STAT 100: Week 12

Ricky's Section

Introductions & 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

Anonymous Feedback

https://docs.google.com/forms/d/e/1FAIpQLSfKv_FGvs0oqm-IvtxKx3Vf6bBzSJE2jamK1gklAzL6NkXE8w/viewform

Upcoming Events...

- **Optional Review**: Monday, 11/25
- **Last day of class for STAT 100**: Wednesday, 12/4
- **ggparty**: Thursday, 12/5 from 11:30 AM to 1:30 PM in Science Center 316
  - RSVP: https://forms.gle/yKw6Wziiy6Lj5guu6
- **Jude's review session**: Thursday, 12/5 (after the **ggparty**)
- **Final exam for STAT 100**: Wednesday, 12/11

Another Workshop (Review Session)

- **Today, 11/20 from 4:30-6 PM in Science Center Hall A!**
- We'll be practicing and reviewing linear regression
  - Same material as my Workshop last week but with different TFs
- We recognize flipped classroom can be difficult, so these are here to help you
  - Along with OH, 1-on-1 OH, Slack, etc.
- Don't wait last minute to study for the Final!

Modified Office Hours

- The OH spreadsheet (on Canvas) has sections for Modified Office Hours on account of the Thanksgiving Break and Reading Period
- https://docs.google.com/spreadsheets/d/1VjMVcc2Ps_sFG2_EhtdpV1ACysbs0Fj4AkI3uYF9Ge8/edit?gid=1087661833#gid=1087661833

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

Picture!

Content Review: Week 12

Big Picture Overview

- We're introducing 3 more tools in our inference toolkit
- These are extensions of things we've seen before
  - Paired t-Test
  - ANOVA
  - Chi-Squared

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!*
- 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
  - This is 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
- **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 wetsuit versus a swimsuit 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 wetsuit versus a swimsuit 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

Paired *t*-test: Code

- As always, our computers do the math for us—we just need to code and interpret!
- **Strategy #1**: Use `t_test()`
  - We're used to this from the tidyverse
  - A paired *t*-test is just a single-mean test on the differences
- **Strategy #2**: Use simulation-based inference
  - Again, very similar to before when we did simulation-based inference for a single mean
  - This isn't my recommended strategy, but if you're curious, check out Slide 19 in Week 12!
- **Strategy #3**: Use `t.test()`
  - This is base R while `t_test()` is tidyverse R, so `t.test()` uses differing syntax
  - This is NOT expected for STAT 100, but if you're curious, check out Slides 17-18 in Week 12!

Paired t-Test: Code for t_test()

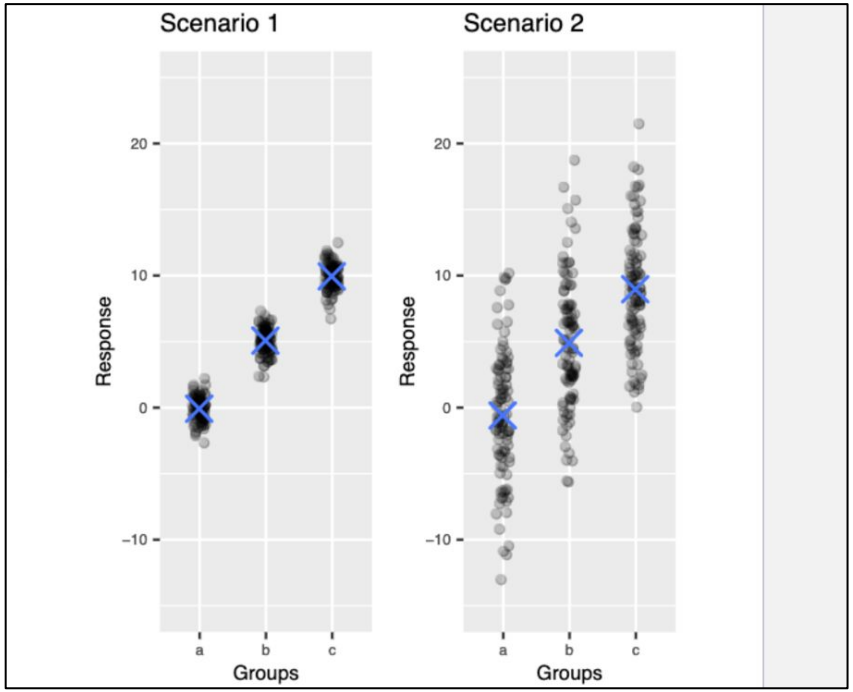
- **General form**: DATASET %>% t_test(response = RESPONSE-VAR.diff)
  - Again, very similar to before, but now we're adding ".diff" because we're interested in the difference for each pair
- **Hypothesis tests**: DATASET %>% t_test(response = RESPONSE-VAR.diff) %>% select(statistic, p_value, estimate)
  - *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)**
  - $\mathbf{H_o}$: $\mu_1 = \mu_2 = \ldots = \mu_k$ (variables are independent)
  - $\mathbf{H_A}$: At least 1 mean is not equal to the rest (variables are dependent)
- Test statistic is **F-statistic**
  - **F = standardardized variance BETWEEN groups / standardized variance WITHIN groups**
- If $\mathbf{H_o}$ is true, **F-statistic** should be roughly equal to 1 (variance between groups should be equal to variance within groups)
- If $\mathbf{H_A}$ 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*

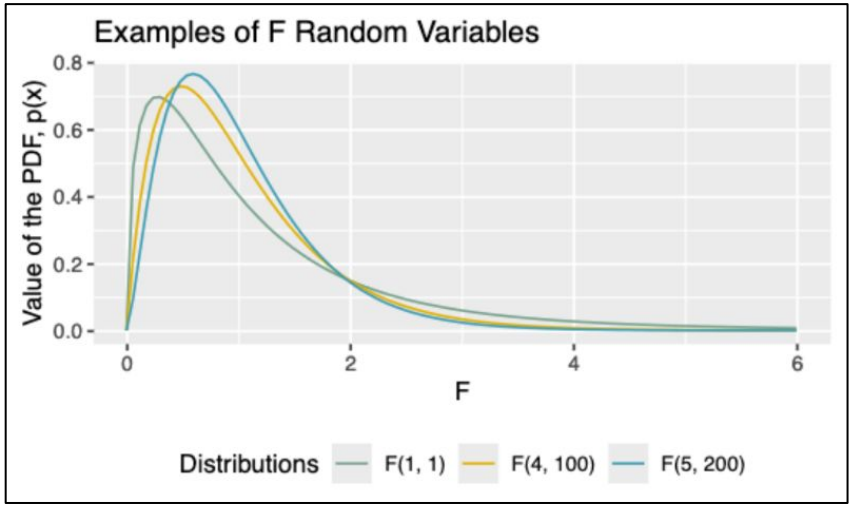
ANOVA: Intuition

- **Scenario 1**, there is little variability WITHIN groups but much more variability BETWEEN groups
  - It's plausible these groups come from different populations



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($\mathbf{df_1}$, $\mathbf{df_2}$)**
  - $\mathbf{df_1 = n_{groups} - 1}$, $\mathbf{df_2 = n_{observations} - n_{groups}}$

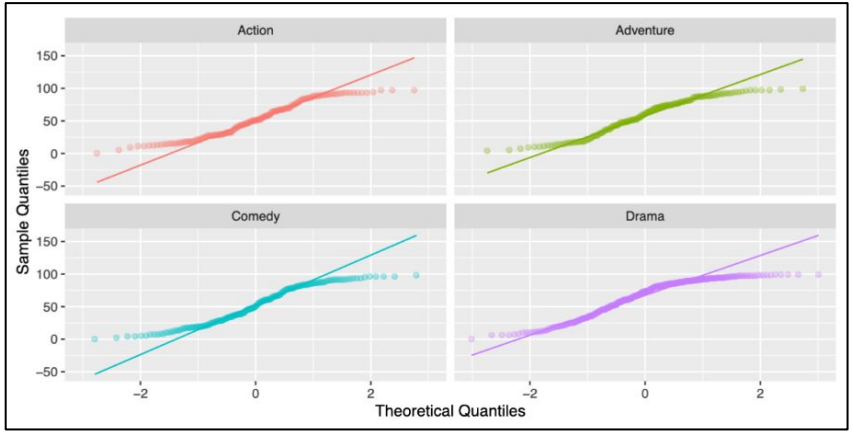


Assumptions for (Theory-Based) ANOVA

- **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**
  - As sample size increases, deviation from normality becomes less of a concern
- **Assumption #3**: **Variability** across groups is about equal
  - We want to see **largest variance / smallest variance < 3**

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
```

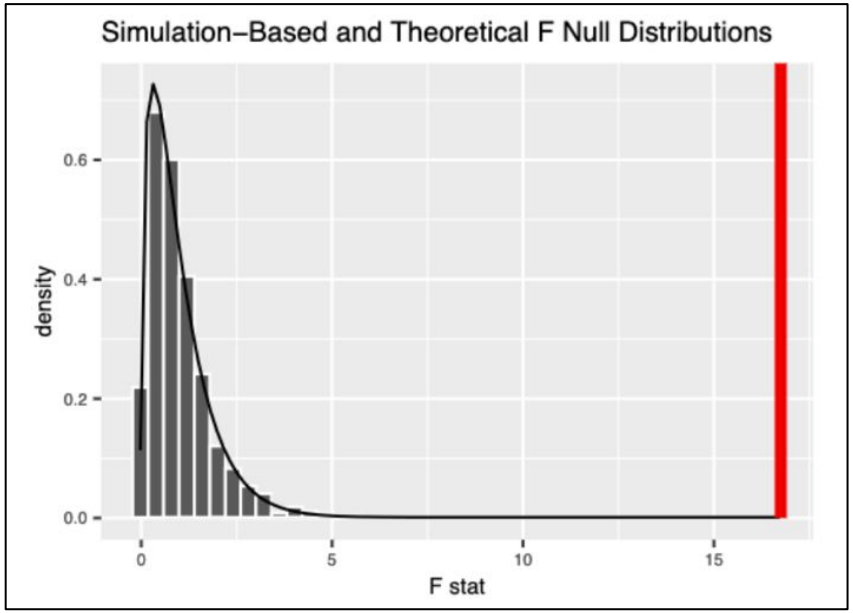
```
movies_subset %>% drop_na(Genre, RottenTomatoes) %>% group_by(Genre) %>% summarize(var = var(RottenTomatoes), n = n())
```

ANOVA: Code

- **Strategy #1**: Tidyverse R
  - movies_anova <- anova(lm(RottenTomatoes ~ Genre, data = movies_subset))
  - tidy(movies_anova)
- **Strategy #2**: Base R
  - 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 the variance between groups is much larger than the variance within groups, suggesting the $\mathbf{H_A}$



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***
- **$\alpha^* = \alpha/K$** , 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, 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(movies_subset$RottenTomatoes, movies_subset$Genre, p.adjust.method = "bonf")

Chi-Squared

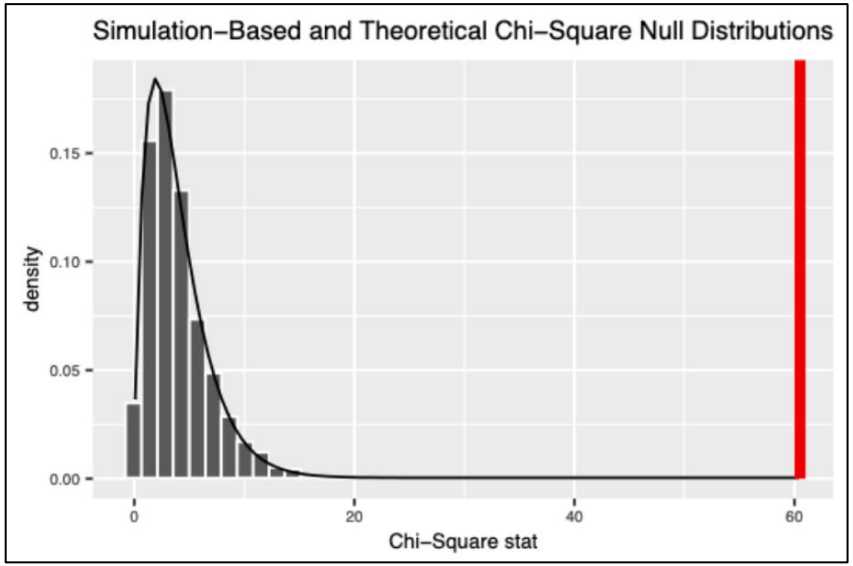
- **Chi-Squared test**: Test for when both **response variable** and **explanatory variable** are **categorical**, and **at least one has more than 2 categories**
  - $\mathbf{H_0}$: The variables are independent
  - $\mathbf{H_A}$: The variables are dependent
- If **response variable** and **explanatory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

Assumptions for Chi-Squared

- **Assumption #1**: Random sampling
- **Assumption #2**: There are at least 10 observations in each cell (check via **data wrangling**)
  - count(DATA-SET, EXPL-VAR, RESPONSE-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 z-score so that negatives are accounted for)
  - $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$
- $\chi^2 \sim \chi^2(df = (r - 1)(c - 1))$ when assumptions are met



Chi-Squared: Code

- **Strategy #1**: Tidyverse R
  - chisq_test(somerville, housing ~ primary_transport)
- **Strategy #2**: Base R
  - 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

- **Strategy #1**: Tidyverse R
  - chisq.test(somerville$primary_transport, somerville$housing) %>% augment() %>% select(-.prop, -.row.prop, -.col.prop, -.std.resid)
- **Strategy #2**: Base R
  - chisq.test(somerville$primary_transport, somerville$housing)$residuals

Recap: Inference Scenarios and Test Statistics

https://drive.google.com/file/d/1rvVsTfhaK_92yWn8DTp-f97SF3tPmkEr/view?usp=sharing

Oral Exam Practice

Person A (Grade Q1 and Q3, Answer Q2 and Q4)

https://docs.google.com/document/d/12xYi5gwnX8UVGKpS3T20K7ywSKGMnpoppPT-1d-ftDQ/edit?usp=sharing

Person B (Grade Q2 and Q4, Answer Q1 and Q3)

https://docs.google.com/document/d/1SvjvF5KS-ocbTUzVJJ4FGL2JTUlEtfKDS99KJv9l9Es/edit?usp=sharing

Solutions

https://docs.google.com/document/d/1IIBXXc7ysfQ_mkwivA42OhClzpoVVXqHqREeWMdQ6Yc/edit?usp=sharing

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