STAT 102: Week 13

Ricky's Section

Introductions and Attendance

Introduction: Name

Question of the Week: Last section, so this one is open-ended! One word to describe how you're feeling? Favorite memory? Something random?

Important Reminders

Upcoming Events

- **Class Lunch** on Tuesday, 04/29 at noon
  - RSVP here!
- **ggparty** on Thursday, 05/01 from noon to 1:30 PM in Science Center 316
  - RSVP here!
- **Review Session** on Monday, 05/12 from 7 to 9 PM in Science Center 706

Apply to be a TF!

- The form is due by Friday, 05/02
  - https://docs.google.com/forms/d/e/1FAIpQLSet4frJdnuQISFFumi8GANjglEl63CTOitEXeAZcjP4AbkPpQ/viewform?usp=sharing
- Let me or any of the other TFs know if you have any question!

Modified Office Hours

- The OH spreadsheet (on Canvas) has sections for Modified Office Hours on account of Reading Period
  - https://docs.google.com/spreadsheets/d/1AgnpomB7qUGtyRTI8Ansothj-2LbZnKJYO7q0br6c18/edit?usp=sharing

Posit Cloud

- If you want help installing R and RStudio onto your computer (in case Posit Cloud doesn't work well), let Julie know!
- It's also helpful to have because our workspace on Posit Cloud expires after this semester

Content Review: Week 12

Big Picture Overview

- We introduce 3 (+2) more tools in our inference toolkit
- These are extensions of things we've seen before
  - Paired t-test
  - ANOVA
  - Chi-squared
- Also…
  - Fisher's exact test
  - Effect size in 2x2 tables

An Introduction to Pairing

- Two-sample numerical data can be **paired** or **unpaired** (i.e., independent)
- Thus far, we've been working with **unpaired**
  - Observations cannot be matched on a one-to-one
  - *Ex: Considering SAT scores for students who studied versus students who did not, we can't match Alice, who studied, with Bob, who didn't—they're completely different people!*
- Now, let's consider studies with **paired** measurements
  - Each observation can be logically matched to another observation in the data
  - *Ex: Considering SAT scores for a group of 10 students before and after studying, we're matching Alice's old score with her new score*

If we want to measure the effect of new wetsuits on swimmers, should we have paired data or unpaired data?

Question:

If we want to measure the effect of new wetsuits on swimmers, should we have paired data or unpaired data?

While both strategies could work, this research question might be best answered with a **paired study**.

It'd be better to keep our swimmers consistent (since everyone has their own velocity, generally). Thus, our data can be paired "before and after."

Paired t-test: Example

- 12 swimmers had their velocity measured using an (old) swimsuit and using a (new) wetsuit
  - These are paired data (e.g., swimmer 1 swimsuit can be matched with swimmer 1 wetsuit)
- Conducting a **non-paired t-test** (what we've done before), we get $\bar{x}_{swimsuit} - \bar{x}_{wetsuit} = 0.0775$ m/s, with a p-value of 0.18
- For **paired t-test**, we look at **đ**, the **sample mean** of differences in velocities
  - If swimmer 1 swam 1.5 m/s with wetsuit and 1.4 m/s with swimsuit, their difference is 0.1 m/s
  - **đ** would be average of 12 differences (e..g,
- **δ** is the **population mean** of difference in velocities (theoretically, for all swimmers—not just 12)

Paired t-test: Example

- **H_o: $\delta = 0$**, the **population mean** difference in swim velocities between swimming with a swimsuit versus a wetsuit equals 0
  - That is, wetsuits do NOT change swim velocities
- **H_A: $\delta \neq 0$**, the **population mean** difference in swim velocities between swimming with a swimsuit versus a wetsuit is non-zero
  - That is, wetsuits DO change swim velocities
- $t = (\bar{d} - \delta_o)/(s_d/\sqrt{n})$, where **t** is our **standardized test statistic** (***t*-score**)
- $t \sim t(df = n - 1)$, where **n** is number of differences/pairs
- $95\%\ CI = \bar{d} \pm (t^* \times s_d/\sqrt{n})$, where **t*** is point on $t(df = n - 1)$ that has area 0.025 to its right (assuming $\alpha = 0.05$)

Paired t-test: Code

- As always, our computers do the math for us—we just need to code and interpret!
- Use `t_test()`
  - We're used to this from the tidyverse
  - A paired *t*-test is just a single-mean test on the differences

Paired t-Test: Code for t_test()

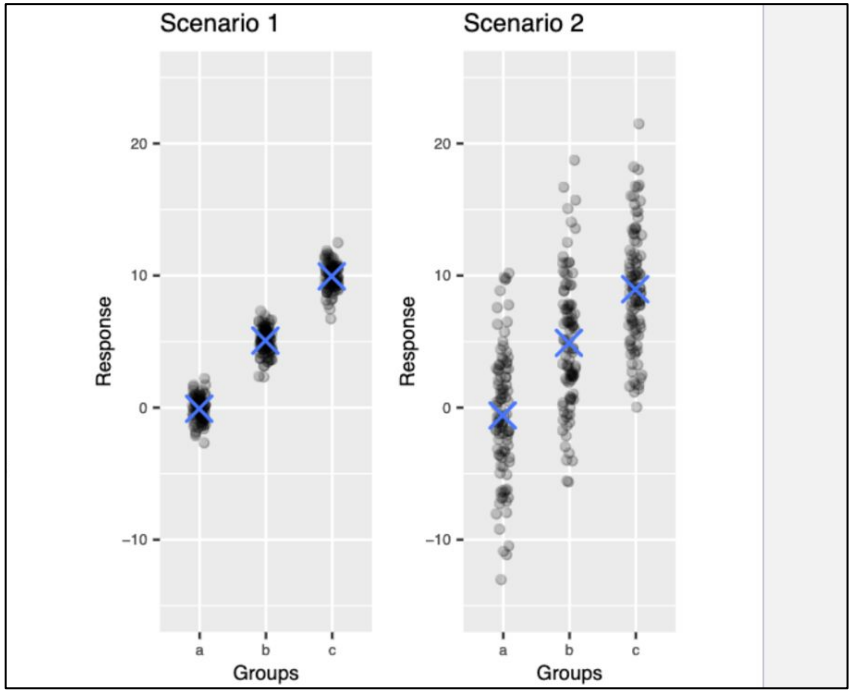
- **General form**: `DATASET %>% t_test(response = RESPONSE-VAR.diff)`
  - *`swim %>% t_test(response = velocity.diff)`*
  - Again, very similar to before, but now we're adding "**.diff**" because we're interested in the difference for each pair
- **Hypothesis tests**: `DATASET %>% t_test(response = RESPONSE-VAR.diff) %>% select(statistic, p_value, estimate)`
  - *`swim %>% t_test(response = velocity.diff) %>% select(statistic, p_value, estimate)`*
- **Confidence intervals**: `DATASET %>% t_test(response = RESPONSE-VAR.diff) %>% select(lower_ci, upper_ci)`
  - *`swim %>% t_test(response = velocity.diff) %>% select(lower_ci, upper_ci)`*

ANOVA: Analysis of Variance

- **ANOVA**: Test for when **response variable** is **numerical** and **explanatory variable** is **categorical (with more than 2 categories)**
  - H_o: $\mu_1 = \mu_2 = ... = \mu_k$ (i.e., variables are independent)
  - H_A: At least 1 mean is not equal to the rest (i.e., variables are dependent)
- Test statistic is **F-statistic**
  - **F = standardardized variance BETWEEN groups / standardized variance WITHIN groups**
- If H_o is true, **F-statistic** should be roughly equal to 1 (variance between groups should be equal to variance within groups)
- If H_A is true, **F-statistic** should be larger than 1
  - *Ex: If F is 3.88, the variance BETWEEN groups is 3.88 times larger than the variance WITHIN groups, which suggests the population means are different*

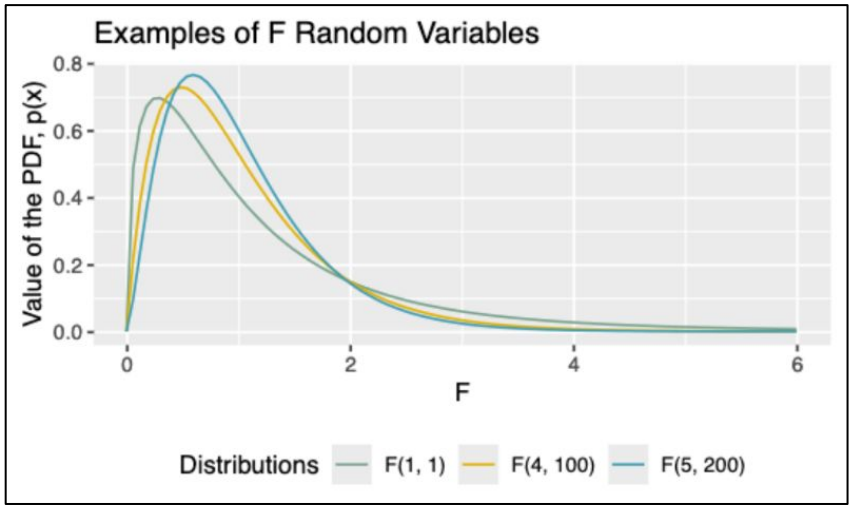
ANOVA: Intuition

- **Scenario 1**, there is little variability WITHIN groups but much more variability BETWEEN groups
  - It's plausible these groups come from different populations
- **Scenario 2**, there is a lot of variability WITHIN groups, so we're less sure… this would correspond to a low **F-statistic**



ANOVA: Theory-Based Inference

- When the ANOVA assumptions (next few slides) are satisfied, the **F-statistic** follows an **F distribution**, with two degrees of freedom: $\mathbf{df_1}$ and $\mathbf{df_2}$
- That is, **F-statistic ~ F($\mathbf{df_1}$, $\mathbf{df_2}$)**
  - $\mathbf{df_1 = n_{groups} - 1}$, $\mathbf{df_2 = n_{observations} - n_{groups}}$
- The p-value is P(F > observed F-statistic)—area to the RIGHT

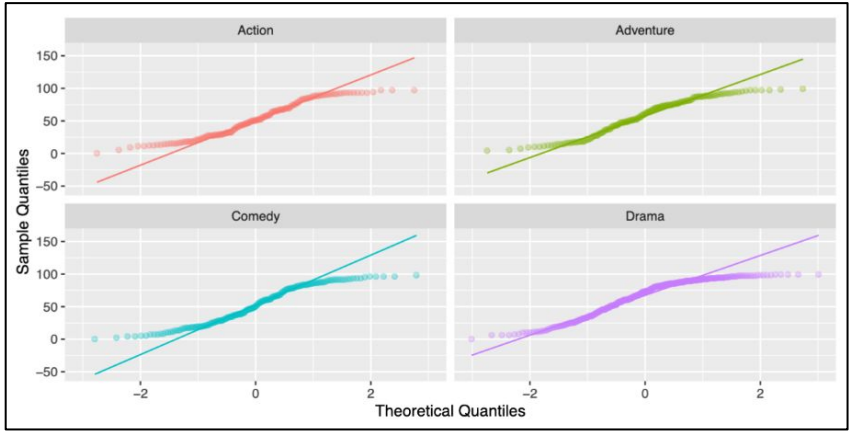


Assumptions for (Theory-Based) ANOVA

- **Assumption #1**: Observations are independent within and across groups
  - Think about study design/context (i.e., read the description)
- **Assumption #2**: Data within each group are approximately **normal**
  - Use **Normal Q-Q plots** (if data are perfectly normal, they follow the line in the Q-Q plot exactly)
  - As **sample size** increases, deviation from normality becomes less of a concern
- **Assumption #3**: **Variability** across groups is about equal
  - The rule of thumb is we want to see **largest variance / smallest variance < 3**, which we can find via data wrangling

Assumption #2: Normality

- Check via **Q-Q plot**, which plots residuals against theoretical quantiles of **normal distribution**
  - If residuals were perfectly **normally distributed**, they'd exactly follow the diagonal
  - We're not looking for perfect—just make sure it's reasonable
- Points should have a linear relationship, with no breaks at tails




```
ggplot(movies_subset, aes(sample = RottenTomatoes, col = Genre)) + geom_qq(alpha = 0.30) + stat_qq_line() + facet_wrap(~ Genre) + labs(y = "Sample Quantiles", x = "Theoretical Quantiles") + guides(col = "none")
```

Assumption #3: Constant Variability

- Check via **data wrangling**
- Remember **variance** is a measure of variability
- We don't expect the **variances** to be exactly the same across groups
  - As a rule of thumb, we want the ratio of largest variance to smallest variance to be less than 3
  - That is, **largest variance / smallest variance < 3**

```
## # A tibble: 4 x 3
##   Genre        var     n
##   <chr>      <dbl> <int>
## 1 Action      724.   170
## 2 Adventure   734.   163
## 3 Comedy      800.   191
## 4 Drama       680.   384
```

```
800/680
```

```
## [1] 1.176471
```

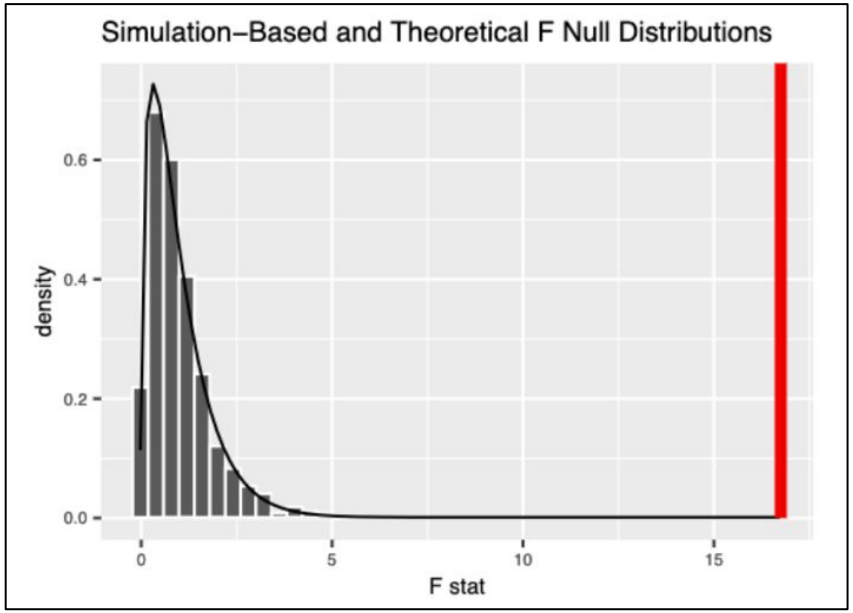
```
movies_subset %>% drop_na(Genre,
RottenTomatoes) %>% group_by(Genre) %>%
summarize(var = var(RottenTomatoes), n = n())
```

ANOVA: Code

- **Strategy #1**: Tidyverse R
  - `ANOVA_MODEL <- anova(lm(Y-VAR ~ X-VAR, data = DATASET))`
  - `tidy(ANOVA_MODEL)`
    - *`movies_anova <- anova(lm(RottenTomatoes ~ Genre, data = movies_subset))`*
    - *`tidy(movies_anova)`*
- **Strategy #2**: Base R
  - `ANOVA_MODEL <- aov(Y-VAR ~ X-VAR, data = DATASET)`
  - `summary(ANOVA_MODEL)`
    - *`movies_anova <- aov(RottenTomatoes ~ Genre, data = movies_subset)`*
    - *`summary(movies_anova)`*

ANOVA: More Intuition

- Remember, when assumptions are met, **F-statistic ~ F(df1, df2)**
- Remember, under $\mathbf{H_0}$, **F-statistic** should be equal to 1
- If **F-statistic** is much higher than 1, the variance between groups is much larger than the variance within groups, suggesting the $\mathbf{H_A}$
- "More extreme" is to the RIGHT!



ANOVA: Afterwards, How Do We Know Which Group Is Different?

- After seeing evidence against $\mathbf{H_o}$ (i.e., 1 of the means is different), how do we see which group is different?
- We'll conduct pairwise t-tests (like what we've been doing before)
- To keep Type I errors in check, we use adjusted alpha level, **α***
  - If we don't, the probability of a Type I error explodes as we do many pairwise t-tests!
- **$\alpha^* = \alpha/K$** , where **K is the total number of possible two-way comparisons**
  - **$K = k(k-1)/2$**, where **k is the total number of groups**
  - *Ex: If α = 0.05, then when there are 4 groups, α* = 0.05/6 = 0.0083*
- Our computers can calculate **α*** for us (the "bonferroni" correction)

ANOVA: Afterwards, Pairwise t-Tests

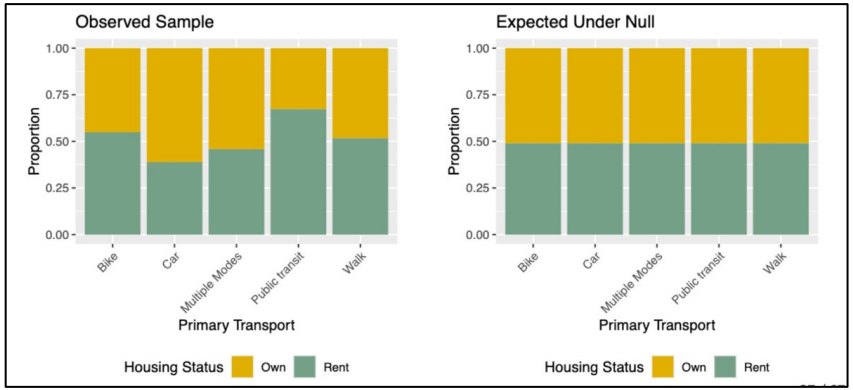
- **Pairwise t-Tests** isn't in tidyverse R, so we're using base R (with different syntax)!
- Remember to use "bonf" if you want to computer to calculate **α***
  - `pairwise.t.test(DATASET$Y-VAR, DATASET$X-VAR, p.adjust.method = "bonf")`
  - *`pairwise.t.test(movies_subset$RottenTomatoes, movies_subset$Genre, p.adjust.method = "bonf")`*

Chi-Squared

- **Chi-Squared test**: Test for when both **response variable** and **explanatory variable** are **categorical**, and **at least one has more than 2 categories**
  - $\mathbf{H_0}$: The variables are independent
  - $\mathbf{H_A}$: The variables are dependent
- If **response variable** and **explanatory variable** were both binary categorical, we'd just use **difference in proportions** (like before)!
- Our test statistic is χ^2 (which, essentially, sums and squares every z-score so that negatives are accounted for)
  - $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$
- The intuition is best explained by graphs and tables...

Chi-Squared: Intuition

- If **variables were actually independent** (i.e., primary transport doesn't affect housing status), then we'd expect a graph to look like the one on the right
- We'd also **expect** certain values
  - We expect (0.0389)(1534) = 59.72 residents who rent homes and use a bike, but we observe 67 residents
  - **Chi-squared** squares and sums these values to see "total extremity"

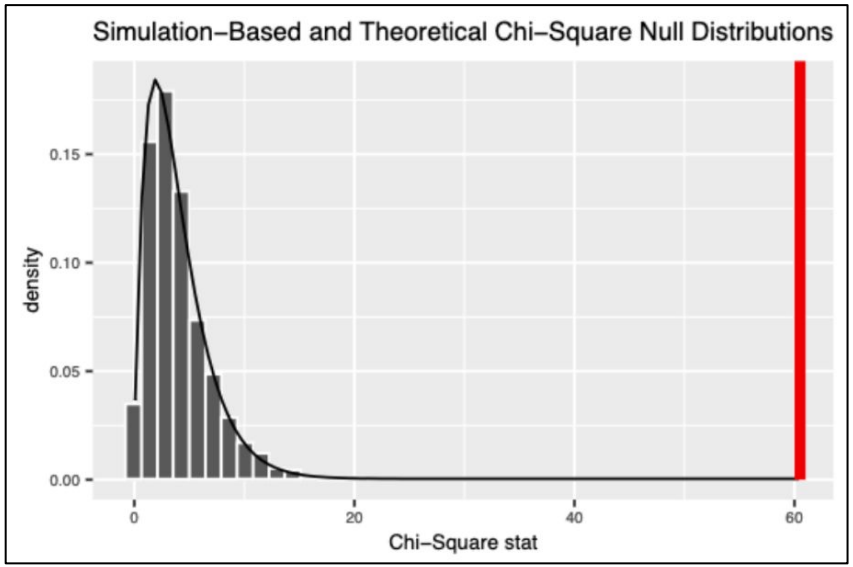


Assumptions for Chi-Squared

- **Assumption #1**: Random sampling
- **Assumption #2**: There are at least 10 observations in each cell (check via **data wrangling**)
  - `count(DATASET, X-VAR, Y-VAR)`
  - *`count(grammar, Education, oxford_comma)`*
- These assumptions must be met for the test statistic to be approximately distributed χ^2 with degrees of freedom $(r-1)(c-1)$, where **r** is the number of rows and **c** is the number of columns

Chi-Square: Intuition

- Our test statistic is χ^2, which essentially sums and squares every (standardized) difference between what we expect and what we observe
  - $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$
- $\chi^2 \sim \chi^2(df = (r - 1)(c - 1))$
  - r = number of rows
  - c = number of columns
- χ^2 quantifies how far the observed results deviate from what is expected under H_o
  - A larger value shows stronger evidence against H_o of independence (thus, "more extreme" is to the RIGHT!)



Chi-Squared: Code

- **Strategy #1**: Tidyverse R
  - `chisq_test(DATASET, Y-VAR ~ X-VAR)`
  - *`chisq_test(somerville, housing ~ primary_transport)`*
- **Strategy #2**: Base R
  - `chisq.test(DATASET$X-VAR, somerville$Y-VAR)`
  - *`chisq.test(somerville$primary_transport, somerville$housing)`*

Chi-Square: Afterwards, Examining Residuals

- We could compare the **observed** versus **expected values** to identify which table cells are contributing the most to the **test statistic**
- Instead of having to look back and forth between two tables, look at the table of **residuals**
- **Residuals** with a **large magnitude** contribute the most to the χ^2 **statistic**
  - If a **residual** is **positive**, the observed value is greater than the expected value
  - If a **residual** is **negative**, the observed value is less than the expected

Chi-Square: Afterwards, Examining Residuals: Code

- **General form**: `chisq.test(DATASET$X-VAR, DATASET$Y-VAR)$residuals`
  - *chisq.test(somerville$primary_transport, somerville$housing)$residuals*

Recap: Inference Scenarios and Test Statistics

https://drive.google.com/file/d/1i1XTclseg1_CPu6eECBuaoKDy6XuNMyn/view?usp=drive_link

Hypergeometric Distribution

- Recall a **random variable** is a mystery box that will "crystalize" to a certain value (with probabilities for each possible value)
  - E.g., if X is the number of STAT 102 students who show up to the ggparty, X could crystalize to 0, 1, 2, ..., 32
- Recall there are **famous types** of random variables, with certain "stories"
- If the story of your random variable matches a famous story, it makes your life easier (i.e., you know the PMF, expected value, etc.)
  - A **Binomial r.v.** counts the number of successes in n independent trials, each with a probability p of success
  - A **Hypergeometric r.v.** counts the number of "desired" draws when we **sample without replacement**

More on the Hypergeometric Distribution

- We draw **n** objects without replacement from a population of **N** objects, where **m** of those are desirable and **N - m** are undesirable… let X be the number of desirable objects we draw
- Thus, X ~ HGeom(m, N − m, n) by the "story" of the **Hypergeometric**
  - *We draw 8 marbles w/o replacement from a jar w/ 25 marbles, where 10 of those are white and 25 - 10 are black… if X is the number of white marbles we draw, then X ~ HGeom(10, 25 - 10, 8)*
- To calculate P(X = k)—probability X crystallizes to some value k—use the following code in R:
  - `dhyper(k, m, N - m, n)`
  - `dhyper(5, 10, 25 - 10, 8)`

Fisher's Exact Test

- **Fisher's exact test**: Test used to analyze whether there's a statistically significant association between **two categorical variables** in a 2x2 contingency table
  - Can be done even when sample size is small (unlike **chi-squared test**!)
- Under H_0, $p_1 = p_2$
  - From the example in class, this would imply individuals in one treatment group are just as likely to be cured as those in the other group
  - If this were true, what is the probability we observe our result (that of the 17 cured individuals, 13 were in the treatment group)? What is the probability we get a result more extreme (here, that 14 were in the treatment group, 15, 16, ...)
  - The **p-value** adds these probabilities (as or more extreme than the one observed) under H_0

Fisher's Exact Test and the Hypergeometric Distribution

- Assuming H_0, given that 17 individuals out of 29 were cured and that 16 individuals were in the fecal infusion group, what is the probability that 13 of the cured individuals were in the fecal infusion group?
- X ~ HGeom(17, 29 - 17, 16) by the "story" of the Hypergeometric
  - P(X = 13) = `dhyper(13, 17, 29 - 17, 16)` = 0.007715441
- To find the **p-value**, we need to add probabilities of the results as or more extreme than those observed
  - One-sided p-value = `phyper(12, 17, 29 - 17, 16, lower.tail = FALSE)` = 0.008401063

Fisher's Exact Test: Code

- Alternatively, we can also use the function `fisher.test()` after defining a table via `matrix()` and `dimnames()`
- Pipe it into `tidy()`, which will give you all the information, including p-value

Treating *C. difficile* infection...

```
#enter the data
infusion.table = matrix(c(13, 3, 4, 9), nrow = 2, ncol = 2, byrow = T)
dimnames(infusion.table) = list("Outcome" = c("Cured", "Uncured"),
                                "Treatment" = c("Fecal Infusion",
                                                "Vancomycin"))

fisher.test(infusion.table, alternative = "greater") %>%
  tidy()
```

```
## # A tibble: 1 x 6
##   estimate p.value conf.low conf.high method                          alternative
##      <dbl>   <dbl>    <dbl>     <dbl> <chr>                           <chr>
## 1     8.85 0.00840     1.74       Inf Fisher's Exact Test for Count~ greater
```

Treating *C. difficile* infection...

```
##      k     prob
## 1    0 0.000000
## 2    1 0.000000
## 3    2 0.000000
## 4    3 0.000000
## 5    4 0.000035
## 6    5 0.001094
## 7    6 0.012036
## 8    7 0.063046
## 9    8 0.177317
## 10   9 0.283708
## 11  10 0.264794
## 12  11 0.144433
## 13  12 0.045135
## 14  13 0.007715
## 15  14 0.000661
## 16  15 0.000024
## 17  16 0.000000
```

```
#P(X leq 5) + P(X geq 13)
phyper(5, 17, 29 - 17, 16) +
  phyper(12, 17, 29 - 17, 16, lower.tail = F)
```

```
## [1] 0.009530323
```

```
#two-sided p-value
fisher.test(infusion.table) %>%
  tidy() %>%
  select(-method)
```

```
## # A tibble: 1 x 5
##   estimate p.value conf.low conf.high alternative
##      <dbl>   <dbl>    <dbl>     <dbl> <chr>
## 1     8.85 0.00953     1.37      78.8 two.sided
```

Questions?

P-Set 9

Thanks for an amazing semester! Wishing you all the best of luck on the final 😁