Lab 11

Taylor Blair

Math 141, Week 12

Due: Before your Week 13 lab meeting (Not during Break)

Goals of this lab

1. Practice inference for categorical data.
2. Practice using the ANOVA test.
3. Practice putting in your own R chunks.

Problems

For most of the R chunks needed, we didn't insert them so that you can get more practice adding your own R chunks.

```r
#Load libraries
library(tidyverse)
library(infer)
library(broom)
library(fmsb)
```

Problem 1

We want to know if there is a relationship between education level and whether or not there is only one true love for each person. In October of 2010, the Pew Foundation ran a telephone survey on a random sample of adults. They asked the following question: "Some people say there is only one true love for each person. Do you agree or disagree?" They also collected demographic information on the participants, such as their education level (no college, some college, or college degree).

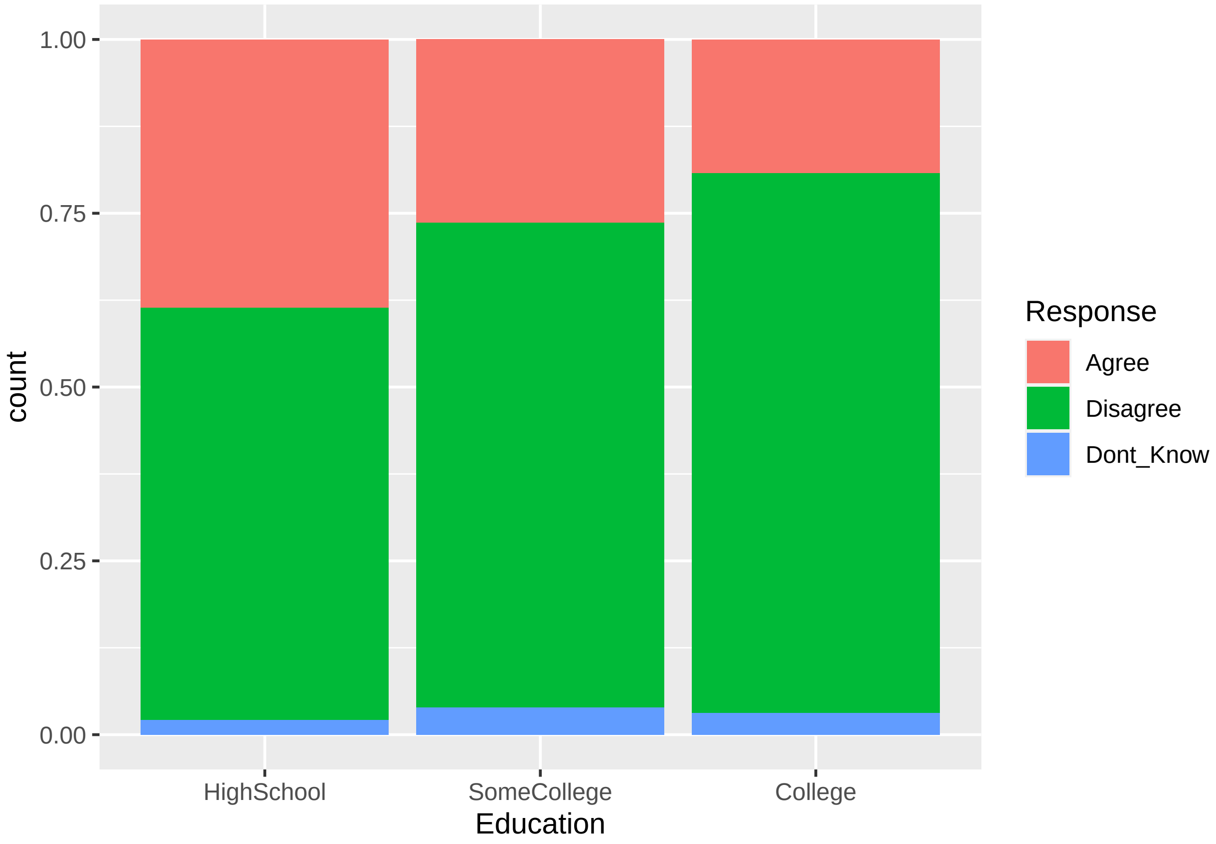
Here are the data from their survey.

```r
# Load data
love_data <- read_csv("/home/courses/math141f19/Data/love.csv")
```

a. Create a useful graph of the response and explanatory variables. Make sure that the order of `Education` is: "High School", "Some College", "College". Based on your graph, does there appear to be a relationship between education level and whether someone believes in the idea of one true love? (Consider reordering Education to make it easier to interpret your graph.)

```r
love_data$Education <- factor(love_data$Education, levels=c("HighSchool", "SomeCollege", "College"))
```

```r
ggplot(love_data, aes(Education, fill=Response)) +
  geom_bar(position = "fill")
```



---

b. State the null and alternative hypotheses.

- Null hypothesis
  - There is not a significant difference between groups belief in one true love.
- Alternative hypothesis
  - There is variance between groups OR the more education the less one belives in "one true love"

---

c. Run the following R code and explain the tables that are generated.

```r
# Table 1
table(love_data$Response, love_data$Education)
```

```
##
##               HighSchool SomeCollege College
##   Agree              363         176     196
##   Disagree           557         466     789
##   Dont_Know           20          26      32
```

```r
# Table 2
chisq.test(table(love_data$Response, love_data$Education))$expected
```

```
##
```

```
##              HighSchool SomeCollege   College
##   Agree       263.20000   187.04000 284.76000
##   Disagree    648.86857   461.11086 702.02057
##   Dont_Know    27.93143    19.84914  30.21943
```

- `table` makes a pivot table of data
- The `chisq.test` code creates a perfect χ^2 distribution. One where each column has an equal distribution but proportional to the original amount of individuals.

---

d. Compare the expected table and observed table. Do they agree? If not, what is the trend?

```
table(love_data$Response, love_data$Education) -
  chisq.test(table(love_data$Response, love_data$Education))$expected
```

```
##
##              HighSchool SomeCollege    College
##   Agree       99.800000  -11.040000 -88.760000
##   Disagree   -91.868571    4.889143  86.979429
##   Dont_Know   -7.931429    6.150857   1.780571
```

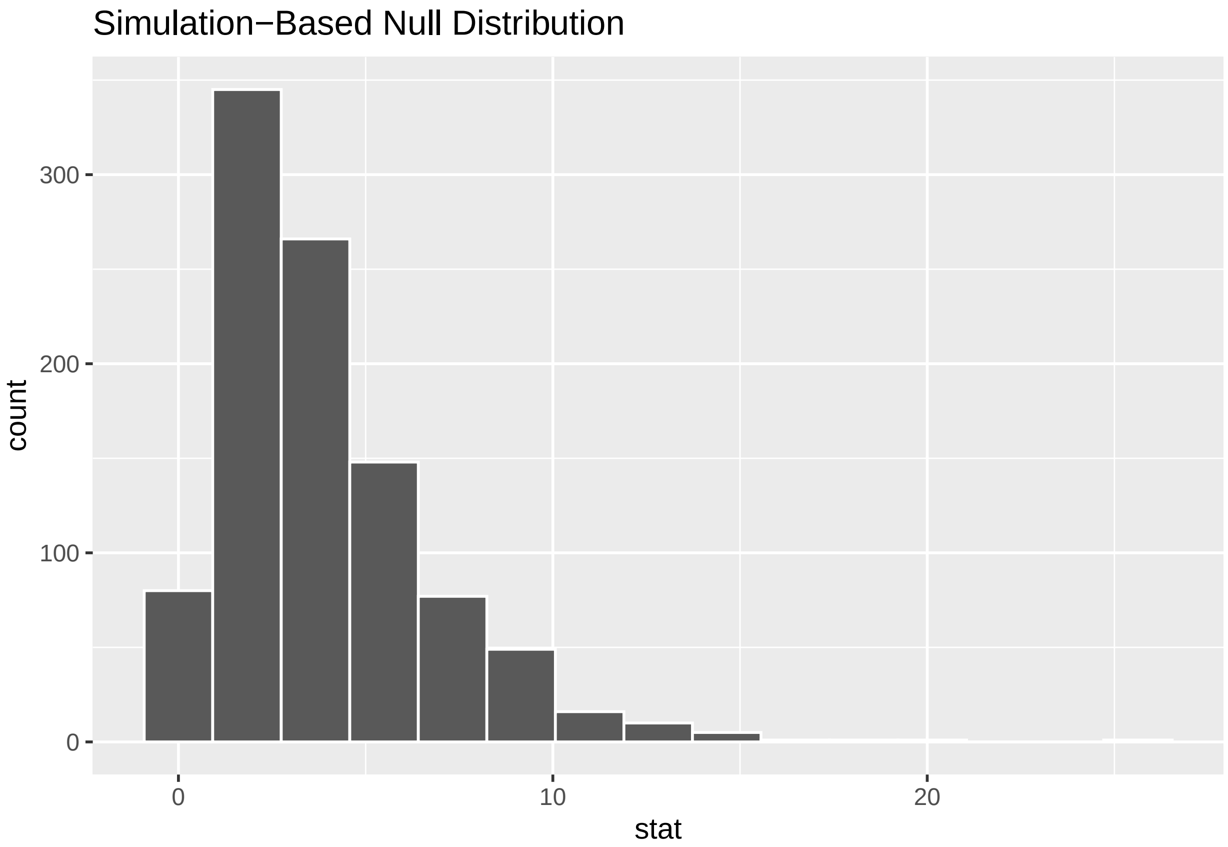
- There does not appear to be a particularly equal expected χ^2 to actual distibution.
- `HighSchool` and college vary wildly, while only someCollege has a distribution where $\hat{y} \approx \bar{y}$

---

e. Find the null distribution for the test statistic using simulation-based methods.

```
null_love <- love_data %>%
  specify(response = Response, explanatory = Education) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")

null_love %>%
  visualise()
```



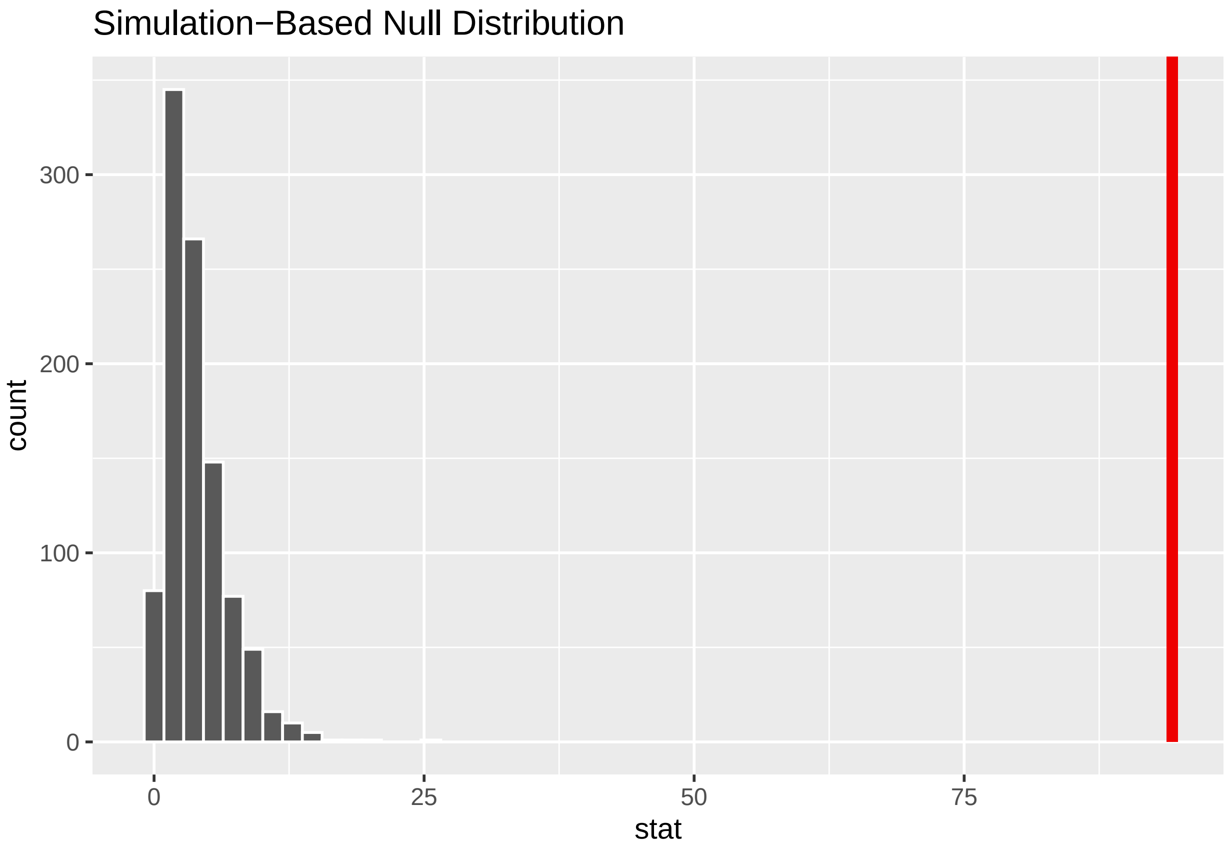
f. Compute the observed test statistic and create a graph of the null distribution. What does this graph and your observed test statistic imply about the sample results?

```
test_stat <- love_data %>%
  specify(response = Response, explanatory = Education) %>%
  hypothesize(null = "independence") %>%
  calculate(stat = "Chisq")

cat("Test statistic is: ", test_stat$stat)
```

```
## Test statistic is:  94.25895
```

```
null_love %>%
  visualise() +
  shade_p_value(obs_stat = test_stat, direction = "right")
```



- 0 of the generated samples were to the right of the observed χ^2 value.
- We can expect that we can reject the null hypothesis.

---

g. Can the distribution of the test statistic be approximated by a known distribution? If so, state the distribution.

- We have a χ^2 test stat of 94.2589542
- There are 4 degrees of freedom (Johnathan was stumped and Simon offered a solution)

---

h. Find the p-value where you approximate the distribution of the test statistic with a chi-squared distribution.

- Using the code `sum(null_love$stat>94)/length(null_love$stat)`
- P-value of: 0

---

i. State your conclusions at a significance level of 0.05.

- Because we have a p-value of 0, it won't matter what significance value we use. We can reject the null and accept the first alternative hypothesis that there is variability between groups

---

j. Based on the expected and observed tables does it appear that more education correlates with a higher belief in one true love? Why or why not?

- The oppisite. The more education the less one believes in “One True Love”
- We have rejected the null so we can say that our results are “stastically significant”
- The alternative seems likely based on the prop stack bar plot

Problem 2

Each semester as part of Math 141, students collect data on Reed senior theses. Because of social distancing requirements, we didn’t collect these data this time around. So for this problem, we will analyze the data collected by last Fall’s Math 141 students. (Note: These data were collected using random sampling techniques.)

I want you to determine if we have evidence that the average number of pages in Reed senior theses varies by division. I have subset the data to only include the three most common divisions (MNS = Mathematics and Natural Sciences, LL = Literature and Languages, and HSS = History and Social Science).

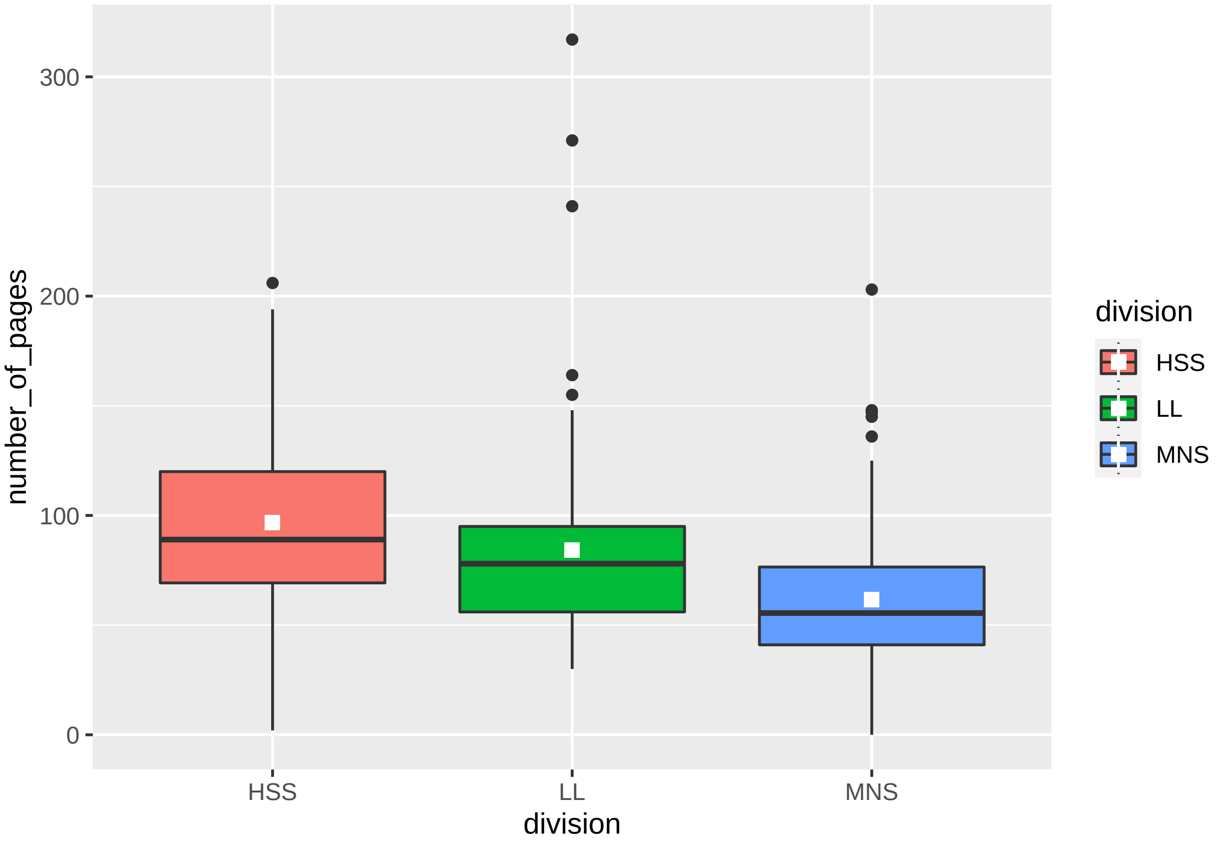
```r
# Load data
thesis_data <- read_csv("/home/courses/math141f19/Data/theses.csv") %>%
  drop_na(division, number_of_pages)
```

a. State the null and alternative hypotheses in words.

- Null H_0
  - The number of pages is identical across departments
  - $HSS = LL = MNS$
- Alternative H_a
  - There is a greater variance between groups than within groups.

b. Create a boxplot to compare the number of pages across division and include the group sample means on the graph. Draw some initial conclusions.

```r
thesis_data %>%
  ggplot(aes(division, number_of_pages, fill=division)) +
  geom_boxplot() +
  stat_summary(fun.y="mean", color="white", shape=15)
```



- Mean is above median because the data is skewed towards larger papers.
- *Mean is the white square*
- Variance between groups but still some overlap.

c. Compute the ANOVA test statistic (F_o). What does this imply about the between group variability compared to the within group variability?

```
thesis_data %>%
  group_by(division) %>%
    summarise(
      count = n(),
      mean = mean(number_of_pages),
      sd = sd(number_of_pages)
    )
```

```
## # A tibble: 3 x 4
##   division count  mean    sd
##   <chr>    <int> <dbl> <dbl>
## 1 HSS        118  96.8  37.2
## 2 LL          93  84.2  45.4
## 3 MNS        168  61.7  31.5
```

```
test_thesis <- summary(aov(number_of_pages ~ division,
            thesis_data))[[1]][1,4]
```

```
test_thesis
```

```
## [1] 32.5649
```

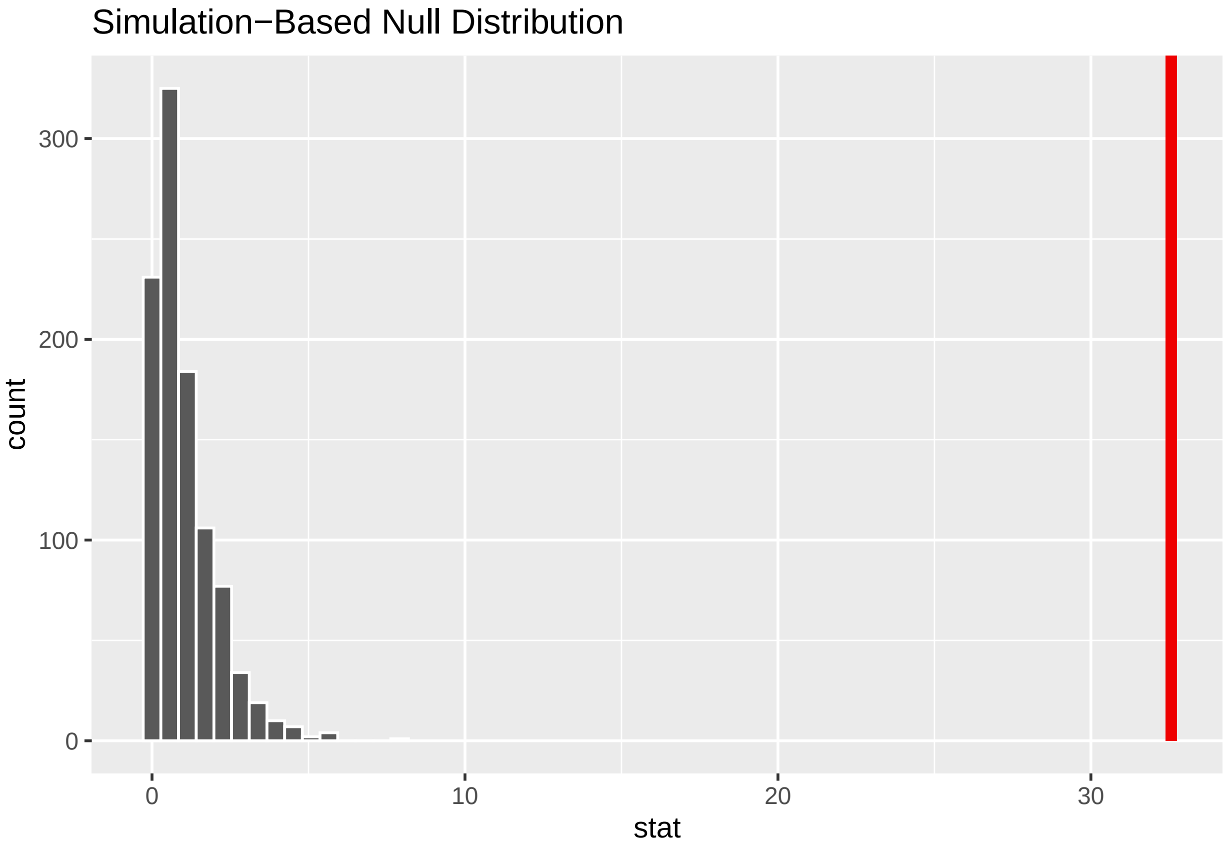
- F-stat is 32.5649018
- There is greater group variability compared to within group value. By ≈ 32 times

---

d. Generate the null distribution of the test statistic. Comparing your observed test statistic to the null distribution, what does this imply about the sample results?

```
null_thesis <- thesis_data %>%
  specify(number_of_pages ~ division) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "F")

null_thesis %>%
  visualise() +
  shade_p_value(test_thesis, "right")
```



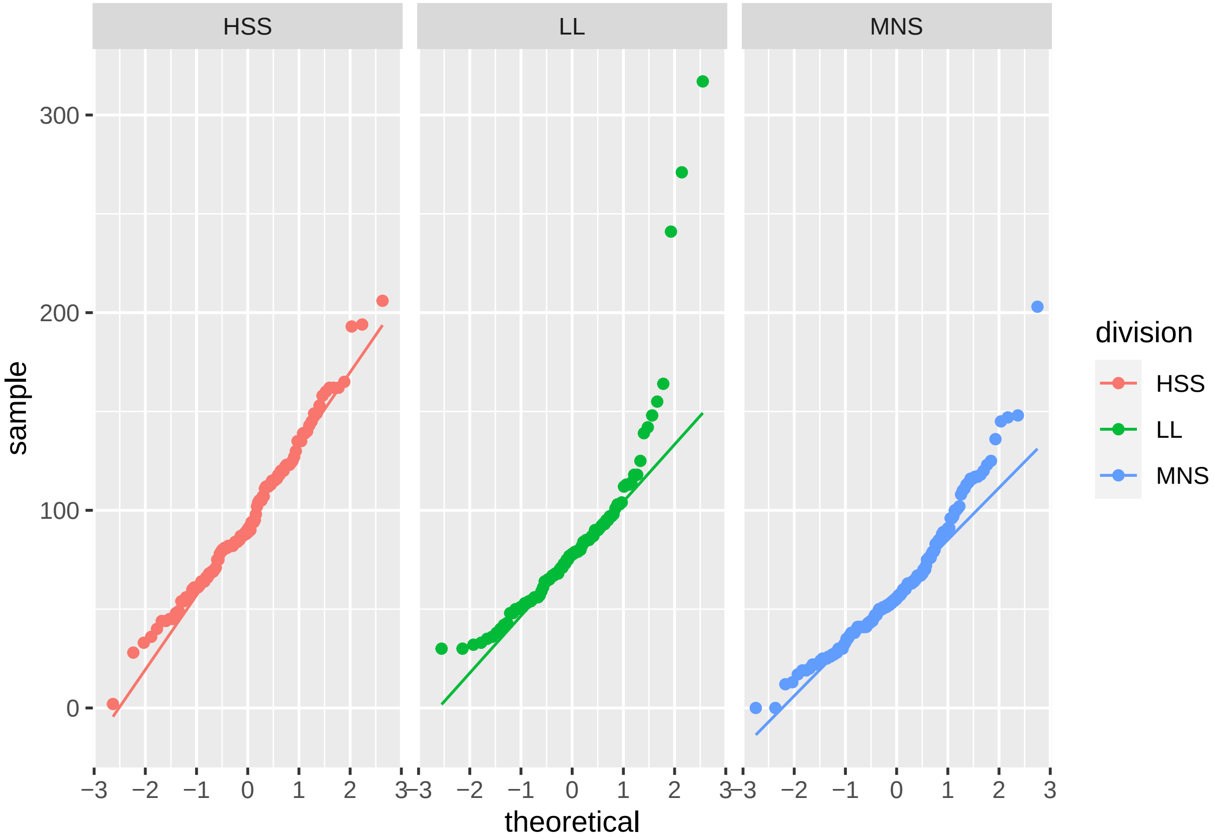
---

e. Check the conditions for whether or not we can approximate the null distribution with an F distribution.

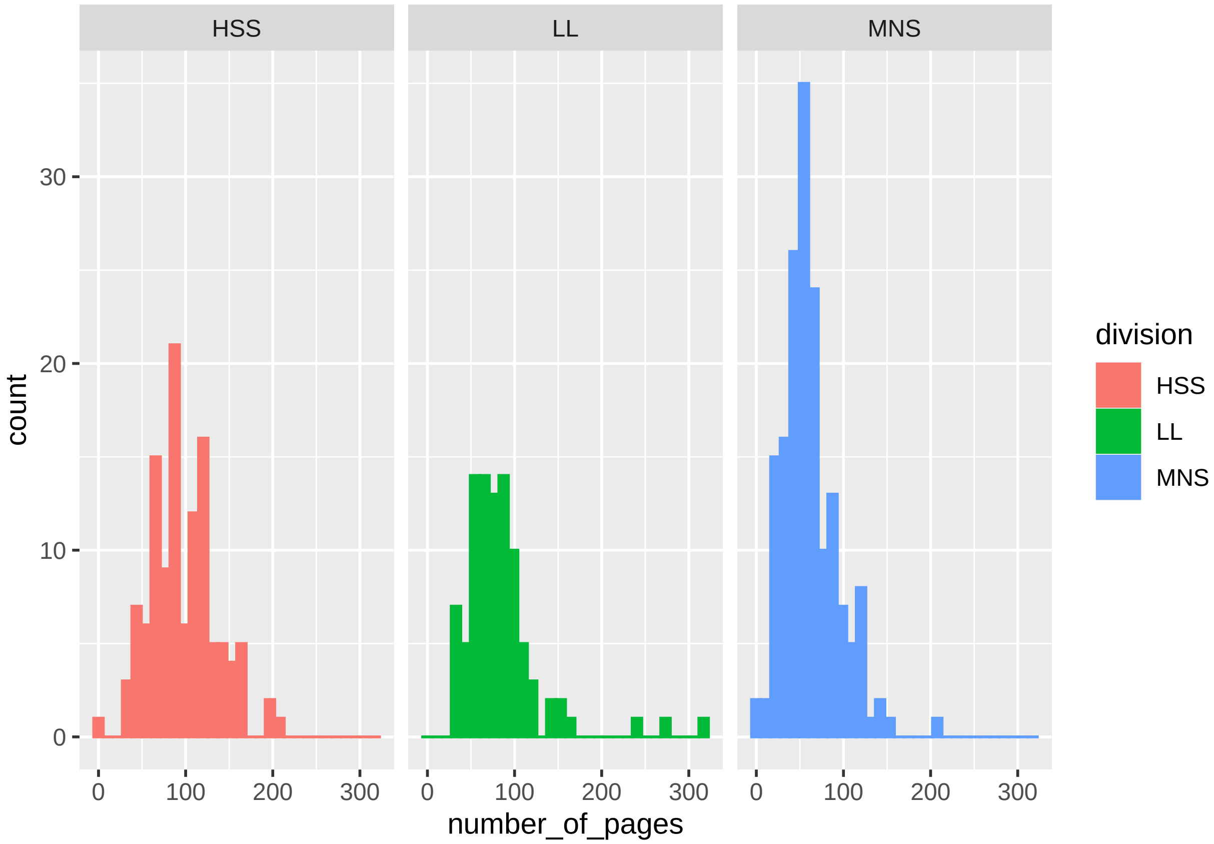
- Not a t-test as we have more than two groups.

- Not a χ^2 as we are not looking at the difference in props
- More than thirty elements

```
thesis_data %>%
  group_by(division) %>%
  ggplot(aes(sample = number_of_pages,
             fill = division,
             color=division))+
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~division)
```



```
# I am leaving this second graph into spite Johnathan K
thesis_data %>%
  group_by(division) %>%
  ggplot(aes(x = number_of_pages,
             fill = division,
             color=division))+
  geom_histogram() +
  facet_wrap(~division)
```



- Based off of the QQ plot, the data is normal shaped with some right skew.

---

f. Produce the ANOVA table (using `aov()`).

```
summary(aov(number_of_pages ~ division, thesis_data))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## division      2  89629   44814   32.56 9.05e-14 ***
## Residuals   376 517436    1376                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

g. Do we have evidence that page lengths, on average, vary by division? Justify your answer.

- Yes, p-value is $9.0538135 \times 10^{-14}$. Which is so well below any threshold, so we can reject our null hypothesis.

---

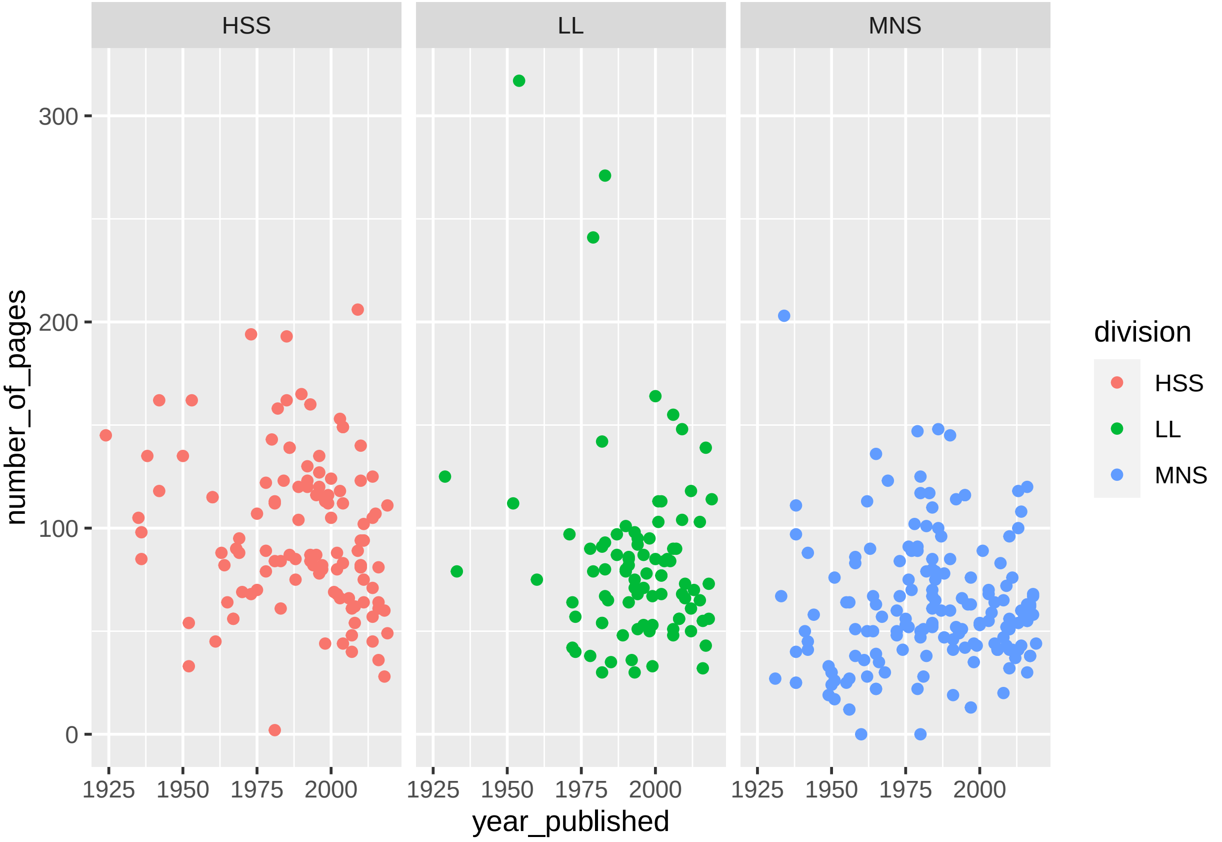
h. What are the effect sizes here? In other words, on average, how does page length differ between HSS and LL, between HSS and MNS, and between HSS and MNS? As the domain expert (i.e. a current or future thesising student), are those effect sizes *practically significant*? Justify your statement.

```
thesis_data %>%
  group_by(division) %>%
    summarise(
      count = n(),
      mean = mean(number_of_pages),
      sd = sd(number_of_pages)
    )
```

```
## # A tibble: 3 x 4
##   division count  mean    sd
##   <chr>    <int> <dbl> <dbl>
## 1 HSS        118  96.8  37.2
## 2 LL          93  84.2  45.4
## 3 MNS        168  61.7  31.5
```

- Yes, looking at a pivot stastics summary table, the means of the number of pages among groups varies.
- However, it might be worth looking into the change in length over time of both the number of theses published by a department and the page length.

```
thesis_data %>%
  group_by(division) %>%
  ggplot(aes(x=year_published,
             y=number_of_pages,
             color=division)) +
  geom_point() +
  facet_wrap(~division)
```



- Varies significantly less as time goes on. I would limit to the last decade in future.
