CLT-based Inference

Dr. Maria Tackett

11.19.19

Click for PDF of slides

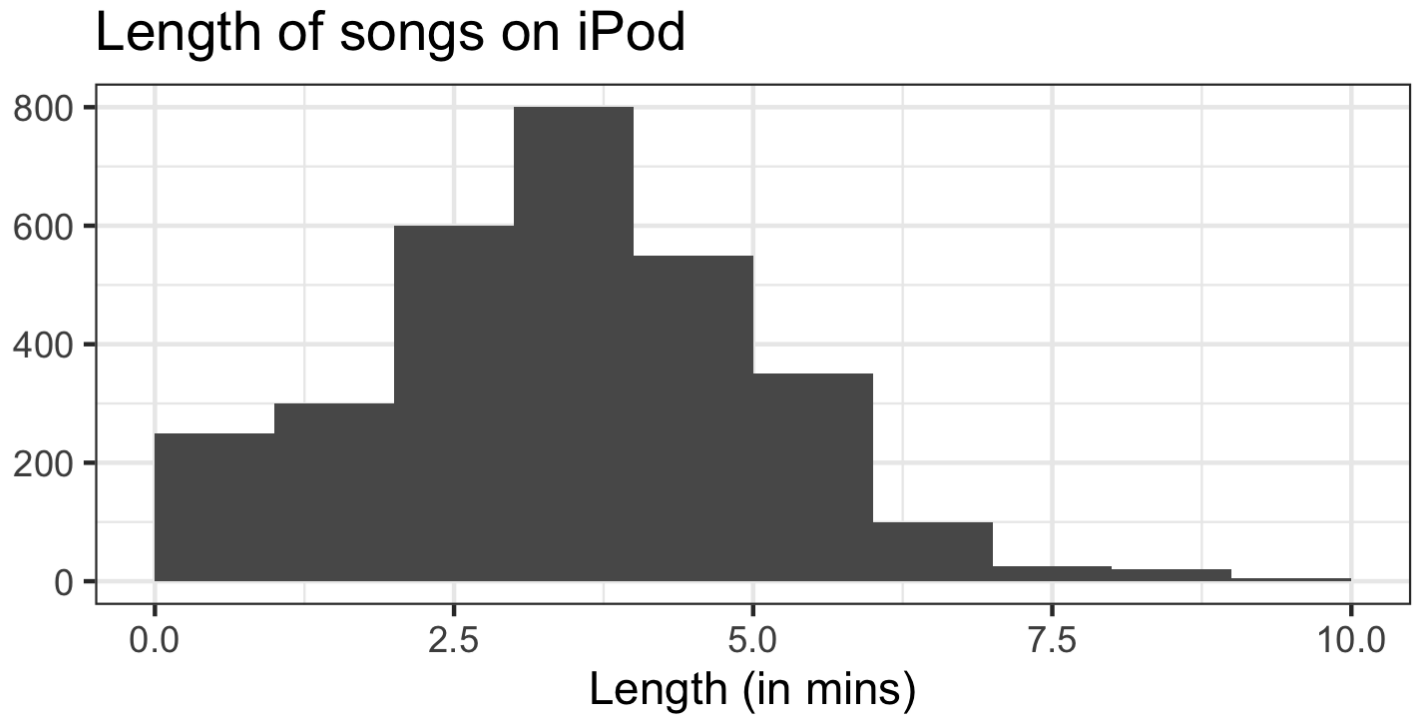
Announcements

- Writing Exercise #3 final revision **due TODAY at 11:59p**
- Exam 02 assigned after class, due Sunday, November 24 at 11:59p
  - Mostly modeling + inference
  - Some exploratory data analysis
  - Use in-line code to write narrative but also show output

Lab 07

- Assignment regraded - see Gradescope & Sakai for updated score
- Question 5 thrown out. Note
  - Adjusted R^2 is only used to compare models in multiple linear regression
  - R^2 is the proportion of variability in Y explained by the model

Suppose my iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for songs on this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. What is the approximate probability that a randomly selected song lasts more than 5 minutes?



I'm about to take a trip to visit friends and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive? Reminder: For songs on this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes.

Hints:

- You know how to find the distribution of $\bar{x}$ (average song length)
- To find probabilities under the normal curve, use the **`pnorm()`** function.

Why do we care?

Knowing the distribution of the sample statistic can help us

- estimate a population parameter as point estimate $\pm$ margin of error, where the margin of error is comprised of a measure of how confident we want to be and how variable the sample statistic is
- test for a population parameter by evaluating how likely it is to obtain to observed sample statistic when assuming that the null hypothesis is true as this probability will depend on how variable the sampling distribution is

Inference methods based on CLT

What is the CLT?

The Central Limit Theorem tells us the distribution of certain sample statistics if necessary conditions are met.

- The distribution of the sample statistic is nearly normal
- The distribution is centered at the (often unknown) population parameter
- The variability of the distribution is inversely proportional to the square root of the sample size

Inference methods based on CLT

If necessary conditions are met, we can also use inference methods based on the CLT:

- use the CLT to calculate the SE of the sample statistic of interest (sample mean, sample proportion, difference between sample means, etc.)
- calculate the **test statistic**, number of standard errors away from the null value the observed sample statistic is
  - Z for proportions
  - T for means, along with appropriate degrees of freedom
- use the test statistic to calculate the **p-value**, the probability of an observed or more extreme outcome given that the null hypothesis is true

Z distribution

Also called the standard normal distribution: $Z \sim N(mean = 0, \sigma = 1)$

Finding probabilities under the normal curve:

```
pnorm(-1.96)
```

```
## [1] 0.0249979
```

```
pnorm(1.96, lower.tail = FALSE)
```

```
## [1] 0.0249979
```

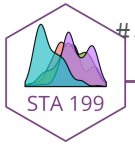
Finding cutoff values under the normal curve:

```
qnorm(0.025)
```

```
## [1] -1.959964
```

```
qnorm(0.975)
```

```
## [1] 1.959964
```

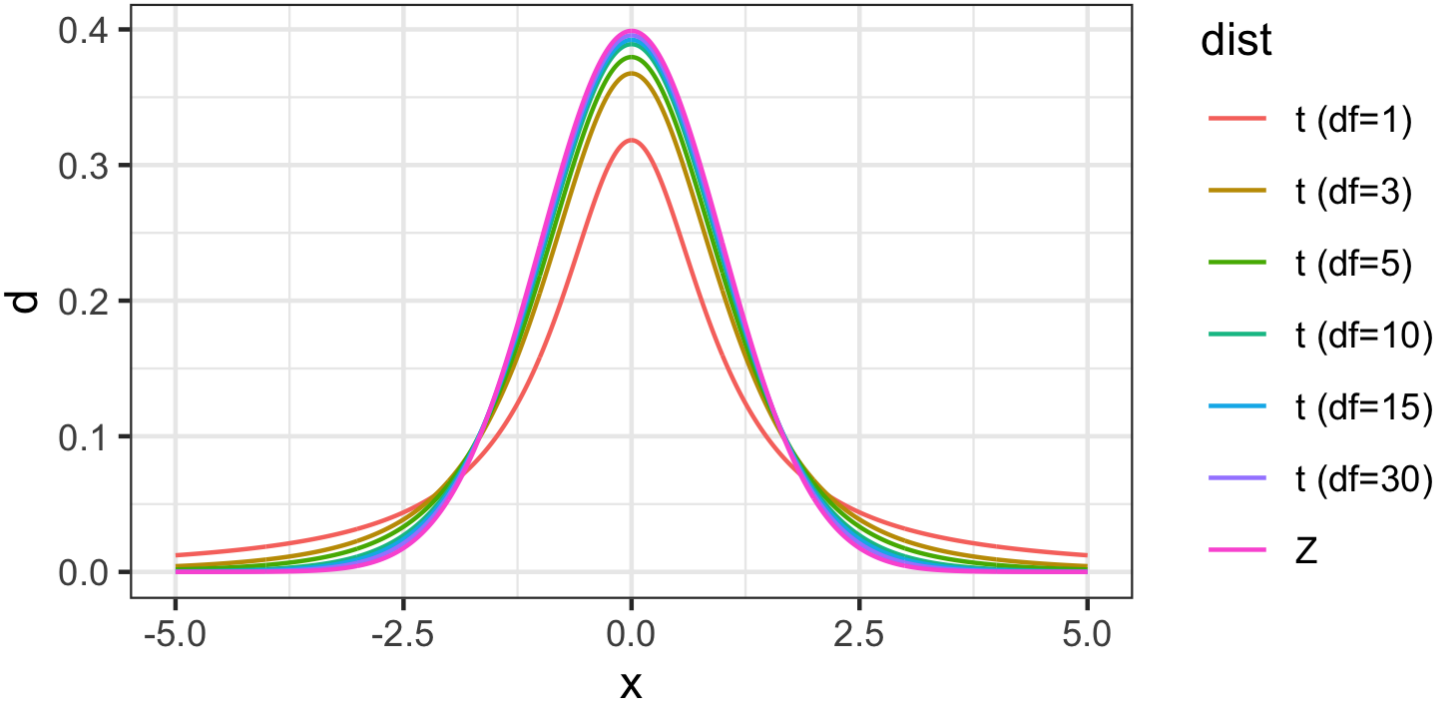


T distribution

- Also unimodal and symmetric, and centered at 0
- Thicker tails than the normal distribution (to make up for additional variability introduced by using s instead of σ in calculation of the SE)
- Parameter: degrees of freedom
  - df for single mean: $df = n - 1$
  - df for comparing two means:

$$df \approx \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \approx min(n_1 - 1, n_2 - 1)$$

T vs Z distributions



T distribution

Finding probabilities under the t curve:

```
pt(-1.96, df = 9)
```

```
## [1] 0.0408222
```

```
pt(1.96, df = 9, lower.tail = FALSE)
```

```
## [1] 0.0408222
```

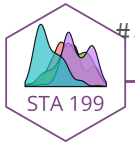
Finding cutoff values under the t curve:

```
qt(0.025, df = 9)
```

```
## [1] -2.262157
```

```
qt(0.975, df = 9)
```

```
## [1] 2.262157
```



Example

Relaxing after work

The GSS asks "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?". Do these data provide convincing evidence that Americans, on average, spend more than 3 hours per day relaxing? Note that the variable of interest in the dataset is **`hrsrelax`**.

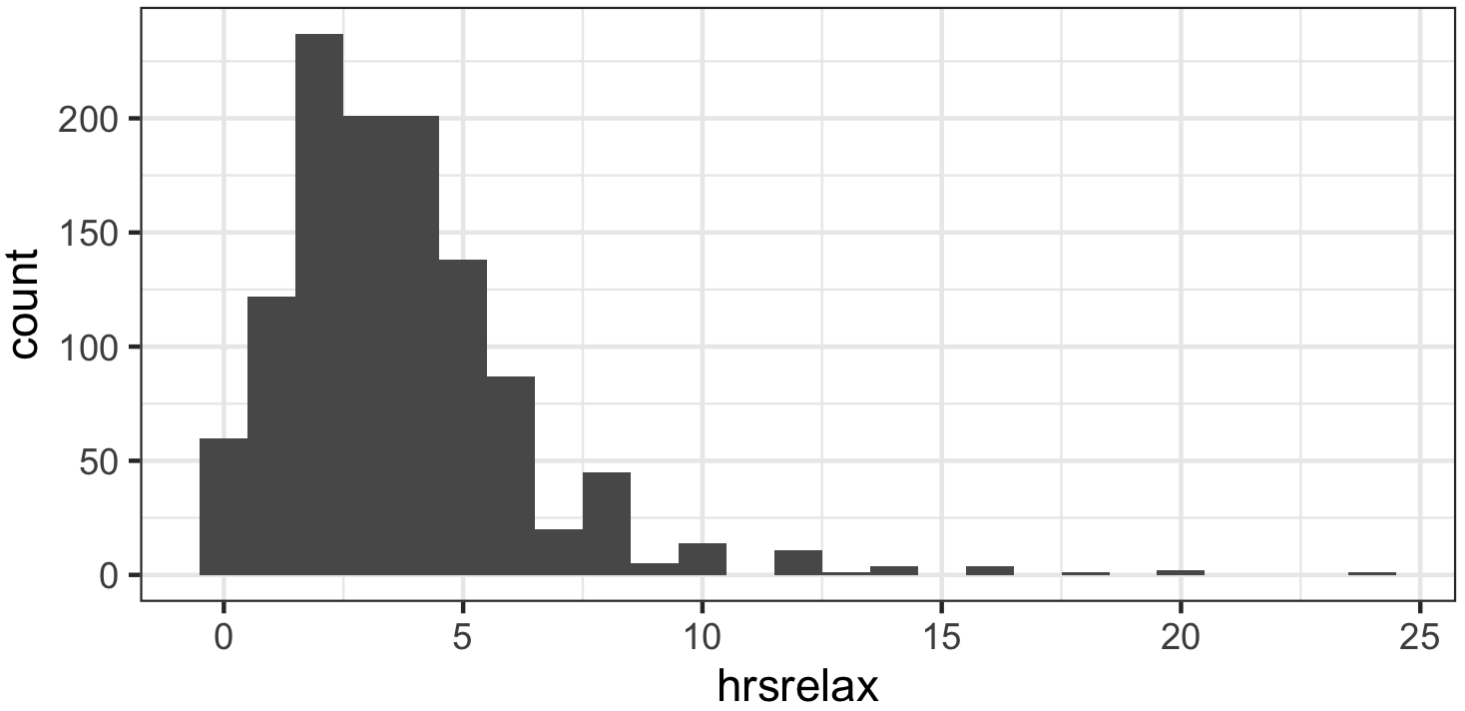
```
gss = read_csv("data/gss2010.csv")

gss %>%
  filter(!is.na(hrsrelax)) %>%
  summarise(x_bar = mean(hrsrelax), med = median(hrsrelax),
            sd = sd(hrsrelax), n = n())
```

```
## # A tibble: 1 x 4
##   x_bar   med    sd     n
##   <dbl> <dbl> <dbl> <int>
## 1  3.68     3  2.63  1154
```

Exploratory Data Analysis

```
ggplot(data = gss, aes(x = hrsrelax)) +
  geom_histogram(binwidth = 1)
```



Hypotheses

What are the hypotheses for evaluating if Americans, on average, spend more than 3 hours per day relaxing?

$$H_0 : \mu = 3$$

$$H_A : \mu > 3$$

Conditions

What conditions must be satisfied to conduct this hypothesis test using methods based on the CLT? Are these conditions satisfied?

Calculating the test statistic

Summary statistics from the sample:

```
## # A tibble: 1 x 3
##    xbar     s     n
##   <dbl> <dbl> <int>
## 1  3.68  2.63  1154
```

And the CLT says:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

How many standard errors away from the population mean is the observed sample mean?

How likely are we to observe a sample mean that is at least as extreme as the observed sample mean, if in fact the null hypothesis is true?

Calculations

```
(se <- hrsrelax_summ$s / sqrt(hrsrelax_summ$n))
```

```
## [1] 0.07740938
```

```
(t <- (hrsrelax_summ$xbar - 3) / se)
```

```
## [1] 8.7876
```

```
(df <- hrsrelax_summ$n - 1)
```

```
## [1] 1153
```

```
pt(t, df, lower.tail = FALSE)
```

```
## [1] 2.720895e-18
```

Conclusion

- Since the p-value is small, we reject H_0.
- The data provide convincing evidence that Americans, on average, spend more than 3 hours per day relaxing after work.

Would you expect a 90% confidence interval for the average number of hours Americans spend relaxing after work to include 3 hours?

Confidence interval for a mean

$$point\ estimate \pm critical\ value \times SE$$

```
t_star <- qt(0.95, df)
pt_est <- hrsrelax_summ$xbar
round(pt_est + c(-1,1) * t_star * se, 2)
```

```
## [1] 3.55 3.81
```

Interpret this interval in context of the data.

Built-in functionality in R

- There are built in functions for doing some of these tests in R:
- However a learning goal is this course is not to go through an exhaustive list of all CLT based tests and how to implement them
- Instead you should try to understand how these methods are / are not like the simulation based methods we learned about earlier

What is similar, and what is different, between CLT based test of means vs. simulation based test?

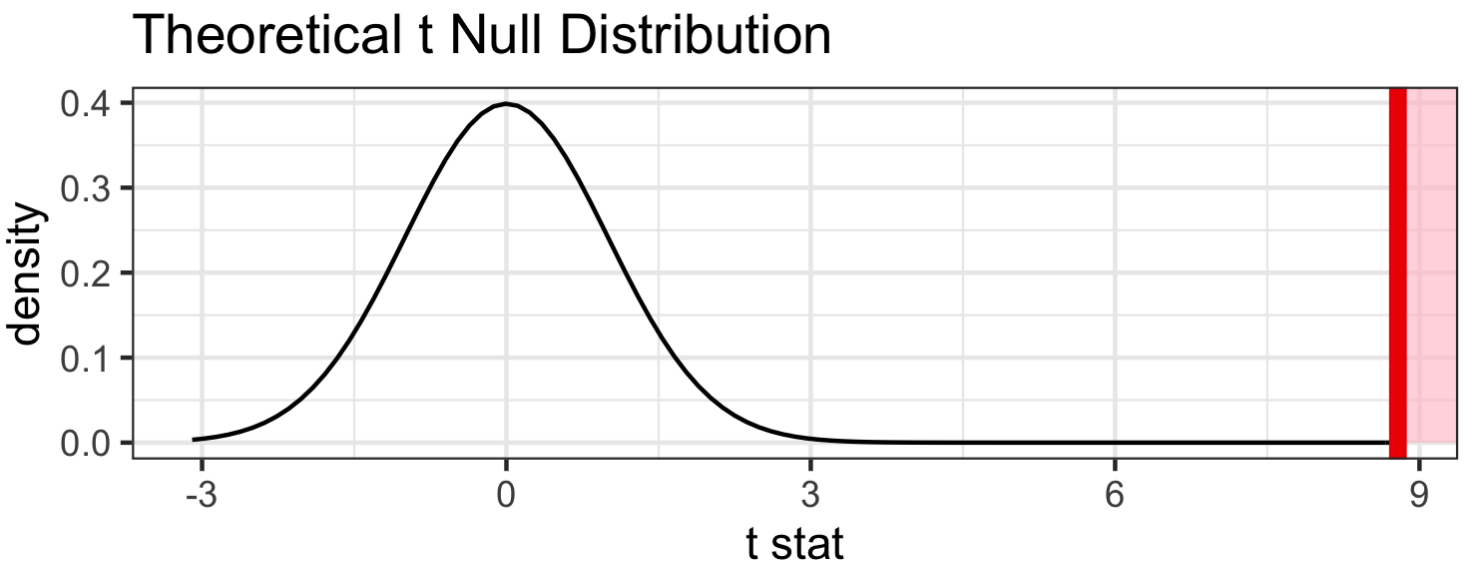
t distribution using `infer`

```
t_null_theor <- gss %>%
  filter(!is.na(hrsrelax)) %>%
  specify(response = hrsrelax) %>%
  hypothesize(null = "point", mu = 3) %>%
  # generate() ## Not used for theoretical
  calculate(stat = "t")
```

t distribution using `infer`

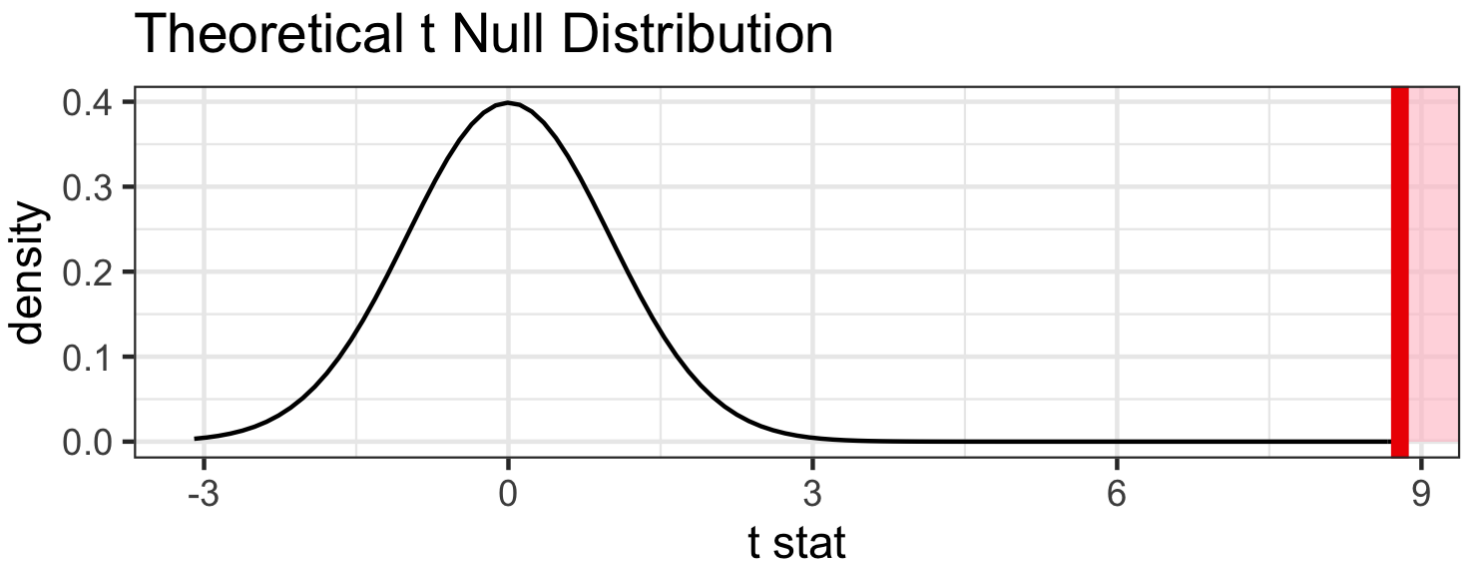
```
visualize(t_null_theor, method = "theoretical") +
  shade_p_value(obs_stat = 8.7876, direction = "greater")
```

```
## Warning: Check to make sure the conditions have been met for the
## theoretical method. {infer} currently does not check these for you.
```



Calculate p-value

```
## Warning: Check to make sure the conditions have been met for the
## theoretical method. {infer} currently does not check these for you.
```



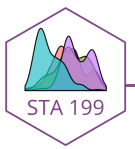

```
df <- hrsrelax_summ$n - 1
pt(8.7886, df, lower.tail = FALSE)
```

```
## [1] 2.698289e-18
```

Hypothesis tests in R

```
# Hypothesis tests
t.test(gss$hrsrelax, mu = 3, alternative = "greater")
```

```
##
##     One Sample t-test
##
## data:  gss$hrsrelax
## t = 8.7876, df = 1153, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 3
## 95 percent confidence interval:
##  3.552813      Inf
## sample estimates:
## mean of x
##  3.680243
```



Confidence intervals in R

```
# Confidence intervals
t.test(gss$hrsrelax, conf.level = 0.90)
```

```
##
##     One Sample t-test
##
## data:  gss$hrsrelax
## t = 47.543, df = 1153, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  3.552813 3.807672
## sample estimates:
## mean of x
##  3.680243
```