for the F-value which depends on both the model and error degrees of the freedom.

We can use this distribution to calculate probabilities of certain outcomes, under different DGPs

```
supernova(Cohab.model)
```

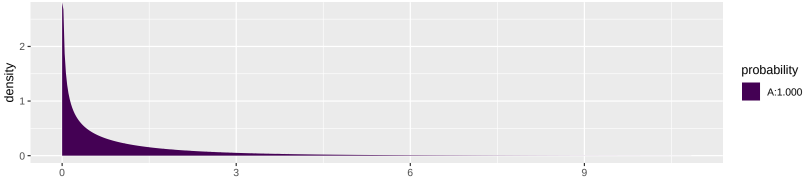
```
##  Analysis of Variance Table (Type III SS)
##  Model: HrsSleep2009 ~ Cohab2009
##
##                                SS   df     MS      F    PRE     p
##  ----- ----------------- -------- ---- ------ ------ ------ -----
##  Model (error reduced) |   37.914    1 37.914 22.373 0.0053 .0000
##  Error (from model)    | 7154.812 4222  1.695
##  ----- ----------------- -------- ---- ------ ------ ------ -----
##  Total (empty model)   | 7192.726 4223  1.703
```

# Using an F distribution

There is a formal mathematical distribution for the F-value which depends on both the model and error degrees of the freedom.

We can use this distribution to calculate probabilities of certain outcomes, under different DGPs

```
xpf(fVal(Cohab.model), 1, 4222)
```



```
## [1] 0.9999977
```

## P-value

We have calculated many probabilities from sampling distributions before!

There is a special name when we think about the probability of getting the outcome from our sample or something more extreme under the null hypothesis (simple model).

We call this the p-value: The probability of getting a group difference as big as we got (or bigger) when there is no group difference in the population.

We can calculate p-value many ways: simulation, mathematical distribution/supernova, **shuffling**

The p-value is used to make a decision: reject the simple model for the more complex model or retain it (due ot Occam's Razor)

In this case, the p-value is very very low, what decision should we make?

# An Alternative Interpretation of Confidence Intervals

```r
confint(Cohab.model)
```

```
##                   2.5 %    97.5 %
## (Intercept)   6.5475863 6.6797779
## Cohab2009TRUE 0.1160757 0.2804167
```

Confidence intervals and model comparison approaches have direct correspondence between them.

If 0 if contained within the confidence interval, we will retain the simple model (fail to reject simple model)

If 0 is not contained within the confidence interval, we will reject the simple model (adopt the complex model)

A confidence interval provides a range of hypotheses (in this case about $\beta_1$) for which we would reject the null hypothesis. For example, we would retain any models which propose that $\beta_1$ is between 0.116 and 0.280, but reject all hypotheses that propose $\beta_1$ is outside of this range.

# Comparing to other types of models

Though most of the time we compare the simple and complex model, we can also compare against other models where $\beta_1$ is fixed to a specific value.

Simple model: $Y_i = \beta_0 + 0 \times X_i + \epsilon_i$

Other model: $Y_i = \beta_0 + 1 \times X_i + \epsilon_i$