Final Exam Review

STA 209, Spring 23

For extra practice, several review problems are provided on this Review Sheet. However, they are not comprehensive, so do not limit your studying to just these problems. While the problems are intended to match the difficulty of those on the exam, the length of this review set does not represent the actual length of the exam. Although detailed solutions will not be posted, you are welcome to talk to me about any of them before the test.

Problem 1

Run the following code to load the `HollywoodMovies` data set.

```
library(Lock5Data)
```

```
## Warning: package 'Lock5Data' was built under R version 4.0.5
```

```
data(HollywoodMovies)
```

This data set contains information on the opening weekend gross income (`OpeningWeekend`), Rotten Tomatoes critics rating (`RottenTomatoes`), the Rotten Tomatoes audience score (`AudienceScore`), the number of theaters showing the movie on opening weekend (`TheatersOpenWeek`), the production budget (`Budget`) and the year the movie was released (`Year`).

a. Create a multilinear model predicting gross opening weekend income using `RottenTomatoes`, `AudienceScore`, `TheatersOpenWeek`, `Budget` and `Year`. Display the output using `get_regression_table` and write down the model's equation.
b. Interpret the slopes of each explanatory variable in context of the model.
c. Create appropriate data visualizations to assist in determining whether inference is reasonable for this multilinear model.
d. Determine which explanatory variables are significant predictors of the opening weekend gross.
e. For any variables that are not significant predictors, create a simple linear regression model for that variable and the response. Is the variable a significant predictor in the simple linear model? If so, explain why this doesn't contradict your findings in the previous part.

---

a.

```
movie_mod <- lm(OpeningWeekend ~ RottenTomatoes + AudienceScore + TheatersOpenWeek + Budget + Year, data

get_regression_table(movie_mod)
```

```
## # A tibble: 6 x 7
##   term             estimate std_error statistic p_value  lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1 intercept        -217.      646.       -0.336   0.737 -1484.    1050.
## 2 RottenTomatoes      0.138     0.033     4.17    0         0.073    0.202
## 3 AudienceScore       0.218     0.05      4.38    0         0.121    0.316
## 4 TheatersOpenWeek    0.008     0.001    13.6     0         0.007    0.009
## 5 Budget                0.262     0.014    18.5     0         0.234    0.29
## 6 Year                  0.091     0.321     0.285   0.776    -0.538    0.72
```

$$\text{OpeningWeekend} = -216.874+0.138\cdot\text{RottenTomatoes}+0.218\cdot\text{AudienceScore}+0.008\cdot\text{TheatersOpenWeek}+0.262\cdot\text{Budget}+0.091$$

b. The slope of each variable indicate by how much the response the OpeningWeekend gross changes, per 1 unit change in the explanatory variable, while all other variables in the model are held constant.

c.

Linearity: The residual plot shows some evidence of non-linearity, as the mean residual follows a slight U shape from left to right.

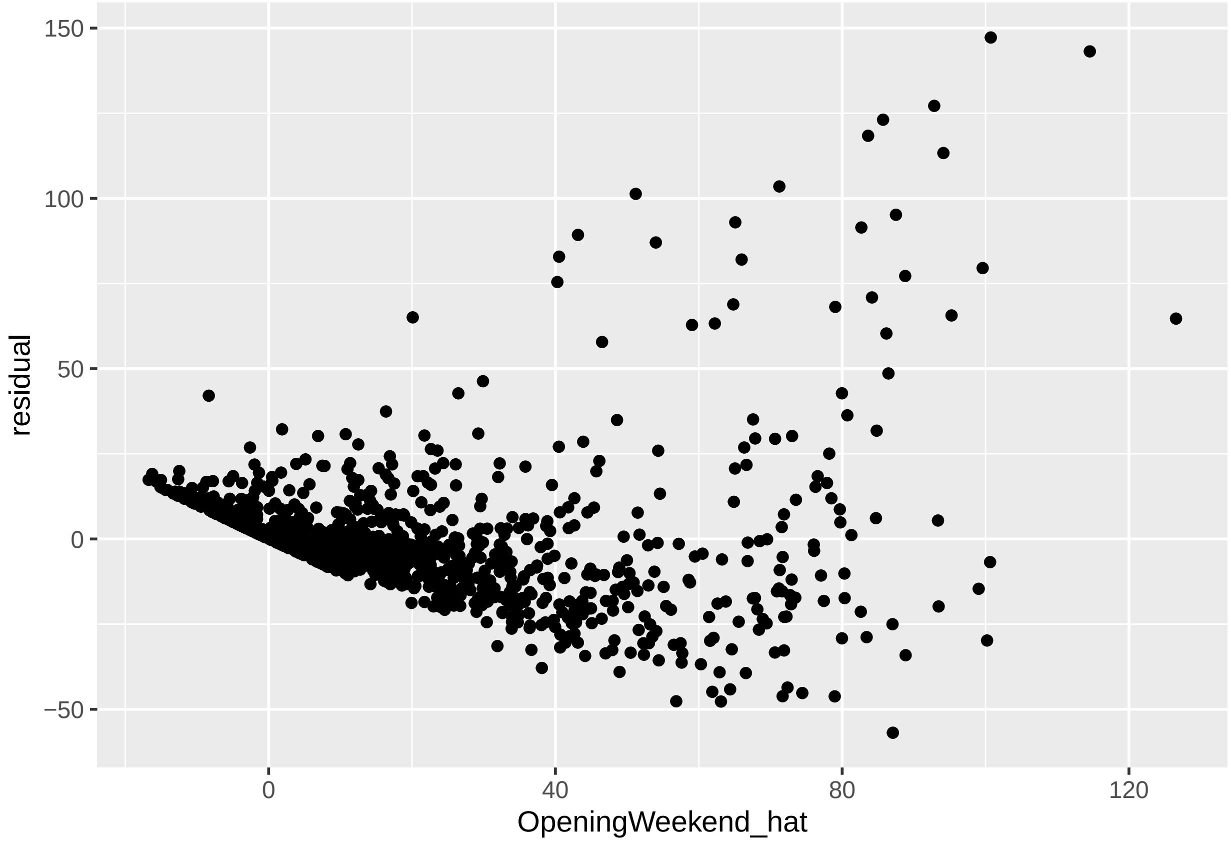
Independence: We aren't explicitly told how this sample was collected. However, we may imagine that movies released in similar time frames may have similar box office gross, and hence, residuals might not be independent.

Normality: The histogram of residuals shows significant right skew, suggesting residuals are not Normally distributed.

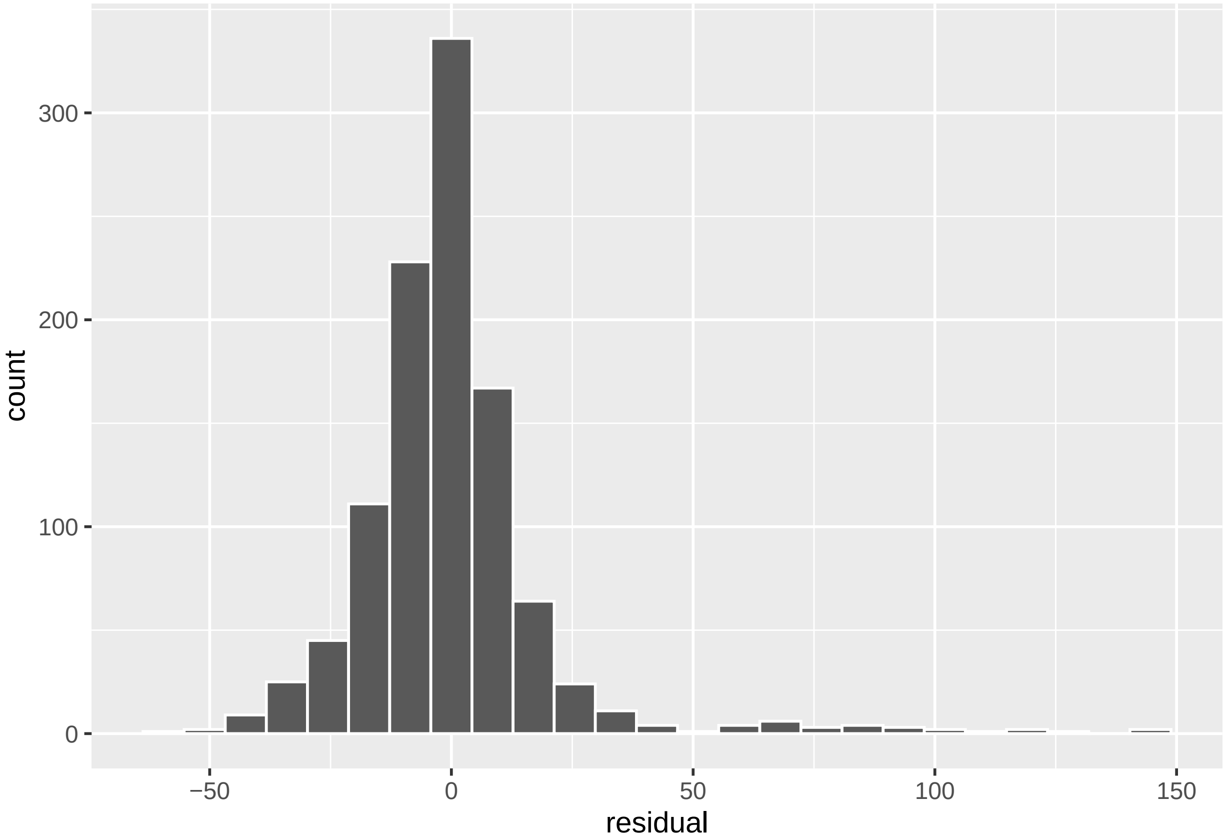
Equal Variability: The residual plot shows clear signs of increasing variability in residuals as OpeningWeeknd gross increases.

```r
movie_res <- get_regression_points(movie_mod)

ggplot(movie_res, aes(x =  OpeningWeekend_hat, y = residual))+geom_point()
```



```r
ggplot(movie_res, aes(x = residual))+geom_histogram(bins = 25, color = "white")
```



d. Every predictor except Year is significant at the 0.001 level.

e. Even in the simple linear regression model, year is not a significant predictor of Opening Weekend.

```r
year_mod <- lm(OpeningWeekend ~ Year, data = HollywoodMovies)
get_regression_table(year_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value  lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1 intercept -280.      845.       -0.332   0.74  -1938.     1377.
## 2 Year         0.148     0.419     0.353   0.724    -0.674     0.97
```

Problem 2

The dataset HomesForSaleCA contains a random sample of 30 houses for sale in California. We are interested in whether there is a positive association between the number of bathrooms and number of bedrooms in each house.

```r
library(Lock5Data)
data("HomesForSaleCA2e")
```

a. Fit a simple linear model and write down the regression equation, treating Baths as the response and Beds as the explanatory variables.

b. How many baths does the model predict a home with 3 bedrooms will have?

c. Create a scatterplot of the relationship between `Baths` and `Beds`. Be sure to account for overplotting.

d. Produce a histogram of the distribution of residuals. Comment on whether the normality condition is satisfied.

e. Produce a residual plot and comment on whether the residuals demonstrate constant variability.

f. Use theory-based methods to test the claim that there is no linear relationship between the variables.

g. Use `infer` to create a 95% confidence interval for correlation between `Baths` and `Beds`. Is it plausible that the two variables have strong positive correlation?

---

a.

```
house_mod <- lm(Baths ~ Beds, data = HomesForSaleCA2e)
get_regression_table(house_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -0.068     0.423    -0.161   0.874   -0.934    0.799
## 2 Beds         0.794     0.124     6.38    0         0.539    1.05
```

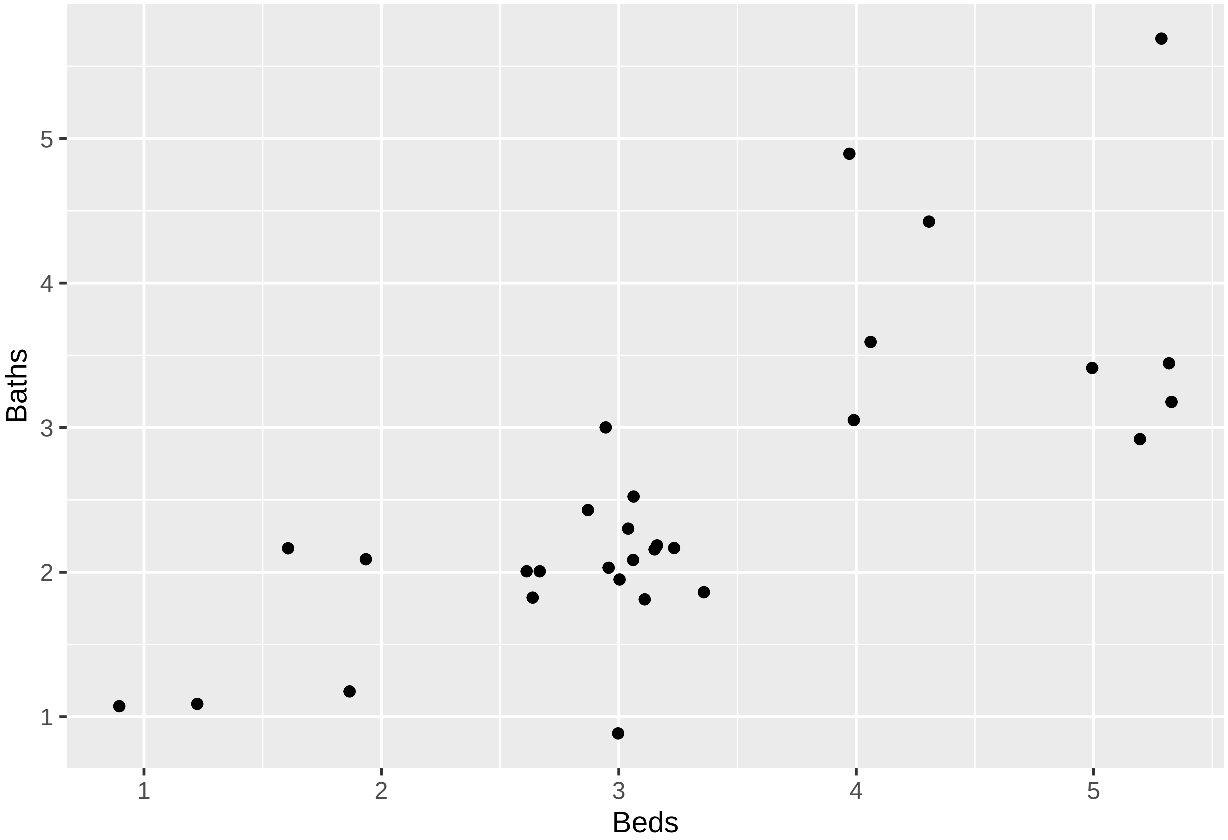
b. The model predicts that a home with 3 bedrooms will have 2.314 bathrooms.

```
-0.068+0.794*3
```

```
## [1] 2.314
```

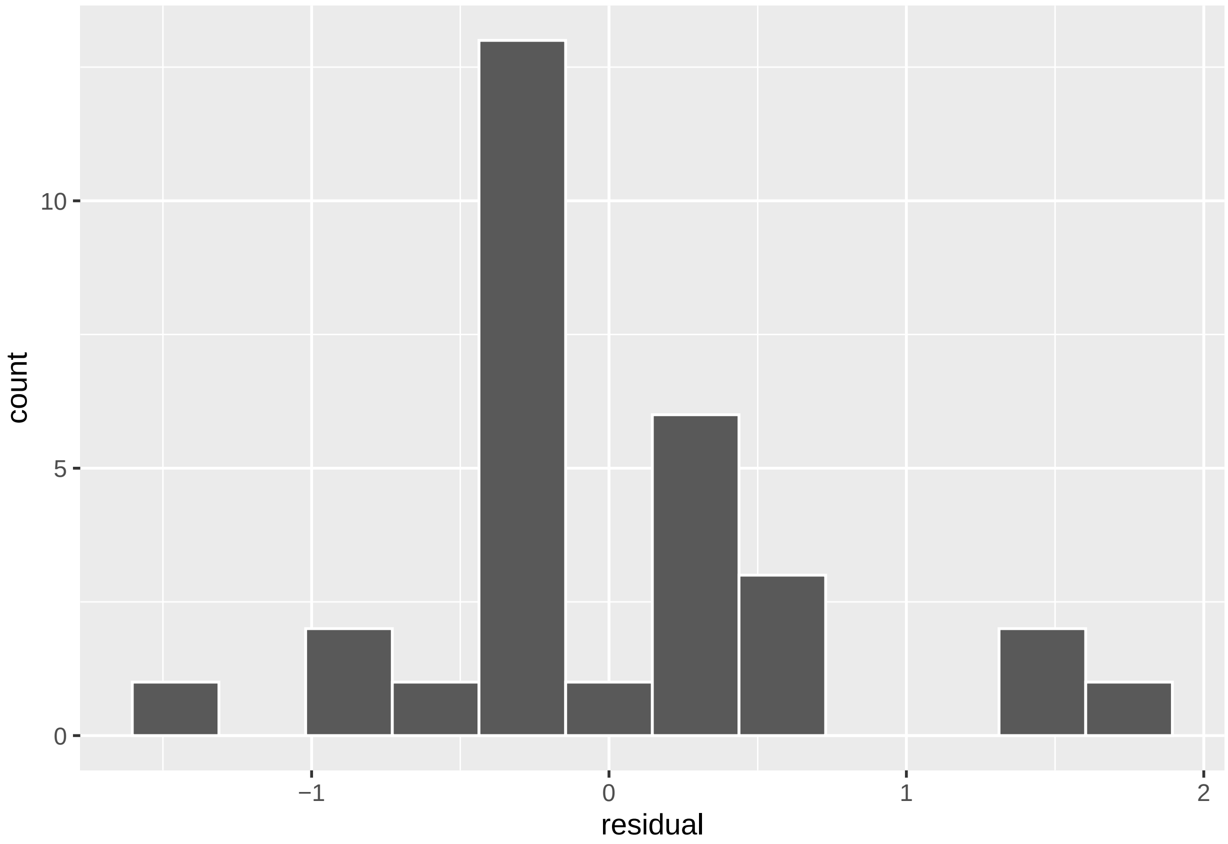
c.

```
ggplot(HomesForSaleCA2e, aes(x = Beds, y= Baths))+geom_jitter()
```



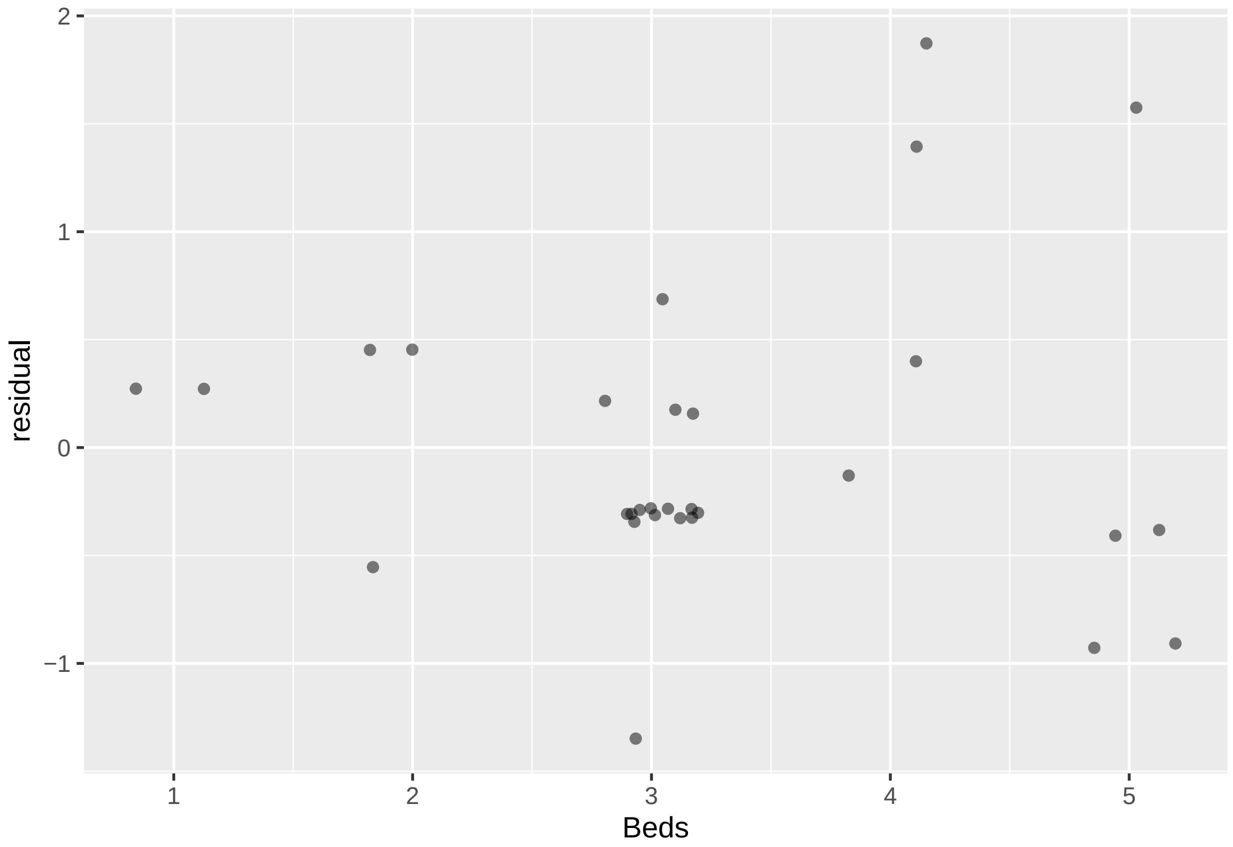
d. The histogram of residuals shows some signs of bimodality, and hence, suggests that the residuals may not be Normally distributed. However, it is difficult to assess Normality with only 30 data points.

```
house_res <- get_regression_points(house_mod)
ggplot(house_res, aes(x = residual))+geom_histogram(bins = 12, color = "white")
```



e. The residuals do not appear to have constant variability. In particular, residuals for houses with fewer bedrooms. have less variability than those for houses with more bedrooms.

```
ggplot(house_res, aes(x = Beds, y = residual))+geom_jitter(width = .2, alpha = .5)
```



f. Based on the regression table, the p-value is approximately 0 for the test of $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. This gives very strong evidence that the slope of the model is non-zero.

```
get_regression_table(house_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -0.068     0.423    -0.161   0.874   -0.934    0.799
## 2 Beds         0.794     0.124     6.38    0         0.539    1.05
```

g.

```
set.seed(221)
boot_dist <- HomesForSaleCA2e %>%
  specify(Baths ~ Beds) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "correlation")
```

```
boot_dist %>%
  get_ci(level =.95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.672    0.869
```

Problem 3

How accurate are lie detectors? The following data involve participants who read either deceptive material or truthful material while hooked to a lie detector. Run the following code to generate the data:

```r
lie_detector<- tibble( lying = rep(c( "yes", "no"), c(48, 48 )),
detector = rep( c("positive", "negative", "positive", "negative"),c(31, 17, 27, 21 )))
```

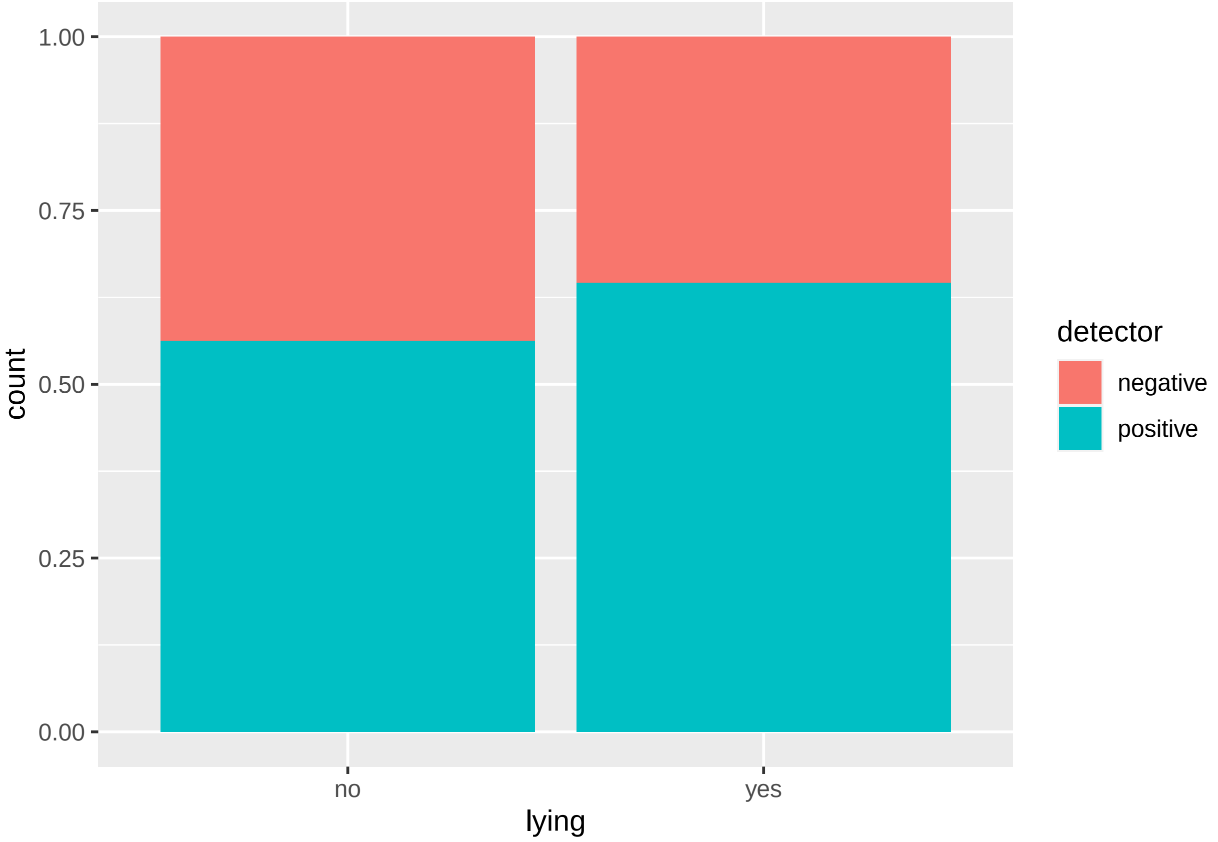
The `lying` variable indicates whether the person read deceptive material, while the `detector` variable indicates whether the detector said the person was lying.

a. Create an appropriate graphic comparing the proportion of people the detector indicated were lying in each group.
b. Find and interpret the 90% confidence interval for the proportion of times the lie detector accurately detects a person lying.
c. Test to see if there is evidence that the lie detector says a person is lying more than 50% of the time, regardless of what the person reads.
d. Test to see if there is a difference in the proportion the lie detector says is lying depending on whether the person is lying or telling the truth.
e. Find and interpret the 95% confidence interval for the difference in the proportion the lie detector says is lying between those lying and those telling the truth.

---

a.

```r
ggplot(lie_detector, aes(x = lying, fill = detector))+geom_bar(position = "fill")
```



b. We first create a new data frame just consisting of those individuals who were lying:

```r
liars <- lie_detector %>%
  filter(lying == "yes")
```

```r
set.seed(323)

liars %>%
  specify(response = detector, success = "positive") %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = .90, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.521     0.75
```

With 90% confidence, the true proportion of liars who the detector correct identifies is between 0.52 and 0.75.

c. Using `infer`, we obtain a p-value of approximately 0.03. At the 0.05 level, this is sufficient evidence to reject the null hypothesis.

```r
obs_prop <- lie_detector %>%
  specify(response = detector, success = "positive") %>%
  calculate(stat = "prop")
```

```r
set.seed(2112)
lie_detector %>%
  specify(response = detector, success = "positive") %>%
  hypothesize(null = "point", p = .5) %>%
  generate(reps = 2000, type = "simulate") %>%
  calculate(stat = "prop") %>%
  get_p_value(obs_stat = obs_prop, direction = "right")
```

```
## The `"simulate"` generation type has been renamed to `"draw"`. Use `type = "draw"` instead to quiet
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0305
```

d. Using `infer`, we obtain a p-value of approximately 0.526, giving insufficient evidence to reject the null hypothesis at the .05 level.

```r
obs_prop_2 <- lie_detector %>%
  specify(response = detector, explanatory = lying, success = "positive") %>%
  calculate(stat = "diff in props", order = c("yes", "no"))
```

```r
set.seed(331)
lie_detector %>%
  specify(response = detector, explanatory = lying, success = "positive") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 2000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("yes", "no")) %>%
  get_p_value(obs_stat = obs_prop_2, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##      <dbl>
## 1   0.526
```

e. Using `infer`, the 95% confidence interval is (-.116, 0.277)

```r
set.seed(331)
lie_detector %>%
  specify(response = detector, explanatory = lying, success = "positive") %>%
  generate(reps = 2000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("yes", "no")) %>%
  get_ci(level = .95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.116    0.277
```

Problem 4

For each situation given below, identify which of the listed tests is most appropriate. If multiple tests are appropriate, list them all.

Tests

- Chi-square test for goodness of fit.
- Chi-square test for independence.
- ANOVA for difference in means
- Test for correlation
- Test for slope of multiple linear regression model
- Test for slope of simple linear regression Model
- 2-sample test for difference in means
- 2-sample test for difference in proportions
- 1-sample test for mean
- 1-sample test for proportion

Situations

a. Three different drugs are being tested on patients who have leukemia and the response variable is white blood cell count.

b. Researchers want to determine whether phone owners do most of their online browsing on their phone. 200 randomly selected phone owners will be surveyed, and the response variable will be whether the respondent use phone for a majority of online browsing.

c. Three different drugs are being tested on patients who are HIV-positive and the respond variable is whether or not the person develops AIDS.

d. Data are collected from 50 towns on number of wood-burning houses and number of people with asthma, and the study is investigating whether there is a linear relationship between the two.

e. Lawmakers want to determine whether support for capital punishment in the US has changed over the last 30 years. They survey a random sample of 2815 individuals and ask whether each favors capital punishment, and compares the results to the responses to the same question in a 1974 survey of 1410 random individuals.

f. The admissions office at a university uses data from high school transcripts such as number of honors courses, number of AP courses, grade in 11th grade English, grade in 9th grade math to develop a model to predict college GPA.

g. A polling agency working in a large city knows the distribution of ethnicity in the city population. They select a sample of 2000 residents and would like to check that the distribution of ethnic groups within their sample is not significantly different from the proportion in the city as a whole.

h. A sabermetrics group wants to know whether American football teams benefit from "Home Field Advantage.' ' A random sample of 20 games from the 2018 regular season is selected, and the group measures the average number of points scored by the home team. In a separate sample of 20 games, the group measured the average number of points scored by the away team.

i. A test is being conducted to see if the average time it takes for a case to go to trial differs between counties in a state. Seven counties will be included and the data will include a random sample of 25 cases from each county.

j. The FDA recommends that the typical adult consume at least 25 grams of fiber per day. To determine whether Americans are following this advice, statisticians survey 200 Americans and ask each to estimate their typical daily fiber consumption.

k. A test is being conducted to see if the proportion of cases that get settled out of court is different between the different counties in the state. Seven counties will be included and the data will include a random sample of 60 cases from each county.

---

a. ANOVA test.

b. 1-sample test for proportion

c. Chi-square test for independence.

d. Test for slope of simple linear regression Model

e. 2-sample test for difference in proportions

f. Test for slope of multiple linear regression model

g. Chi-square test for goodness of fit.

h. 2-sample test for difference in means

i. ANOVA for difference in means

j. 1-sample test for mean

k. Chi-square test for independence.

Problem 5

The Comprehensive Assessment of Outcomes in Statistics (CAOS) exam is an online multiple-choice test on concepts covered in a typical introductory statistics course. Students can take a pretest version before instruction and then a posttest version after instruction. Scores on the pretest and posttest for a random sample of 10 students with one instructor can be obtained by running the following code:

```
library(Lock5Data)
data(CAOSExam)
```

a. Was the data gathered using matched-pairs design, or by collecting 2 independent samples?

b. What information would you need to know about the population in order to ensure it is appropriate to use theory-based methods for statistical inference on this sample?

c. Compute and interpret a 95% confidence interval for the improvement in mean CAOS scores between the two exams.

d. The developers of the CAOS exam give benchmark data based on a very large number of students taking the pretest and posttest. The mean score on the pretest was 44.9 and the mean on the posttest was 54.0. Treat these values as population means, is there evidence at the 5% level that the instructor's students have a mean score on the posttest that is higher than the overall average?

e. Using the parameters from the previous part, can we conclude that the instructor began with stronger students than average?

f. Finally, can we conclude, at the 5% significance level, that the mean *improvement* from pretest to posttest for students with this instructor is higher than the overall norm?

---

a. The data appears to come from a matched pairs design, since each student contributed a pre and post-test score

b. Since the sample size is small (10 cases), we need to know whether scores are approximately Normally distributed.

c. Using Theory-based methods, the 95% confidence interval for the mean difference is (9.44, 18.56)

```
CAOSExam %>% mutate(diff = Posttest - Pretest) %>% summarize(mean_diff = mean(diff), sd_diff = sd(diff)
```

```
##   mean_diff  sd_diff  n
## 1        14 6.368324 10
```

```
t_star <- qt(.975, df = 9)

lower <- 14 - t_star*6.37/sqrt(10)
upper <- 14 + t_star*6.37/sqrt(10)

lower
```

```
## [1] 9.443177
```

```
upper
```

```
## [1] 18.55682
```

d. Let μ be the mean of the posttest scores for these students. We test the hypotheses:

$$H_0 : \mu = 54 \qquad H_a : \mu > 54$$

The mean posttest score for these students was 60.25, with standard deviation 9.96.

```
CAOSExam %>% summarise(mean = mean(Posttest), sd = sd(Posttest))
```

```
##    mean       sd
## 1 60.25 9.961732
```

Using theory-based methods, the test statistic is 1.98

```
test_stat <- (60.25 - 54 )/(9.96/sqrt(10))
test_stat
```

```
## [1] 1.984361
```

and the associated p-value is 0.039. This is sufficient evidence to reject H_0 at the 0.05 level, although not at the 0.01 level. The sample gives moderately strong evidence that the instructor's students have a mean score on the posttest that is higher than the overall average.

```
1-pt(test_stat, df = 9)
```

```
## [1] 0.03925321
```

e. Using the parameters from the previous part, can we conclude that the instructor began with stronger students than average?

Let μ be the mean of the pretest scores for these students. We test the hypotheses:

$$H_0 : \mu = 44.9 \qquad H_a : \mu > 44.9$$

The mean pretest score for these students was 46.25 , with standard deviation 9.30.

```
CAOSExam %>% summarise(mean = mean(Pretest), sd = sd(Pretest))
```

```
##    mean       sd
## 1 46.25 9.298297
```

Using theory-based methods, the test statistic is 0.46

```
test_stat <- (46.25 - 44.9 )/(9.30/sqrt(10))
test_stat
```

```
## [1] 0.4590403
```

and the associated p-value is 0.329. This is not sufficient evidence to reject H_0 at the 0.10 level. The sample does not give us evidence that the instructor began with stronger students than average.

```
1-pt(test_stat, df = 9)
```

```
## [1] 0.3285452
```

f. The mean difference in the population is $54 - 44.9 = 9.1$. We'll test the hypotheses

$$H_0 : \mu = 9.1 \qquad H_a : \mu > 9.1$$

Using theory-based methods, our test statistic is 2.43

```
test_stat <- (14 - 9.1)/(6.37/sqrt(10))
test_stat
```

```
## [1] 2.432521
```

and the associated p-value is 0.018. This is sufficient evidence to reject H_0 at the 0.05 level, although not at the 0.01 level. The sample gives moderately strong evidence that improvement from pretest to posttest for students with this instructor is higher than the overall norm.

```
1-pt(test_stat, df = 9)
```

```
## [1] 0.0189129
```

Problem 6

A survey was given to a sample of high school seniors in Pennsylvania between 2010 and 2019, and includes many different variables. The data is in `PASeniors`. *Some light data wrangling was performed to remove 4 students that did not provide an answer for which superpower they would prefer*

```
library("Lock5Data")
PASeniors <- PASeniors %>%
  filter(Superpower != "") %>%
  mutate(Superpower = droplevels(Superpower))
```

One of the variables in the survey asks the students to indicate which in a list of five superpowers (`Superpower`). The results are shown in the table below, aggregated by self-reported gender (`Gender`).

```
table(PASeniors$Superpower, PASeniors$Gender) %>% addmargins()
```

```
##
##                   Female Male Sum
##   Fly                 51   69 120
##   Freeze time         58   71 129
##   Invisibility        39   41  80
##   Super strength       5   14  19
##   Telepathy           68   37 105
##   Sum                221  232 453
```

a. Create an appropriate visualization showing the relationship between `Superpower` and `Gender`.

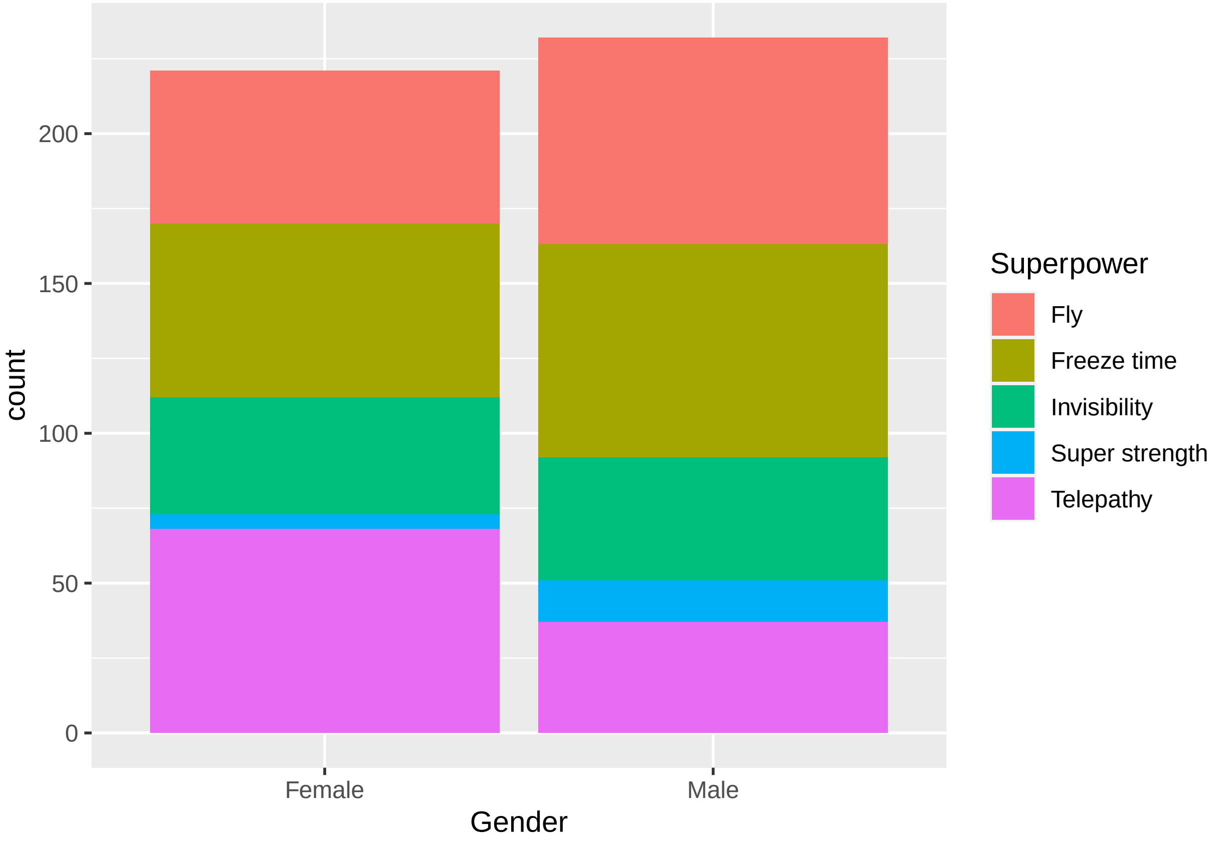
b. Use `infer` to perform a hypothesis test to see if the distribution of superpower preferences is different between males and females for high school seniors in Pennsylvania. Make a conclusion at the 0.01 level.

c. Which cells contribute the most to the chi-square statistic? For these cells, describe how the observed counts compare to the expected counts for male and female students.

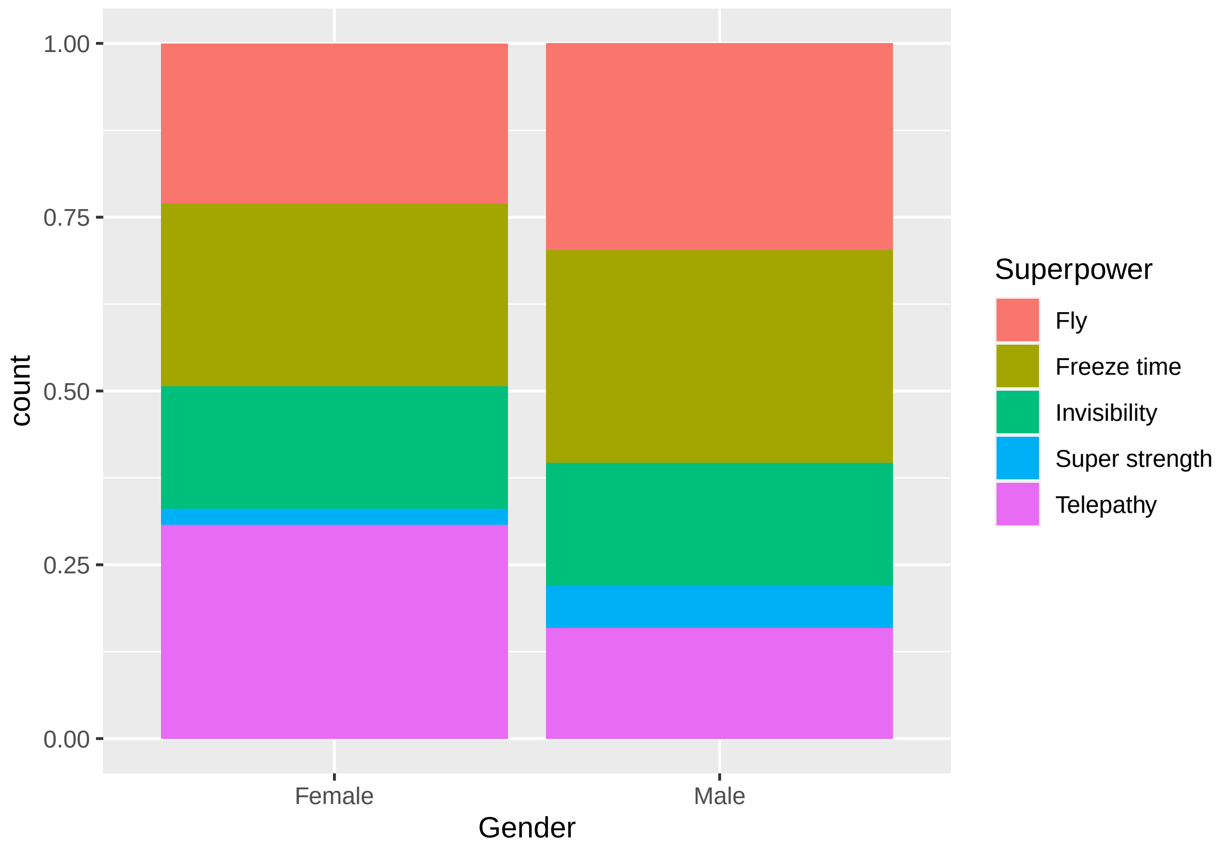
---

a.

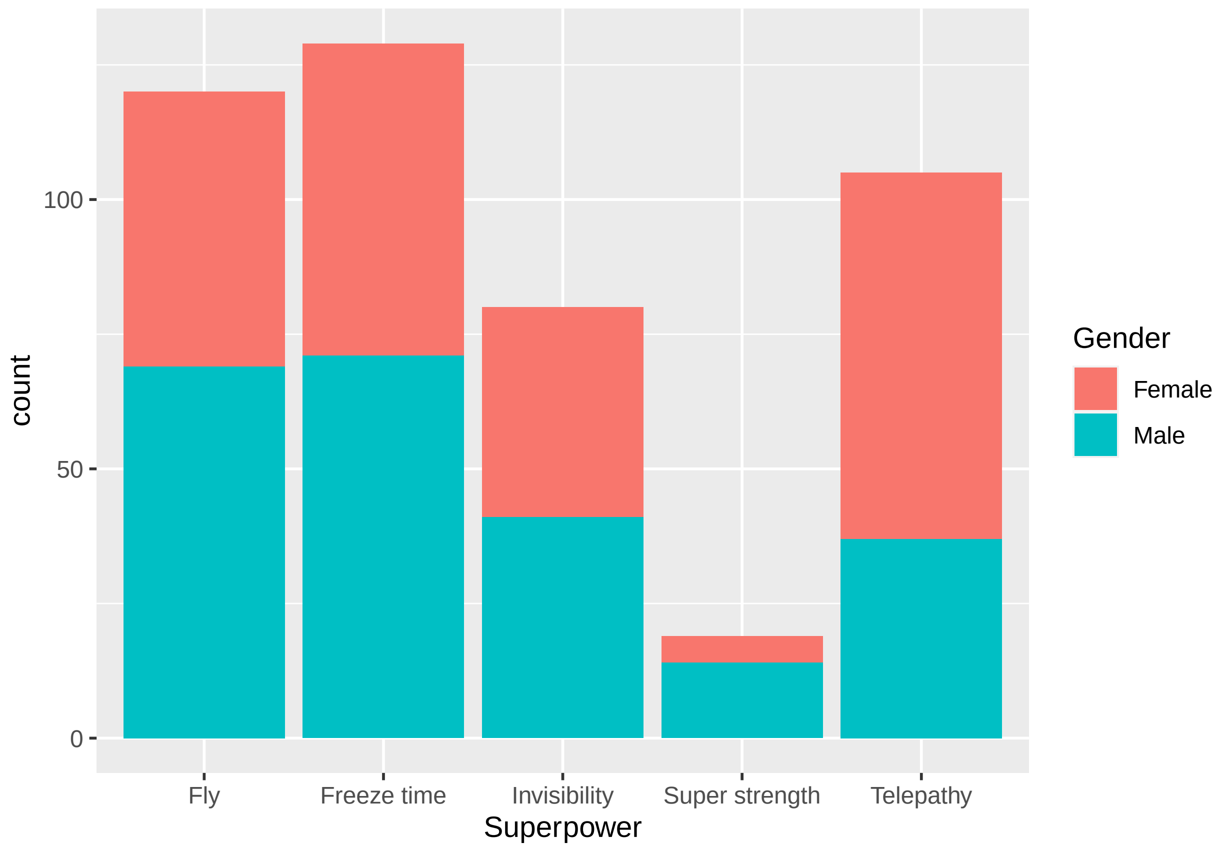
```
ggplot(PASeniors, aes(x = Gender, fill = Superpower))+geom_bar()
```



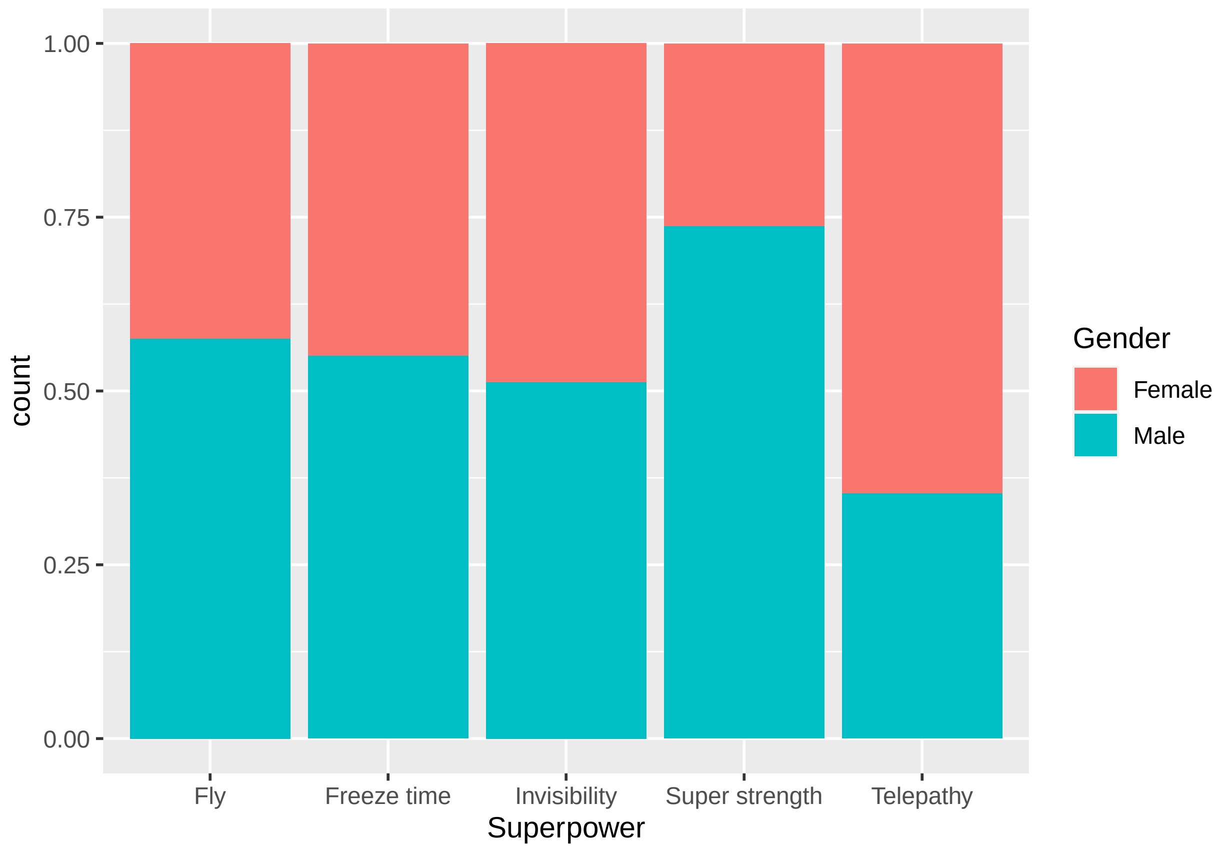
```
ggplot(PASeniors, aes(x = Gender, fill = Superpower))+geom_bar(position= "fill")
```



```
ggplot(PASeniors, aes(fill = Gender, x = Superpower))+geom_bar()
```



```
ggplot(PASeniors, aes(fill = Gender, x = Superpower))+geom_bar(position= "fill")
```



b. Using `infer` to created a null-distribution, the p-value of this sample is approximately 0.0016. This gives us sufficient evidence at the 0.01 level to reject the null hypothesis that distribution of superpower preferences is the same between males and females for high school seniors in Pennsylvania.

```
set.seed(3343)

chi_sq_stat <- PASeniors %>%
  specify(response = Superpower, explanatory = Gender) %>%
  hypothesise(null = "independence") %>%
  calculate(stat = "Chisq")

PASeniors %>%
  specify(response = Superpower, explanatory = Gender) %>%
  hypothesise(null = "independence") %>%
  generate(reps = 5000, type = "permute" ) %>%
  calculate(stat = "Chisq") %>%
  get_p_value(obs_stat = chi_sq_stat, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0016
```

c. Using the table shown in the statement of the problem, the telepathy and super strength cells seemed to have the greatest relative difference between genders.

Problem 7

The previous problem introduces a survey given to a sample of high school seniors in Pennsylvania. Two of the variables in the survey are `HangHours`, the number of hours per week spent hanging out with friends, and `SchoolPressure`, the amount of pressure felt due to schoolwork (None, Very little, Some, or A lot). We wish to test whether the amount of school pressure felt by students is related to the mean time hanging out with friends

a. What are the explanatory and response variables? Is each categorical or quantitative?

b. Create an appropriate data visualization showing the relationship between these variables.

c. Compute the mean and standard deviation in number of hours spent handing out with friends for each level of School Pressure. Additionally, compute the number of students in each level of School Pressure.

d. Use `infer` to compute the F statistic for this sample. What does the size the F statistic suggest about the relative variability between groups versus within groups?

e. Use `infer` to perform an appropriate test assessing whether the mean time spend hanging out with friends are equal among all groups. What is the conclusion of this investigation?

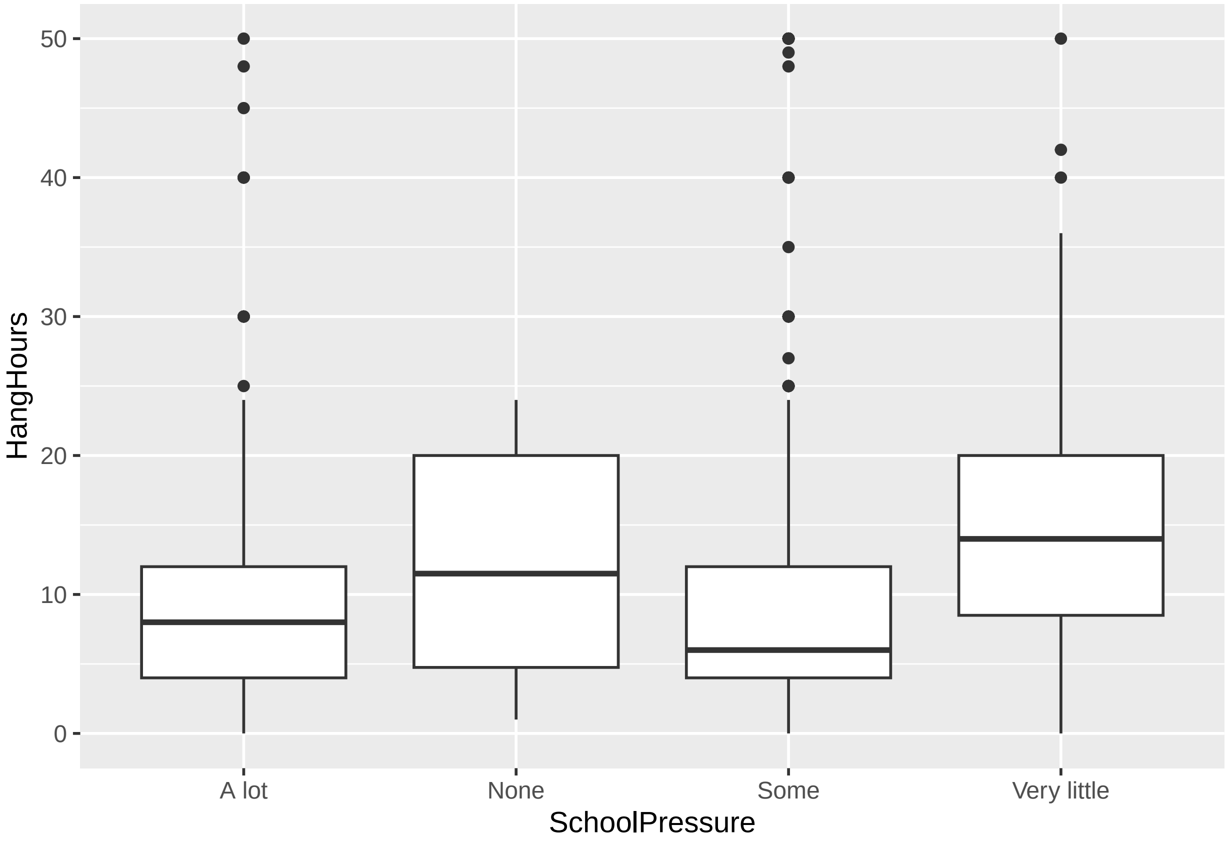
---

a. The explanatory variable is SchoolPressure (categorical), while the response variable is HangHours (quantitative).

b.

```
ggplot(PASeniors, aes(x = SchoolPressure, y = HangHours))+geom_boxplot()
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_boxplot()`).
```



c.

```
PASeniors %>% group_by(SchoolPressure) %>%
  drop_na(HangHours) %>%
  summarize(mean_hang = mean(HangHours), sd_hang = sd(HangHours), n = n())
```

```
## # A tibble: 4 x 4
##   SchoolPressure mean_hang sd_hang     n
##   <fct>              <dbl>   <dbl> <int>
## 1 A lot               9.89    8.94   178
## 2 None               12.2     8.50    16
## 3 Some               11.0    11.6    196
## 4 Very little        15.9    11.3     55
```

d. The F stat for this sample is 4.69, which suggests that the variability between groups is about 4.7 times larger than the variability within groups.

```
F_stat <- PASeniors %>%
  specify(HangHours ~ SchoolPressure) %>%
  hypothesise(null = "independence") %>%
  calculate(stat = "F")
```

```
## Dropping unused factor levels  from the supplied explanatory variable 'SchoolPressure'.
```

```
## Warning: Removed 8 rows containing missing values.
```

```
F_stat
```

```
## Response: HangHours (numeric)
## Explanatory: SchoolPressure (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  4.69
```

e. The p-value for the ANOVA test is 0.0044, which gives strong evidence (at the 0.01 level) that there is a difference in mean number of hours spent hanging out, among the different School Pressure levels.

```
set.seed(300)
PASeniors %>%
  specify(HangHours ~ SchoolPressure) %>%
  hypothesise(null = "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "F") %>%
  get_p_value(obs_stat = F_stat, direction = "right")
```

```
## Dropping unused factor levels  from the supplied explanatory variable 'SchoolPressure'.
```

```
## Warning: Removed 8 rows containing missing values.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0044
```

Problem 8

For each of the following questions, choose the type(s) of variable(s) that the question pertains to, the appropriate graphs that could be used to visualize the data, and the appropriate type of statistic that could be used to summarize the data.

a. What is the current public opinion on assault rifle bans? Do a majority support it?

b. Is there an association between the diameter of a plate used and the amount of food consumed?

c. Do people who take a daily multivitamin live longer than those who do not?

d. How many hours do Reed college students sleep each night?

e. Are men or women more likely to support restrictions on abortion?

f. Do college graduates who take at least 1 statistics course earn more 1 year after graduation than those who do not?

g. How far away are stars in the Milky Way galaxy?

h. What percentage of football field goal attempts from 35 yards or more are successful?

i. Is there a linear relationship between how long a child is breastfed and the child's weight at age 2?

j. Is there an association between the color of a car and whether that car has been pulled over for speeding?

---

a. **Variable**: Support ban on assault rifles (Binary Categorical) **Graphic**: Bar Chart. **Statistic**: Proportion

b. **Variable**: Diameter of plate (quantitative), food consumed (quantitative). **Graphic**: Scatterplot **Statistic**: Slope of regression equation

c. **Variable**: Lifespan (quantitative), consume multivitamin (Binary Categorical). **Graphic**: Side-by-side boxplot **Statistic**: Difference in means

d. **Variable**: Hours slept (Quantitative). **Graphic**: Histogram **Statistic**: Mean

e. **Variable**: Support abortion (Binary categorical), Gender (Binary categorical; based on the wording of this question) **Graphic**: Filled Barchart **Statistic**: Difference in proportions

f. **Variable**: Earnings after 1 year (quantitative), take statistics class (Binary Categorical). **Graphic**: Side-by-side boxplot **Statistic**: Difference in means

g. **Variable**: Distance to stars (quantitative) **Graphic**: Histogram **Statistic**: Mean

h. **Variable**: Success of field goal attempt (Binary categorical) **Graphic**: Bar chart **Statistic**: Proportion

i. **Variable**: Amount of time breastfed (quantitative), Weight of Child (quantitative) **Graphic**: Scatterplot **Statistic**: Correlation

j. **Variable**: Car color (multi-level categorical), pulled over for speeding (Binary Categorical). **Graphic**: Filled Barchart **Statistic**: Chi-Square statistic

---

Problem 9

In a study investigating how students use laptops in class, researchers recruited 45 student volunteers at one university. On average, students cycled through 65 active windows per lecture. The researchers developed a rubric to distinguish productive class-related applications from distracting ones, and recorded the proportion of distracting windows active, and the percent of time that was spent on distracting windows. They found, on average 62% of open windows were distracting, and that students had distracting windows open 42% of the time. Finally, the study measured how students performed on a test of related material, and concluded that students who spent more time on distracting websites generally had lower test scores.

a. What is the implied population of this study? What variables and parameters did the researchers want to study?
b. What was the sampling method?
c. Is this an experiment or observational study?
d. From the information given, explain why we cannot conclude that students who spend more time on distracting websites during class will have lower grades as a result.
e. Discuss how the investigation could be modified in order to make a causal conclusion about the entire population of students.

---

a. The implied population consists of students at universities. The researchers measure whether a student is using distracting applications, the amount of time spent on distracting windows, and the test score for a student on related material. They can estimate the *proportion* of students using distracting applications, the *mean* time each student spends on distracting windows, and the *mean* test score.

b. Since the researchers collected volunteers, the sampling method was voluntary / convenience.

c. The explanatory variables were not randomized among the students; therefore, this is an observational study, rather than an experiment.

d. Because this was an observational study, we cannot make causal conclusions based on our investigation. Perhaps the students who viewed distracting apps during class would have tended to have lower test scores regardless of whether they viewed the distracting apps; perhaps they did better on the test by viewing distracting apps than they would have if they had not been viewing distracting apps (for example, if these students would have elected not to come to class at all if they were not permitted to view apps, they may have performed even worse on the tests).

e. In order to facilitate causal conclusions, researchers would need to design a randomized experiment, where students are randomly assigned to either view distracting apps, or not view distracting apps, during class. It may be difficult to do so in a blinded fashion.

In order to generalize the results from the sample to the population, the researchers would need to collect a random sample, rather than a convenience sample.
