SDS 192 Introduction Data Sciences: Midterm II
Friday October 4th 5pm through Sunday October 6th 11:55pm, 2019

Question	Points	Points Deducted
1	$9 = 5 + 2 + 2$	
Total	34	

Instructions:

- This midterm lasts **140 minutes**, including transit time to the writing area. **Timestamps will be verified; exams whose timestamps indicate more than 140 minutes will be penalized.**
- There are 85 pages in this midterm. Please restaple the pages **in the correct order**.
- This exam is closed-book, to be individually completed and without the aid of the internet or mobile phones. You are allowed colored pens/pencils.
- In case of potential errors or ambiguity on the exam, please note them, state your assumptions, and use your best judgement.

Honor Code Statement: Smith College expects all students to be honest and committed to the principles of academic and intellectual integrity in their preparation and submission of course work and examinations. Students and faculty at Smith are part of an academic community defined by its commitment to scholarship, which depends on scrupulous and attentive acknowledgment of all sources of information, and honest and respectful use of college resources.

Dishonest Examination Behavior: The unauthorized giving or receiving of information during examinations or quizzes (this applies to all types, such as written, oral, lab or take-home) is dishonest examination behavior.

Signature: I have read the above instructions and agree to abide by the Honor Code in taking this exam. I will not speak with anyone about the exam until after the midterm period is over.

(Printed Name)

(Signature)

1 Short Answer

a) In terms of the "Grammar of Graphics" what is a statistical graphic?

b) What is the chief difference between a barplot and a histogram?

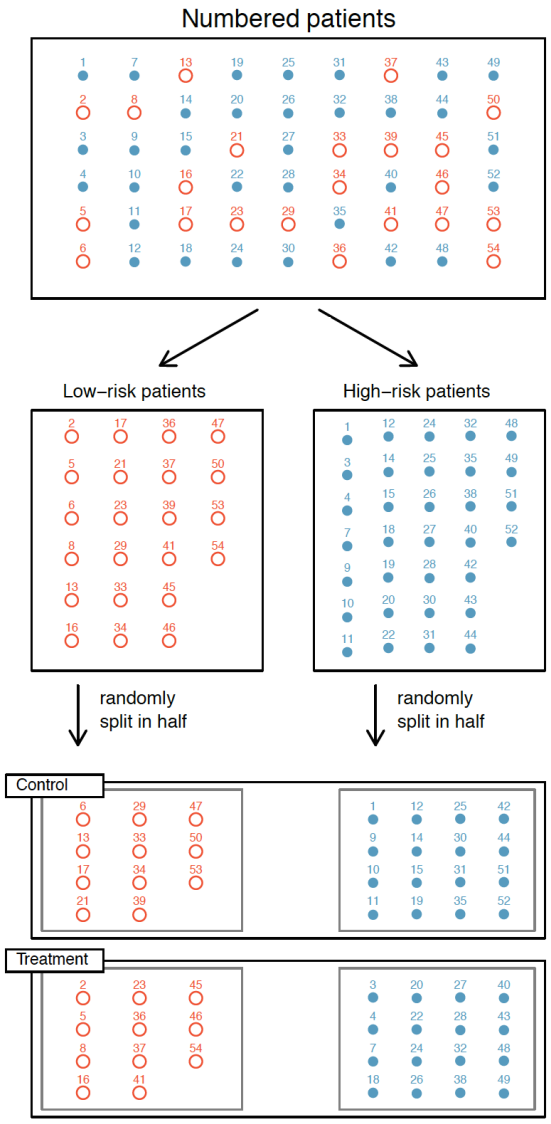
c) What is **overplotting**? Name the two ways seen in class to deal with it.

c) We saw in class that there are two ways to create a barplot using `ggplot`: either using `geom_bar()` or using `geom_col()`. Illustrate the need for these two different ways using simple data frame examples.

d) A smoother allows one to focus on a trend in graphic by separating out the A from the B. What are:
A: _____

B: _____

a) Researchers are looking at the effect of a drug on heart attacks. They first split patients in the study into low-risk and high-risk groups, then randomly assign half the patients from each group to the control and the other half to the treatment, as shown in the figure below.



This is an example of: _____.

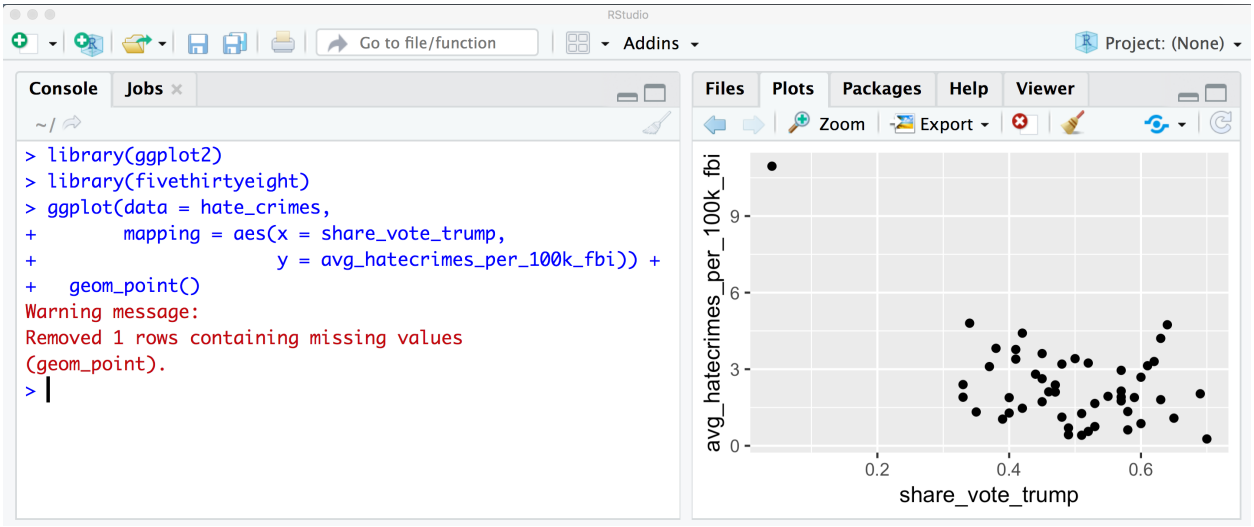
b) Alfred Kinsey, a sexologist in the 1940's, wants to study the sexual behavior of American males. After interviewing a number of males, he asks them to refer other men they know and conducts interviews with them. He repeats this process until 5500 men are interviewed. From an analysis of this data, he declares that "10% of all American males are exclusively homosexual." Comment on his research design specifically in two sentences or less, using language from this course.

a) Say a class of 30 introductory statistics students took a test where the median score was 17 and the interquartile range was 0. What are the 25th and 75th percentiles of test scores? Hint: A picture may help.

b) Recall in Problem Set 02 you created the following scatterplot visualizing the relationship between

- the proportion that voted Trump in the 2016 election
- the average annual hate crimes per 100,000 population between 2010-2015 as reported by the FBI

for the 50 states and the District of Columbia (DC). Here is what RStudio looks like after running the necessary code in the Console:



Other than by counting the number of points, based on the above output how can we know the number of points that are in above plot?

d) For which of the following pairs of variables would you visualize with a scatterplot? Circle which pairs.

- Pair 1: "Distance from school in miles" and "mode of transportation to school (bike, walking, bus)"
- Pair 2: "Number of years at a job" and "Salary"
- Pair 3: "Years experience playing an instrument" and "number of mistakes made playing a song"
- Pair 4: "Number of years since a person retired" and "favorite sport"

c) What is the chief difference between an experiment being blinded vs double-blinded?

d) An analysis of Middlebury faculty salaries shows that on average women get paid significantly less than men. However, an astute statistician observes that

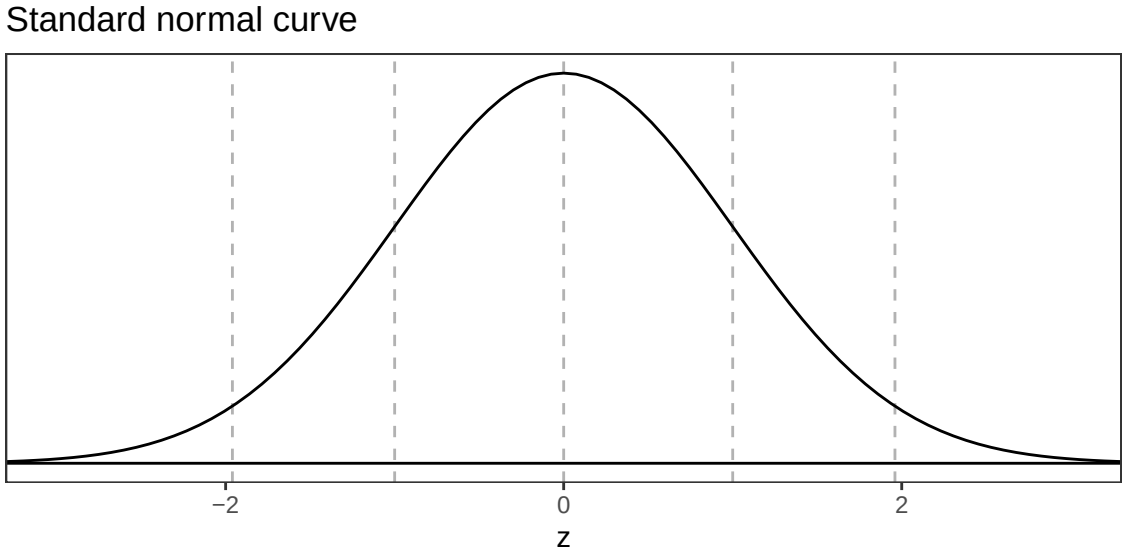
- Younger faculty tend to get paid less than older faculty due to seniority.
- Amongst the younger faculty, there is better representation of women because of shifts in the labor pool and hiring practices.

In this example, age is an example of what?

e) (Continued from part d) Thus an analysis of Middlebury faculty salaries should _____ for age.

c) The example from class where we studied the causal effect of shoes on the likelihood of waking up with a headache is an observational study. Why is it an observational study? **Answer in one sentence**.

d) Below we have a standard Normal Z-curve along with 5 vertical dashed lines at z = -1.96, -1, 0, 1, and 1.96 cutting the x-axis into 6 segments. In the plot below, write down the 6 proportion of values under the Z-curve in each of the 6 segments. Hint: Your 6 proportions should sum to 100%.



a) Analysis of Variance (ANOVA) compares k group means for the following hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
vs. $$H_A : \text{At least one of the } k \text{ means is different}$$

For example, in class we compared the mean life expectancy of countries in $k = 5$ continents. What other statistical technique covered in this course would allow us to similarly compare group means?

b) Say you are in a hypothesis testing situation and your estimate of the standard error is *underestimating* the true value of the standard error. Which of the following statements are true:

I: You will be more likely to incorrectly reject H_0.

II: You will be more likely to incorrectly fail to reject H_0.

III: It makes no difference.

c) Designed experiments, clinicial trials, and A/B tests are concerned with random X whereas polls and surveys are concerned with random Y. What are X and Y?

d) A *test statistic* is a X of the population parameter of interest used for hypothesis testing. What is X?

f) The *null distribution* used in hypothesis testing for computing p-values is the X distribution of the test statistic assuming Y. What are X and Y?
i) Name all 4 conditions for inference for regression.

j) Say you have data on the price of three kinds of fruit in a data frame `fruits`. Is this data in "tidy" format? If not, rewrite the data frame so that it is.

Date	Type	Price
2009-01-01	Apple	\$1.74
2009-01-02	Apple	\$1.73
2009-01-01	Orange	\$1.72
2009-01-02	Orange	\$1.74
2009-01-01	Melon	\$1.74
2009-01-02	Melon	\$1.75

a) For each of these five regression scenarios, name an appropriate visualization (along with any distinguishing features) that graphically summarizes the relationship between the outcome variable y and the explanatory/predictor variable(s).

1. Simple linear regression with one numerical predictor
2. Simple linear regression with one categorical predictor
3. Multiple regression with two numerical predictors
4. Multiple regression with one categorical and one numerical predictor

5. Multiple regression with two categorical predictors

b) Consider the following hypothetical study. Say you collect two variables of information from a population of interest: y = life expectancy and x = annual income out of college measured in units of thousands of dollars. You find that

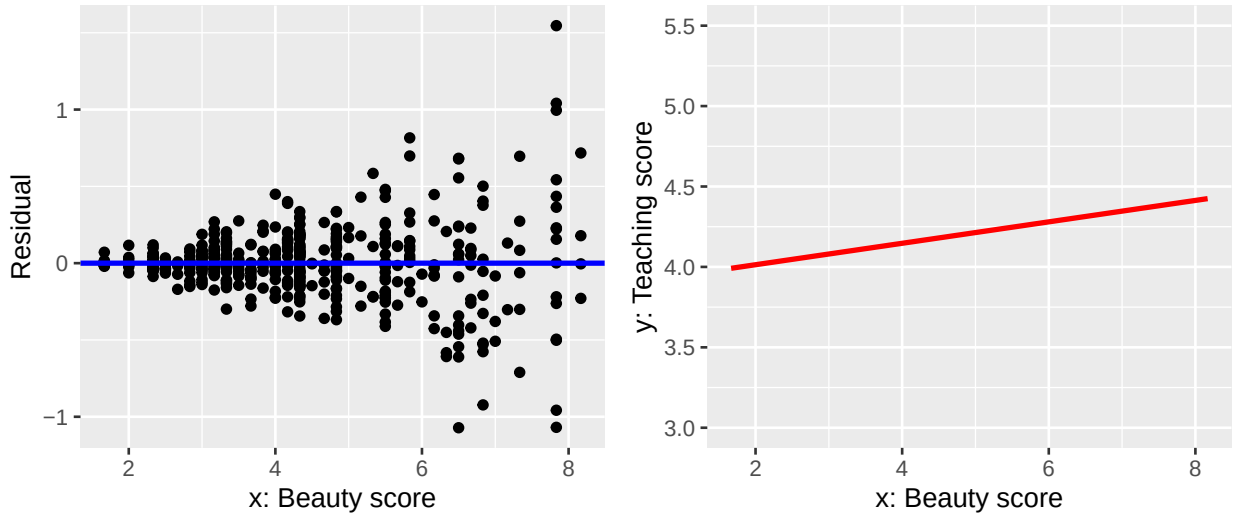
1. the correlation coefficient is 0.25
2. the fitted regression line $\widehat{y} = 45 + 0.5x$

Write down what the following two quantities would be if x was not measured in units of thousands of dollars, but measured in units of dollars:

1. the correlation coefficient
2. the fitted slope b_1 of the regression line $\widehat{y} = b_0 + b_1 x$

c) Name one situation when doing data analysis/modeling where log 10-transformations are useful? Answer in **20 words** or less.

d) Say we perform a regression to model an instructors' teaching score as a function of their beauty score, and obtain the following residual plot on the left which exhibits heteroskedasticity. Draw a *rough* sketch of what the scatterplot of x and y would look like given that the red line is the fitted regression line.



e) Say we perform a residual analysis of a regression model and find that the residuals exhibit very strong heteroskedasticity as above. What implications does this have for the results of our analysis?

d) Name two different reasons (one for each type of study we've seen in class) why it's difficult to establish the causal effect of college on future earnings.

f) For each scenario, determine the random sampling method used by the managers at a large company who wished to know the percentage of employees who feel "extremely satisfied" to work there:

1. Use the company email directory to contact 150 employees from among those employed for less than 5 years, 150 from among those employed for 5–10 years, and 150 from among those employed for more than 10 years.
2. Use the company email directory to contact every 50th employee on the list.
3. Select several divisions of the company at random. Within each division, draw a simple random sample of employees to contact.

2 Five Named Graphs

TABLE 2.4: Summary of Five Named Graphs

	Named graph	Shows	Geometric object	Notes
1	Scatterplot	A	`geom_point()`	
2	Linegraph	B	`geom_line()`	Used when there is a H
3	Histogram	C	`geom_histogram()`	Facetted histograms show I
4	Boxplot	D	`geom_boxplot()`	
5	Barplot	E	`geom_` F when counts are not pre-counted, `geom_` G when counts are pre-counted	Stacked, side-by-side, and faceted barplots show the joint distribution of 2 J

What are:

3 Barplots

Say you have a data frame called `example` that has 3 variables, 4 rows, and a header row.

fruit	city	number
apple	Toronto	5
apple	Montreal	7
orange	Toronto	4
orange	Montreal	3

Draw what the visualization resulting from the following code looks like:

```
ggplot(data = example, mapping = aes(x = fruit)) +
  geom_bar() +
  labs(x = "Fruit type", y = "Number")
```

Draw what the visualization resulting from the following code looks like:

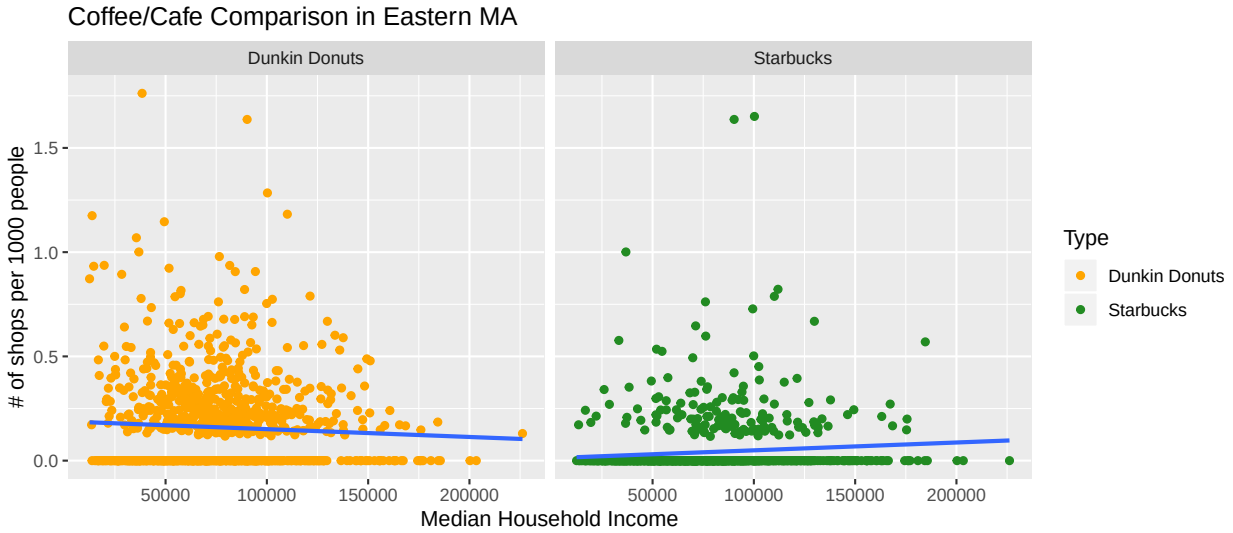
```
ggplot(data = example, mapping = aes(x = fruit, y = number)) +
  geom_col() +
  labs(x = "Fruit type", y = "Number")
```

Draw what the visualization resulting from the following code looks like:

```
ggplot(data = example, mapping = aes(x = fruit, y = number, fill = city)) +
  geom_col(position = "dodge") +
  labs(x = "Fruit type", y = "Number", fill = "City")
```

4 America Runs on Starbucks?

A researcher from eastern Massachusetts is a big Starbucks fan. She has a suspicion that Starbucks tend to locate in richer neighborhoods, while this is not the case for Dunkin Donuts. She writes code to pull data from the internet about all 1024 census tracts (areas where decennial census data are collected) in Eastern Massachusetts. She summarizes her results in the following graphic:



a) Write out the elements of the "Grammar of Graphics" that need to be specified to create this graphic. You do not need to specify the axes labels, the plot title, nor the regression lines

c) Assuming there are no missing data, how many rows are in the data frame that we input into the `ggplot()` function?

d) Does the presented visualization support or contradict the researcher's suspicion? Why?

5 Boxplots

The asking prices for a sample of 15 textbooks currently being sold are listed below. For convenience, the data have been ordered:

-11	12	15	20	20	30	30	30	30	40	40	40	40	41	51

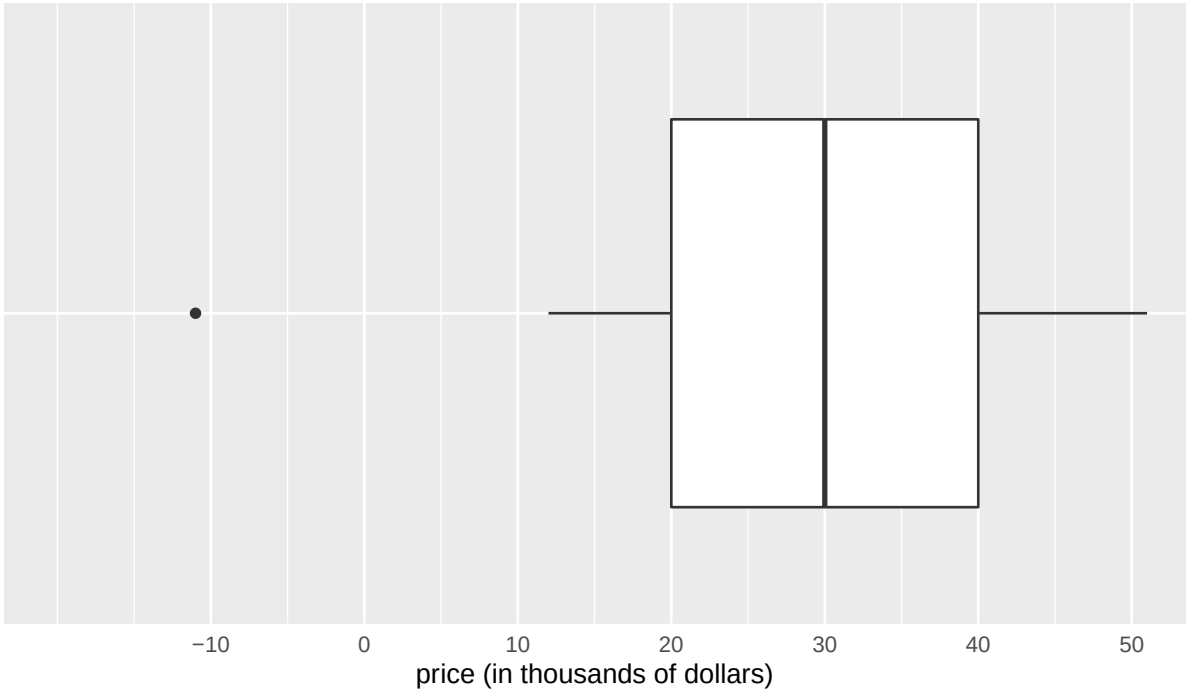
Furthermore, the following three *summary statistics* have been computed:

1st quartile	Average	3rd quartile
20	29	40

a) What is the interquartile range (IQR) for this data?

b) Is the IQR a measure of center or a measure of spread of a numerical variable? Circle your response.

c) Draw the boxplot for this dataset. Be sure to mark all relevant numerical values:



6 Car Prices

The asking prices (in thousands of dollars) for a sample of 15 cars currently being sold are listed below. For convenience, the data have been ordered:

11	12	20	25	25	30	30	30	30	35	35	35	35	39	39

Furthermore, the following three *summary statistics* have been computed:

1st quartile	2nd quartile	3rd quartile
25	30	35

a) What is the interquartile range (IQR) for this data?

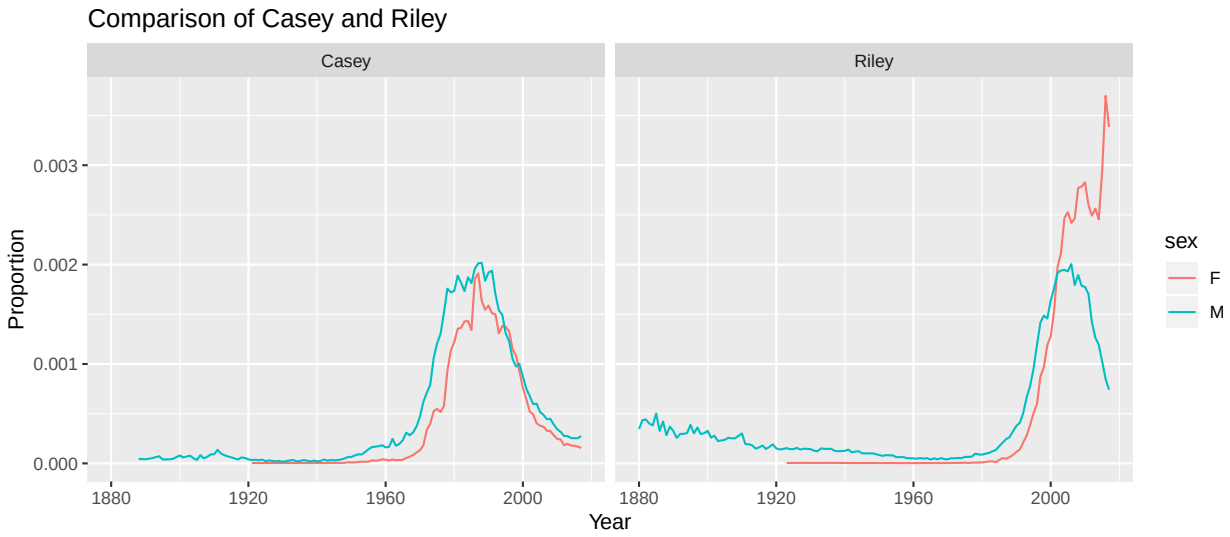
b) Is the IQR a measure of center or a measure of spread of a numerical variable? Circle your response.

c) Draw the boxplot for this dataset:

7 Babynames

The `babynames` dataset contains all babynames used more than 5 times for any given year, split by sex (as recorded by the Social Security), for the years 1880 through 2015. Here is a preview of the first 10 rows:

year	sex	name	n	prop
1880	M	Riley	41	0
1881	M	Riley	47	0
1882	M	Riley	54	0
1883	M	Riley	45	0
1884	M	Riley	47	0
1885	M	Riley	58	0
1886	M	Riley	39	0
1887	M	Riley	46	0
1888	M	Riley	37	0
1888	M	Casey	6	0



Furthermore, the following three *summary statistics* have been computed:

1st quartile	2nd quartile	3rd quartile
25	30	35

a) What is the interquartile range (IQR) for this data?

b) Is the IQR a measure of center or a measure of spread of a numerical variable? Circle your response.

c) Draw the boxplot for this dataset:

8 Family income by city

NPR recently posted an article titled "How Much (Or Little) The Middle Class Makes, In 30 U.S. Cities." It included the image on the following page.

a) This image most closely resembles what statistical visualization we've seen?

b) Which city has the third highest mean family income?

c) Which four cities have the highest income disparity in the US?

d) Quantify this income disparity for only one of the four chosen cities in part c) using a summary statistic of your choice.

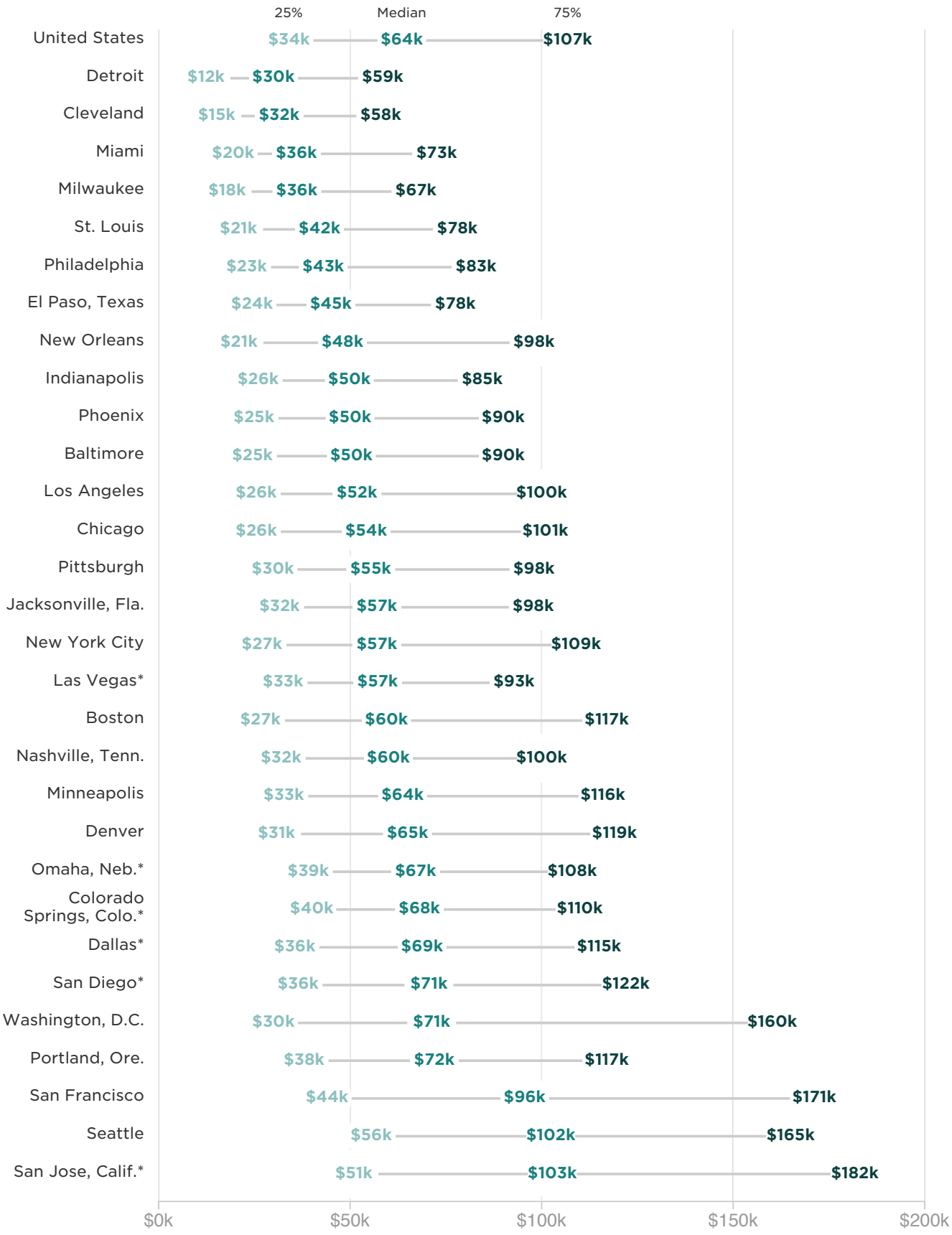
e) What proportion of Nashville families had a family income of $100K or more?

f) What proportion of Nashville families had a family income of $80K or more?

WRITE YOUR RESPONSES BELOW:

What Is Middle Class?

Family income by city, 2013



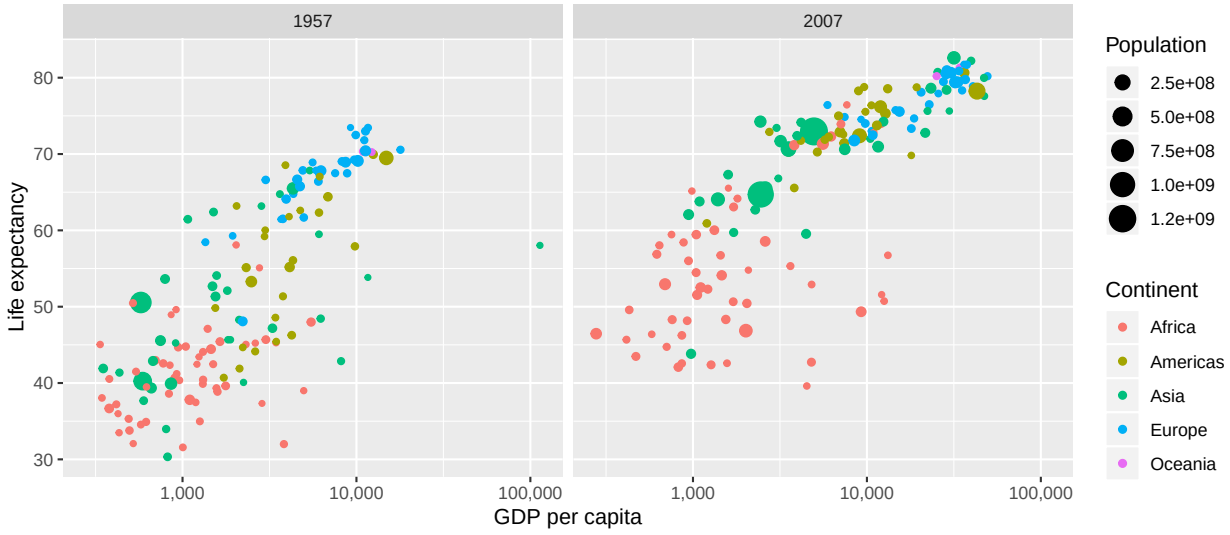
9 Gapminder

Consider a subset of the `gapminder` dataset we've seen numerous times in class:

```
library(gapminder)
gapminder_subset <- gapminder %>%
  filter(year %in% c(1957, 2007))
gapminder_subset
```

```
## # A tibble: 284 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1957    30.3  9240934      821.
##  2 Afghanistan Asia       2007    43.8 31889923      975.
##  3 Albania     Europe     1957    59.3  1476505     1942.
##  4 Albania     Europe     2007    76.4  3600523     5937.
##  5 Algeria     Africa     1957    45.7 10270856     3014.
##  6 Algeria     Africa     2007    72.3 33333216     6223.
##  7 Angola      Africa     1957    32.0  4561361     3828.
##  8 Angola      Africa     2007    42.7 12420476     4797.
##  9 Argentina   Americas   1957    64.4 19610538     6857.
## 10 Argentina   Americas   2007    75.3 40301927    12779.
## # ... with 274 more rows
```

Using this data, we can create the following plot:



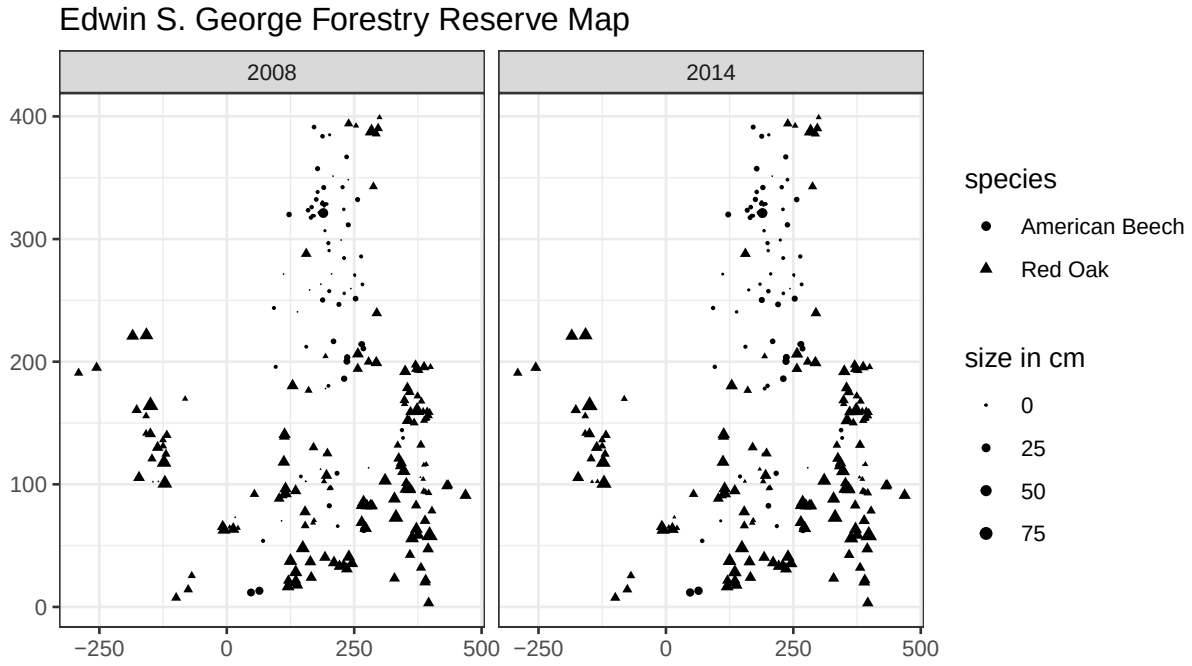
a) Write out **in bullet point form** all the elements of the "Grammar of Graphics" that need to be specified in a `ggplot()` function call to create this graphic. Note

- You don't need to write code, you only need to specify all components of the graphic.
- There is no need to specify the x and y axes labels.

b) Why does the x-axis increase on a multiplicative scale (1000, 10000, 100000) instead of an additive scale (Ex: 1000, 2000, 3000)?

10 Maps

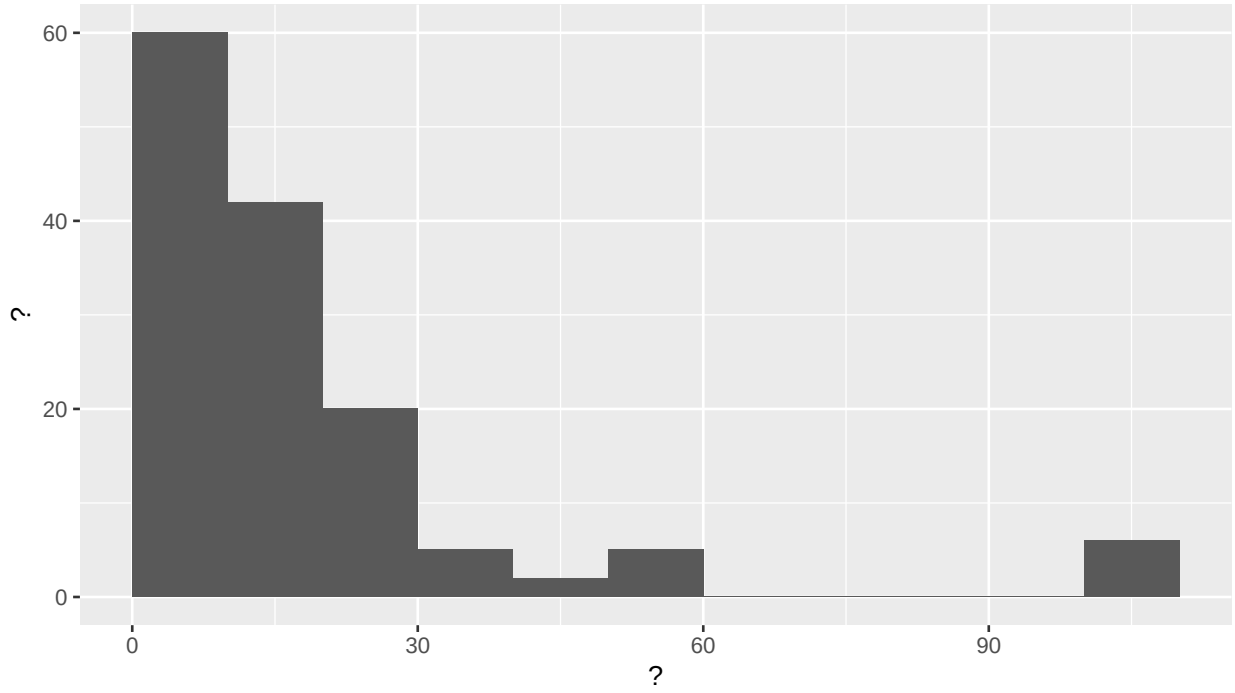
Say you are walking around the Forestry Department at the University of Michigan and you see the following visualization posted on a bulletin board.



a) The geometric object of this visualization is "points." Map all the aesthetic attributes of this visualization to variables of a hypothetical dataset as is required by the "grammar of graphics" and identify all other additional components that you need to specify to create this visualization. You may treat the title of the visualization and the legends as given.

11 Histograms

In a statistics class with 140 students, the professor records how much money (in dollars) each student has in their possession during the first class of the semester. The histogram shown below represents the data they collected:



a) What is the variable on the horizontal (x) axis?

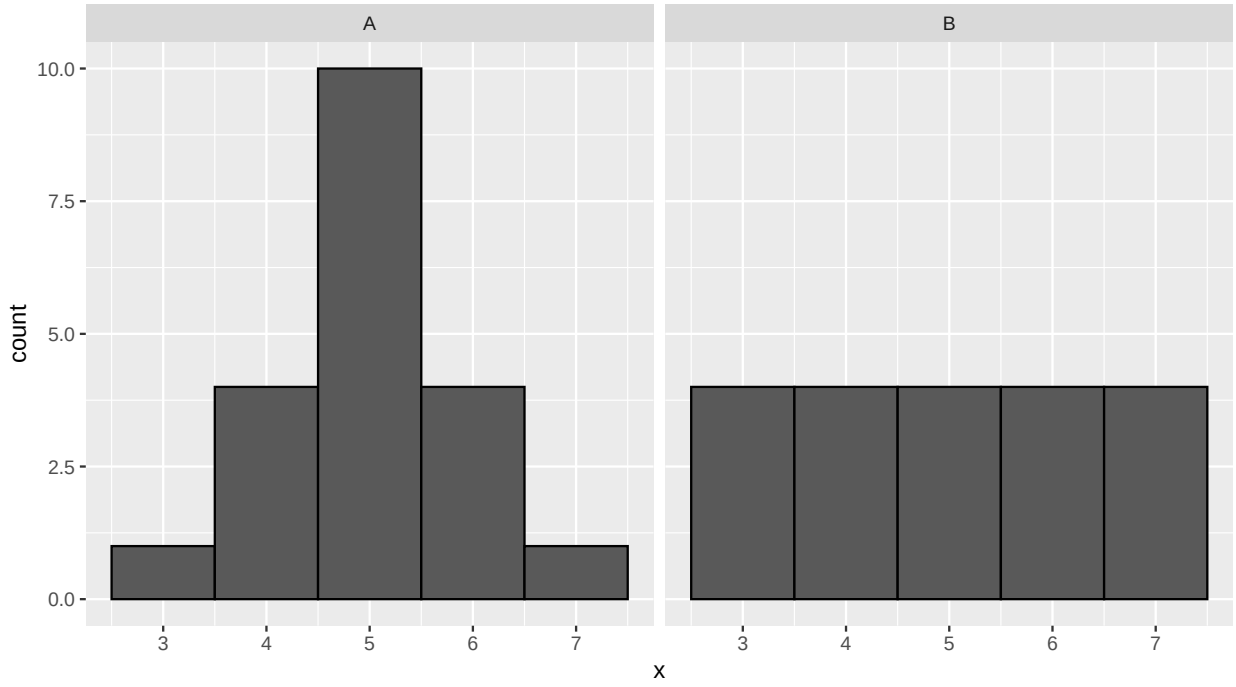
b) What is the quantity on the vertical (y) axis?

c) The height of the second bar is 42. What does that tell us? Say precisely in **one sentence**.

d) Fill in the blanks: The median amount of money possessed is between \$ _____ and \$ _____. Show work or briefly explain your reasoning.

12 Histograms

a) Which of these two histograms exhibits more variability in the variable `x`?



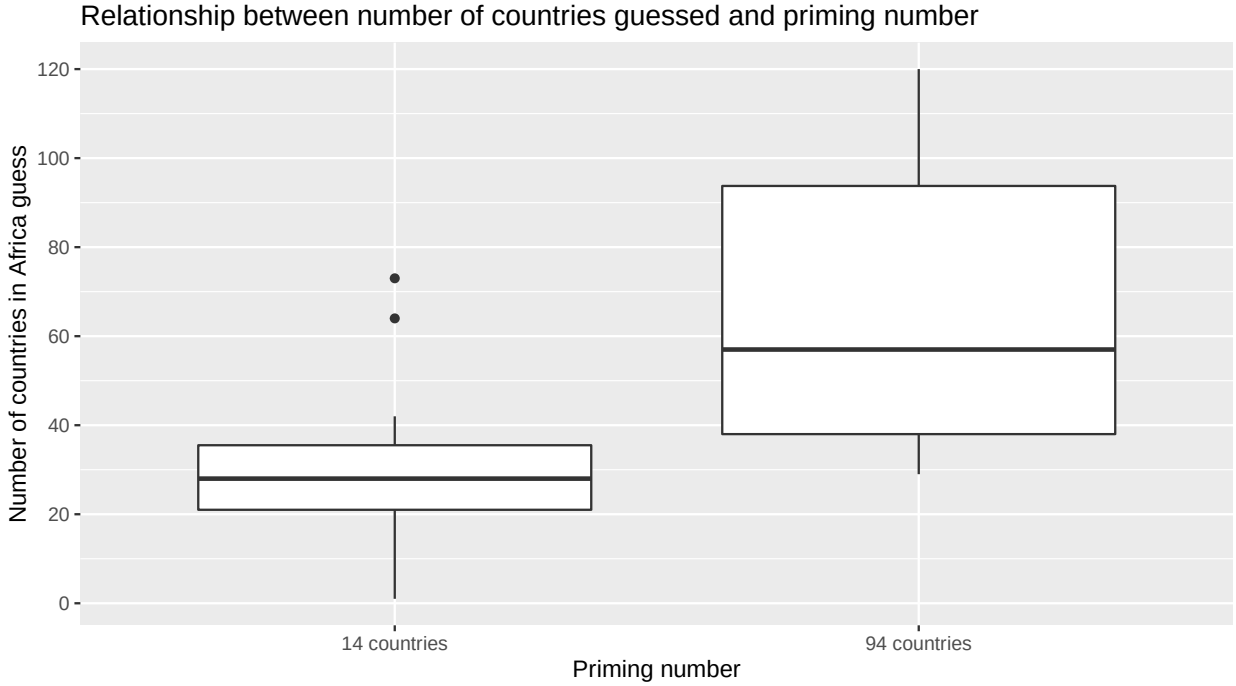
b) Write the rough pseudo-code of the `ggplot()` command that created the above plot, in particular specifying all elements of the "Grammar of Graphics." If you feel confident writing the code directly, then feel free to do so.

13 Exploratory data analysis via visualizations

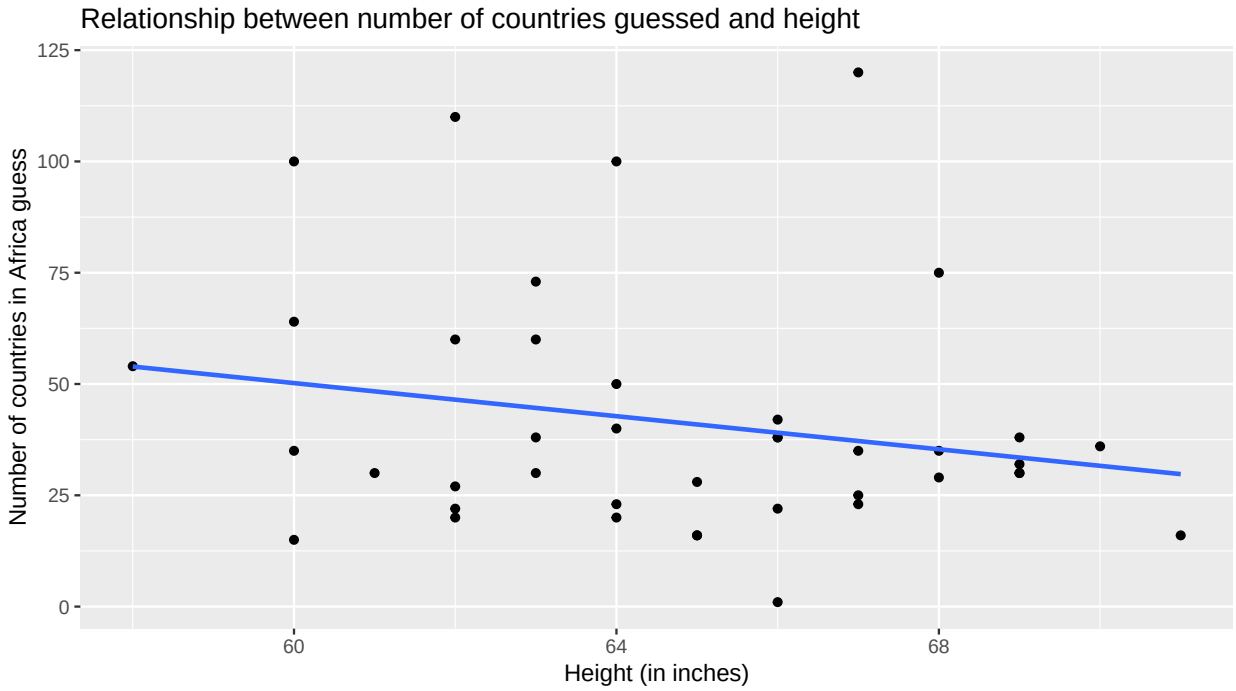
Continuing the previous `africa` question, for the remainder of this midterm let the outcome variable y be the number of countries a student guesses.

a) Name an ideal exploratory data visualization for the relationship between y and `height`.

b) We present an exploratory boxplot of the relationship between y and `priming`. It is a fact that there is more variation in responses amongst the students primed with the number 94. How is this apparent in the visualization? Compute the approximate values of a *summary statistic* we've seen in class to justify your answer.



c) The following graphic is created by the (incomplete) code snippet below.




```
ggplot(africa, aes(AAA, BBB)) +
  geom_CCC() +
  geom_DDD(method = "lm", se = FALSE) +
  labs(x = "Height (in inches)", y = "Number of countries in Africa guessed")
```

What precise code should be in place of `AAA`, `BBB`, `CCC`, and `DDD` in order to create this plot?

d) While an exploratory scatterplot of the relationship between y and `year` would be valid since `year` is numerical, why would a (vertical) boxplot with `year` on the x-axis also be acceptable *for this particular dataset*? Answer in one sentence.

14 Babynames

Recall the `babynames` dataset that contains all babynames used more than 5 times for any given year, split by sex, for the years 1880 through 2015. Here is a preview of the first 10 rows:

year	sex	name	n	prop
1880	F	Mary	7065	0.07
1880	F	Anna	2604	0.03
1880	F	Emma	2003	0.02
1880	F	Elizabeth	1939	0.02
1880	F	Minnie	1746	0.02
1880	F	Margaret	1578	0.02
1880	F	Ida	1472	0.02
1880	F	Alice	1414	0.01
1880	F	Bertha	1320	0.01
1880	F	Sarah	1288	0.01

START WRITING YOUR RESPONSES WHERE INDICATED BELOW.

a) Write the pseudocode that is going to compute the total number of babies born between 1950 and 2000 that are named "Riley."

b) You want to compare the degree to which the names "Casey" and "Riley" have been "unisex" names for all years between 1950 to 2000, in other words the focus is on the degree to which the names have been used by both sexes. Write the pseudocode for the data wrangling and specify all the elements of the grammar of graphics that is going to generate an appropriate visualization. Hint: Draw the visualization first.

15 Weather Data

a) The `weather` data set in the `nycflights13` package contains hourly meterological data for the three NYC airports (EWR, JFK, and LGA) for every day in 2013. We present a snapshot of the data below, but only for the first 6 rows in the data set. What variables are needed to uniquely identify each observation?

origin	year	month	day	hour	temp	humid	wind_speed	precip	pressure	visib
EWR	2013	1	1	1	39	59	10.4	0	1012	10
EWR	2013	1	1	2	39	62	8.1	0	1012	10
EWR	2013	1	1	3	39	64	11.5	0	1012	10
EWR	2013	1	1	4	40	62	12.7	0	1012	10
EWR	2013	1	1	5	39	64	12.7	0	1012	10
EWR	2013	1	1	6	38	67	11.5	0	1012	10

b) Write down the arithmetic operation you would enter into a calculator to compute the number rows that the `weather` data set has (not including the header row). An example of an arithmetic operation is $10 \times 7 + 6$.

16 "Tidy" Data

Say your collaborator in biology sends you an Excel spreadsheet with the contents below. This data is **not** in tidy format. Rewrite this spreadsheet data so that it is in tidy format. To minimize writing you may use abbreviations. For example in place of "Allactaga balikunica" write "Al.Ba."

species	tail length
Allactaga balikunica	**177.32**
Allactaga bullata	165.2
Allocricetulus eversmanni	**18.64**
Apodemus uralensis	84.89
Arvicola amphibius	105.14
bold = needs checking	
yellow = from different source	

17 Alcohol consumption globally

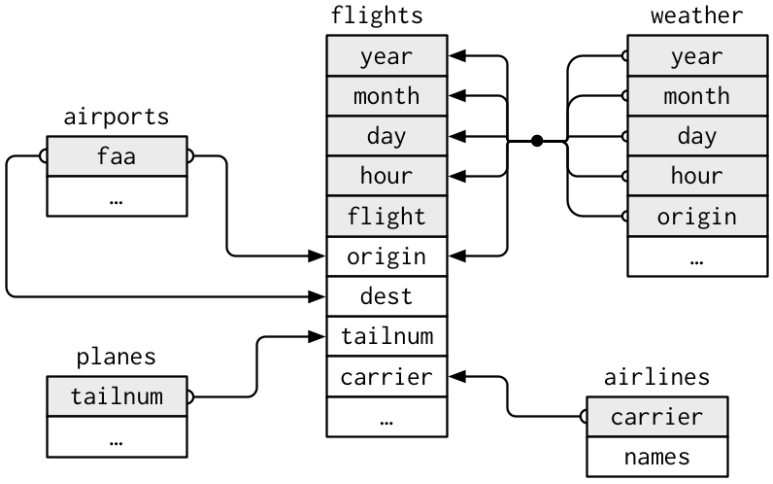
a) The following table of presents average annual alcohol consumption per individual for two countries. The table is not in "tidy" format. Rewrite it so that it is.

country	beer_servings	spirit_servings	wine_servings
Afghanistan	0	0	0
Albania	89	132	54

b) The above table only has data for the first two alphabetically listed countries. However the complete dataset has information for all 193 countries. A social scientist wants a visualization that will allow the reader to compare beer vs spirit vs wine consumption worldwide. Draw a rough sketch of an effective graphic to display this. There is no need to write the code/pseudocode that will construct this graphic, just sketch the graphic.

18 NYC flights

Recall the `airports`, `planes`, `flights`, `weather`, and `airlines` datasets in the `nycflights13` package containing information about all 336,776 domestic flights that left one of three NYC airports (EWR, JFK, and LGA) in 2013. Furthermore, recall the following graphic showing how theses datasets are related.



Also, consider the following R output given the names of the variables/ columsn in each dataset.

```
library(nycflights13)
names(airports)

## [1] "faa"   "name"  "lat"   "lon"   "alt"   "tz"    "dst"   "tzone"

names(planes)

## [1] "tailnum"      "year"         "type"         "manufacturer"
## [5] "model"        "engines"      "seats"        "speed"
## [9] "engine"

names(flights)

##  [1] "year"           "month"          "day"            "dep_time"
##  [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
##  [9] "arr_delay"      "carrier"        "flight"         "tailnum"
## [13] "origin"         "dest"           "air_time"       "distance"
## [17] "hour"           "minute"         "time_hour"

names(weather)

##  [1] "origin"     "year"       "month"      "day"        "hour"
##  [6] "temp"       "dewp"       "humid"      "wind_dir"   "wind_speed"
## [11] "wind_gust"  "precip"     "pressure"   "visib"      "time_hour"

names(airlines)

## [1] "carrier" "name"
```

a) Which datasets are you going to need to compute the available seat miles (sum over all flights of the number of seats $\times$ the number of miles flown) for United Airlines in 2013?

b) Write down the pseudocode that is going to merge the `flights` and `weather` datasets so that on top of information for all 336,776 flights, we have information of the weather conditions at the time of each flight's departure.

c) Write the pseudocode that will output a table displaying the median departure delay for each airline leaving Newark (airport code `EWR`).

19 Titanic

You are presented with data on the Titanic disaster of 1912 in a data frame `Titanic`, which cross-classifies survival vs death by class, sex, and age. Write down the *pseudocode* of the commands that will output a table comparing survival vs death counts for the following three scenarios:

a) by sex

b) by sex and class and age

c) to answer the question if the "women and children"-first policy of the White Star Line Company (the company that ran the Titanic) held true or not.

Note: you don't need to calculate the output table, just write the pseudocode that would produce it where the more concise the pseudocode the better. Here is what the `Titanic` data looks like:

Class	Sex	Age	Survived	n
1st	Male	Child	No	0
2nd	Male	Child	No	0
3rd	Male	Child	No	35
Crew	Male	Child	No	0
1st	Female	Child	No	0
2nd	Female	Child	No	0
3rd	Female	Child	No	17
Crew	Female	Child	No	0
1st	Male	Adult	No	118
2nd	Male	Adult	No	154
3rd	Male	Adult	No	387
Crew	Male	Adult	No	670
1st	Female	Adult	No	4
2nd	Female	Adult	No	13
3rd	Female	Adult	No	89
Crew	Female	Adult	No	3
1st	Male	Child	Yes	5
2nd	Male	Child	Yes	11
3rd	Male	Child	Yes	13
Crew	Male	Child	Yes	0
1st	Female	Child	Yes	1
2nd	Female	Child	Yes	13
3rd	Female	Child	Yes	14
Crew	Female	Child	Yes	0
1st	Male	Adult	Yes	57
2nd	Male	Adult	Yes	14
3rd	Male	Adult	Yes	75
Crew	Male	Adult	Yes	192
1st	Female	Adult	Yes	140
2nd	Female	Adult	Yes	80
3rd	Female	Adult	Yes	76
Crew	Female	Adult	Yes	20

20 Unisex Names

Write the pseudocode that is going to generate an appropriate visualization to compare the trends in the "unisex"iness (not a measure of gender ambiguous sexiness, but rather the degree to which a name is used by both sexes) of the names "Casey" and "Riley" from 1950 to 2014. As a hint, here are the first 10 rows of the `babynames` dataset.

year	sex	name	n	prop
1880	F	Mary	7065	0.07
1880	F	Anna	2604	0.03
1880	F	Emma	2003	0.02
1880	F	Elizabeth	1939	0.02
1880	F	Minnie	1746	0.02
1880	F	Margaret	1578	0.02
1880	F	Ida	1472	0.02
1880	F	Alice	1414	0.01
1880	F	Bertha	1320	0.01
1880	F	Sarah	1288	0.01

21 Exploratory data analysis via data wrangling

Recall the Google Forms survey you completed where:

- Students with an odd birthday (Ex: Nov 15th) were first asked if there are more or less than **14 countries** in Africa and then asked to guess how many countries there are in Africa.
- Students with an even birthday (Ex: Nov 14th) were first asked if there are more or less than **94 countries** in Africa and then asked to guess how many countries there are in Africa.

Let's refer to the numbers 14 and 94 as "priming" numbers since survey participants were "primed" with them in order to influence the number of countries they guessed. Furthermore all students were also asked their height (in inches), their graduation year (2019, 2020, 2021, or 2022), and whether or not they had previously been to Africa. A total of 41 students responded and the results are saved in a data frame `africa` with 41 rows:

```
## # A tibble: 41 x 5
##     year height been_to_africa priming      countries
##    <dbl>  <dbl> <chr>          <chr>            <dbl>
##  1  2021     70 No             14 countries        36
##  2  2020     67 No             94 countries       120
##  3  2021     69 No             14 countries        30
##  4  2021     60 Yes            14 countries        64
##  5  2021     66 No             14 countries         1
##  6  2021     66 No             14 countries        22
##  7  2022     65 No             14 countries        16
##  8  2021     64 No             94 countries       100
##  9  2022     68 No             94 countries        29
## 10  2021     62 No             94 countries       110
## # ... with 31 more rows
```

a) Write the pseudocode that will allow you to wrangle `africa` to obtain the median number of countries guessed for each of the two priming groups:

```
## # A tibble: 2 x 2
##   priming      median_guess
##   <chr>               <dbl>
## 1 14 countries           28
## 2 94 countries           57
```

b) Write the pseudocode that will allow you to wrangle `africa` to obtain only the year, priming group, and number of countries guessed for only the first-year students (class of 2022):

```
## # A tibble: 4 x 3
##    year priming      countries
##   <dbl> <chr>            <dbl>
## 1  2022 14 countries        16
## 2  2022 94 countries        29
## 3  2022 14 countries        30
## 4  2022 14 countries        27
```

c) Write the pseudocode that will allow you to wrangle `africa` so that the rows are reordered from the largest number of countries guessed to the smallest (note we only show the first 5 out of 41 rows in the output below):

```
## # A tibble: 5 x 5
##    year height been_to_africa priming      countries
##   <dbl>  <dbl> <chr>          <chr>            <dbl>
## 1  2020     67 No             94 countries       120
## 2  2021     62 No             94 countries       110
## 3  2021     64 No             94 countries       100
## 4  2021     60 No             94 countries       100
## 5  2019     68 No             94 countries        75
```

22 Regression with a numerical explanatory variable

Recall our teaching evaluation dataset seen in class. We're interested in fitting a model of teaching score (evaluated by students) as a function of instructor age.

```
ggplot(data = evals, mapping = aes(x = age, y = score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Instructor age", y = "Teaching score")
```


```
model_score <- lm(score ~ age, data = evals)
get_regression_table(model_score)

## # 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    4.46      0.127     35.2    0        4.21     4.71
## 2 age         -0.006     0.003     -2.31   0.021   -0.011   -0.001
```

a) Interpret the `intercept` term in the `estimate` column of the regression table.

b) Give the precise interpretation of the slope for `age` in the `estimate` column of the regression table.

c) The regression line visualized in the above figure is considered the "best fitting line" through these points. By what criteria do we mean "best"?

d) What is the correlation coefficient of `age` and `score`? Is it positive or negative?

e) Consider the first row of the following output. Write down the equation that computes the first value of `score_hat`: `4.25`.

```
model_points <- get_regression_points(model_score)
model_points

## # A tibble: 463 x 5
```

```
##       ID score   age score_hat residual
##    <int> <dbl> <int>     <dbl>    <dbl>
##  1     1   4.7    36      4.25    0.452
##  2     2   4.1    36      4.25   -0.148
##  3     3   3.9    36      4.25   -0.348
##  4     4   4.8    36      4.25    0.552
##  5     5   4.6    59      4.11    0.488
##  6     6   4.3    59      4.11    0.188
##  7     7   2.8    59      4.11   -1.31
##  8     8   4.1    51      4.16   -0.059
##  9     9   3.4    51      4.16   -0.759
## 10    10   4.5    40      4.22    0.276
## # ... with 453 more rows
```

f) Write down the equation that computes the first value of `residual`: `0.452`.

g) Write down the data wrangling pseudocode to apply to `model_points` to compute the value of the criteria described in part c).

23 Seattle House Prices

Recall the Seattle House Prices dataset you saw in the DataCamp course "Modeling with Data in the Tidyverse." where the sale price of 21,613 homes sold between May 2014 and May 2015 in King County WA is provided along with other information; let's only consider 3 of these variables: sale price, the square footage of living space, and the condition of the house (1 = worst, ..., 5 = best). Before we begin this question, let's perform a little data wrangling.

```
library(moderndive)
house_prices <- house_prices %>%
  mutate(
    log10_price = log10(price),
    log10_size = log10(sqft_living)
  ) %>%
  select(log10_price, log10_size, condition)
```

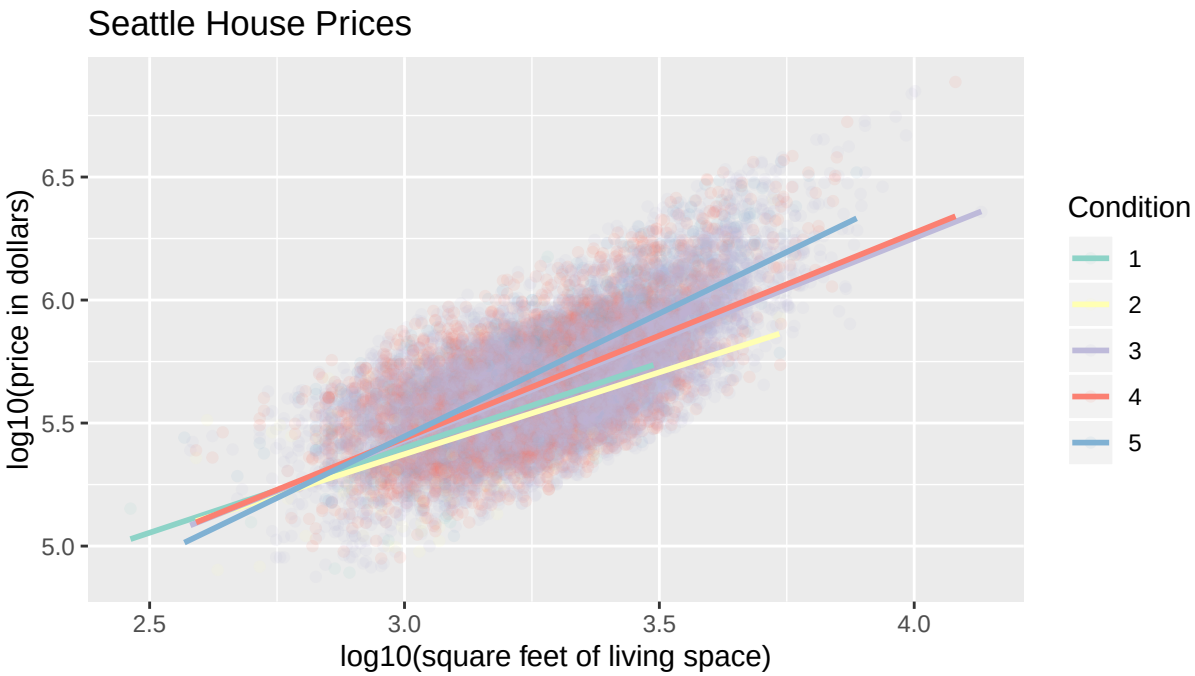
Now let's look at a random sample of 5 out of the 21,613 rows:

log10_price	log10_size	condition
5.3	3.1	3
5.8	3.3	5
5.6	3.3	3
5.8	3.3	4
5.5	3.1	3

We are interested in modeling the outcome variable y = log10 of house price in dollars as a function of two explanatory variables:

1. x_1: numerical explanatory/predictor variable log10 of the square footage of the house
2. x_2: categorical explanatory/predictor variable condition

You fit an interaction model, both graphically and using a regression model. Note the last 4 rows of the regression table got truncated; they should read `log10_size:condition2` through `log10_size:condition5`.




```
house_price_model <- lm(log10_price ~ log10_size * condition, data = house_prices)
get_regression_table(house_price_model)

## # A tibble: 10 x 7
##    term              estimate std_error statistic p_value lower_ci upper_ci
##    <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
##  1 intercept             3.33     0.451     7.38    0        2.45     4.22
##  2 log10_size            0.69     0.148     4.65    0        0.399    0.98
##  3 condition2            0.047    0.498     0.094   0.925   -0.93     1.02
##  4 condition3           -0.367    0.452    -0.812   0.417   -1.25     0.519
##  5 condition4           -0.398    0.453    -0.879   0.38    -1.29     0.49
##  6 condition5           -0.883    0.457    -1.93    0.053   -1.78     0.013
##  7 log10_size:condi~    -0.024    0.163    -0.148   0.882   -0.344    0.295
##  8 log10_size:condi~     0.133    0.148     0.893   0.372   -0.158    0.424
##  9 log10_size:condi~     0.146    0.149     0.979   0.328   -0.146    0.437
## 10 log10_size:condi~     0.31     0.15      2.07    0.039    0.016    0.604
```

a) Why did we `log10()` transform the house price and house size in square feet variables first?

b) Using the numerical values in the above regression table, write the equation for the line for houses of

condition 1.

c) Using the numerical values in the above regression table, write the equation for the line for houses of condition 5.

d) Say a house get puts on the market in Seattle and you know nothing other than its size is 1000 square feet and it is of condition 5. What is the above model's prediction of this house's sale price in dollars?

e) Say instead you ran the following regression model below. Write down what all the elements of the `term` variable would be in the resulting regression table.

```
house_price_model <- lm(log10_price ~ log10_size + condition, data = house_prices)
get_regression_table(house_price_model)
```

24 Regression model using priming number

Continuing the previous `africa` question, we fit a regression where y is the number of countries guessed and x indicates which "priming group" a student was a part of:

```
model_countries_priming <- lm(countries ~ priming, data = africa)
get_regression_table(model_countries_priming)

## # 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              29.5      4.18      7.05       0     21.1     38.0
## 2 priming94 countri~     34.7      7.16      4.84       0     20.2     49.2
```

a) What does the `intercept` term in the `estimate` column of the regression table tell us? Answer in one sentence.

b) What does the `priming94 countr~` term in the `estimate` column of the regression table tell us? Answer in one sentence.

c) Say instead of using only two priming numbers, we used three: 0, 14, and 94 countries. In other words, we assigned students to one of three priming groups. Write down what the three terms in the left-most `term` column of the above regression table would now be.

d) Say you perform data wrangling to compute the mean number of countries guessed for each of the two priming groups. What are `XXX` and `YYY` in the table below? Your answers should be numerical values. Show your work.

```
## # A tibble: 2 x 2
##   priming      mean_guess
##   <chr>        <chr>
## 1 14 countries XXX
## 2 94 countries YYY
```

e) Say we run the following code and focus only on the first two rows out of the output (out of 41), corresponding to the first two students in the `africa` dataset. What are `XXX`, `YYY`, `AAA`, and `BBB` below? Your answers should be numerical values. Show your work.

```
get_regression_points(model_countries_priming)
```

```
## # A tibble: 2 x 5
##      ID countries priming      countries_hat residual
##   <int>     <dbl> <chr>        <chr>         <chr>
## 1     1        36 14 countries XXX           YYY
## 2     2       120 94 countries AAA           BBB
```

f) Do you think the number of countries guessed by those primed by "14" differs *significantly* from the number of countries guessed by those primed with "94"? Why? You will receive full credit for merely making a good faith attempt at answering. A "right answer" is not expected as you don't have the tools to answer this question . . . yet.

25 Regression model using height

Continuing the previous `africa` question, say you run the following regression instead, using `height` instead of `priming` as the explanatory/predictor variable:

```
model_countries_height <- lm(countries ~ height, data = africa)
get_regression_table(model_countries_height)

## # 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   162.       86.9       1.86   0.07    -14.1   338.
## 2 height       -1.86      1.34     -1.39   0.174    -4.57    0.854
```

a) Interpret the `intercept` term in the `estimate` column of the regression table, both mathematically and practically speaking ("practically" meaning in context of the data).

b) Give the precise interpretation of the slope for `height` in the `estimate` column of the regression table.

c) Say we run the following code and present only the first row of the output (out of 41 rows), corresponding to the first student in `africa`. What are `XXX` and `YYY`? Your answers should be numerical values. Show your work.

```
get_regression_points(model_countries_height)
```

```
## # A tibble: 1 x 5
##      ID countries height countries_hat residual
##   <int>     <dbl>  <dbl> <chr>         <chr>
## 1     1        36     70 XXX           YYY
```

d) Based on the regression model above, someone predicts that someone of height 54 inches will guess 62 countries. Why might this prediction inappropriate? Base your answer only on the various output of the analysis/model so far, and not prior knowledge or hypotheses you may have about the relationship between height and knowledge of the number of countries in Africa.

e) What would it mean for the relationship between height and the number of countries guessed if the slope for `height` in the table above were 0? Answer in practical and not mathematical terms ("practical" meaning in context of the data).

f) Do you think the observed slope for `height` of `-1.86` is *significantly* different from 0? Why? You will receive full credit for merely making a good faith attempt at answering. A "right answer" is not expected as you don't have the tools to answer this question . . . yet.

26 Teaching Evals

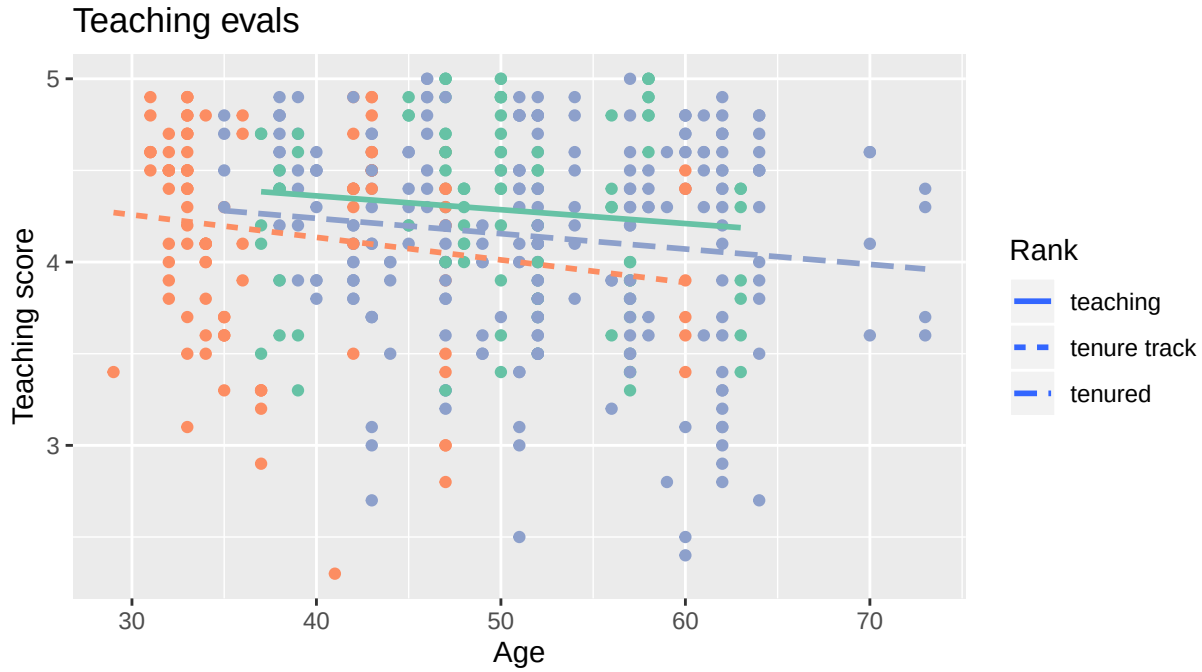
Recall the teaching evaluation data from class. We are interested in modeling the outcome variable $y =$ teaching `score` as a function of two explanatory variables:

1. x_1: numerical explanatory/predictor variable of the instructor's `age`
2. x_2: categorical explanatory/predictor of the instructor's `rank`: `teaching`, `tenure track` (AKA junior professor), or `tenured` (AKA senior professor)

Let's look at a random sample of 5 out of the 463 rows of this dataset:

score	age	rank
3.1	33	tenure track
4.1	45	tenured
5.0	46	tenured
4.8	52	tenured
3.7	50	tenured

Here is a visualization of the interaction model ...



. . . and the corresponding regression table. Note that the second to last row got truncated and should read `age:ranktenure track`.

```
model_1 <- lm(score ~ age * rank, data = evals)
get_regression_table(model_1)
```

```
## # 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              4.66      0.391    11.9     0        3.90     5.43
## 2 age                   -0.008     0.008    -0.979   0.328   -0.023    0.008
## 3 ranktenure track      -0.036     0.471    -0.077   0.939   -0.962    0.89
## 4 ranktenured           -0.09      0.442    -0.203   0.84    -0.958    0.779
## 5 age:ranktenure tr~    -0.005     0.01     -0.465   0.642   -0.025    0.015
## 6 age:ranktenured       -0.001     0.009    -0.095   0.924   -0.018    0.016
```

a) Using the numerical values in the above regression table, write the equation of the regression line for instructors with the rank `teaching`. Since you do not have a calculator, you do not need to perform the arithmetic. However write down all the arithmetic operations you would enter into a calculator.

b) Using the numerical values in the above regression table, write the equation of the regression line for instructors with the rank `tenured`. Since you do not have a calculator, you do not need to perform the arithmetic. However write down all the arithmetic operations you would enter into a calculator.

c) Interpret the slope for `age` -0.008.

d) Say there is a tenure track instructor of age 50 who just joins UT Austin. What is the fitted value $\widehat{y} = \widehat{\text{score}}$ for this instructor? i.e. What is their predicted teaching score? You **must** write a single numerical answer. Justify your answer. Hint: you don't need perform any arithmetic.

e) Say instead you ran the following regression model code. Write down what all the elements of the `term` variable in the resulting regression table.

```
model_2 <- lm(score ~ age + rank, data = evals)
get_regression_table(model_2)
```

f) Why might *some* people consider it reasonable to choose `model_2` over `model_1` to explain teaching score?

27 Life Expectancy

Note: for this question, you do not need to do the arithmetic (adding, subtracting, multiplying, dividing), but rather write down what you would enter into a calculator if you had one. Let's consider the `gapminder` development data, but only for the year 2007. Let's look at a random sample of 5 out of the 142 rows of this dataset:

country	continent	lifeExp
Togo	Africa	58
Sao Tome and Principe	Africa	66
Congo, Dem. Rep.	Africa	46
Lesotho	Africa	43
Bulgaria	Europe	73

We are interested in modeling the relationship between the outcome variable y = life expectancy in years and the categorical explanatory variable x = continent. You fit a following regression and obtain the following regression table rounded to the nearest integer:

```
## # A tibble: 5 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             54.8      1.02     53.4        0     52.8     56.8
## 2 continentAmericas     18.8      1.8      10.4        0     15.2     22.4
## 3 continentAsia         15.9      1.65      9.68       0     12.7     19.2
## 4 continentEurope       22.8      1.70     13.5        0     19.5     26.2
## 5 continentOceania      25.9      5.33      4.86       0     15.4     36.4
```

a) What is the fitted value $\widehat{y}$ of life expectancy in years for any given country in:

1. Africa
2. Asia
3. Europe

b) What is the residual for the following three countries?

1. Nambia
2. Iran
3. Italy

c) What is the mean life expectancy for countries in the following continents:

1. Africa
2. Asia
3. Europe

28 Sampling Scenarios

Consider the three scenarios below

- **Scenario 1**: You want to know the proportion of the balls in a sampling bowl of 2400 balls that are red. To this end, you mix the bowl first and use a shovel with 50 slots to pull out 50 balls. We observe that 20 of them are red.
- **Scenario 2**: We want to know the average year of minting of **all** pennies currently being used in the US. To this end, you go to Florence Bank in Downtown Northampton and ask the cashier to exchange a ten dollar bill for 1000 pennies. We observe that the average year of minting of these pennies is 2013.56
- **Scenario 3**: The instructor of SDS/MTH 220 wants to know what the effects are of priming with the numbers 14 and 94 on the number of countries Smith students guess are in Africa. To this end he conducts a priming experiment with all 38 of his students as done in class. He obtains the following fitted regression line based on the regression table output below:

$$\begin{aligned} \widehat{y} &= b_0 + b_1 \times x \\ \widehat{\text{countries}} &= b_0 + b_1 \times \mathbb{1}\,(\text{primed with 94}) \\ \widehat{\text{countries}} &= 29.5 + 34.7 \times \mathbb{1}\,(\text{primed with 94}) \end{aligned}$$

```
model_countries_priming <- lm(countries ~ priming, data = africa)
get_regression_table(model_countries_priming)

## # 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              29.5      4.18      7.05       0     21.1     38.0
## 2 priming94 countri~     34.7      7.16      4.84       0     20.2     49.2
```

a) On the next page there is a table. For all cells with a question mark, fill in what those values should be.

Scenario	1	2	3
Population	$N = ?$	$N = ?$	$N = ?$
Population parameter name	?	?	Population slope
Population parameter mathematical notation	?	?	β_1
Sample size	$n = ?$	$n = ?$	$n = ?$
Point estimate name	?	?	Fitted slope
Point estimate mathematical notation	?	?	b_1
Point estimate numerical value	?	?	?

b) Is the point estimate for the population parameter in Scenario 1 a good one? Why or why not? Answer in three sentences or less.

c) Is the point estimate for the population parameter in Scenario 2 a good one? Why or why not? Answer in three sentences or less.

d) Is the point estimate for the population parameter in Scenario 3 a good one? Why or why not? Answer in three sentences or less.

29 Sampling Distribution

a) Recall the virtual `bowl` consisting of 2400 balls from the `moderndive` package. Let's show the first 10 rows of the data set:

```
bowl
```

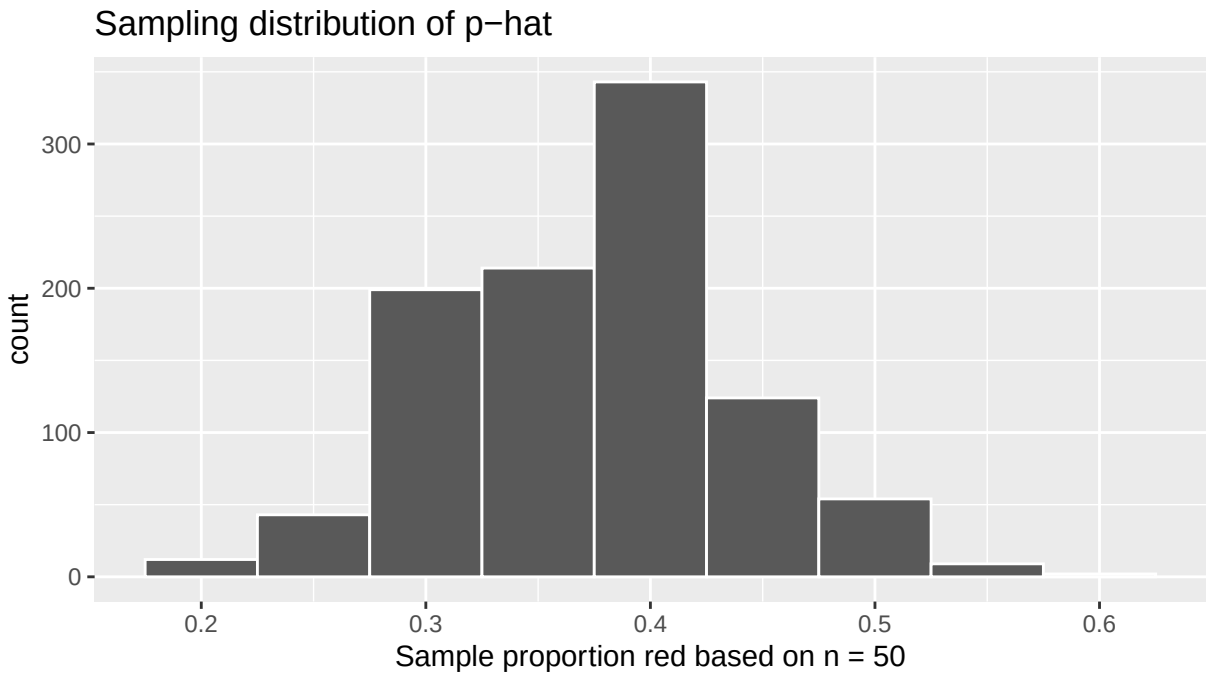
```
## # A tibble: 2,400 x 2
##    ball_ID color
##      <int> <chr>
##  1       1 white
##  2       2 white
##  3       3 white
##  4       4 red
##  5       5 white
##  6       6 white
##  7       7 red
##  8       8 white
##  9       9 red
## 10      10 white
## # ... with 2,390 more rows
```

From this virtual bowl we can draw virtual samples using the `rep_sample_n()` function. For example:

```
bowl %>%
  rep_sample_n(size = 3, reps = 2)
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [2]
##   replicate ball_ID color
##       <int>   <int> <chr>
## 1         1    2287 white
## 2         1     599 white
## 3         1     108 white
## 4         2     846 red
## 5         2     390 red
## 6         2     344 white
```

Recall that we ran a simulation creating the sampling distribution of $\widehat{p}$ based on 1000 samples of size $n = 50$ drawn using the virtual shovel:



Write out the pseudocode that will produce the visualization of the above sampling distribution. Feel free to write in actual code if you like. Hint: your pseudocode should start with `bowl` and use the `rep_sample_n()` function from earlier.

b) What will happen to the above sampling distribution if we used a virtual shovel with $n = 100$ slots?

c) What is the standard deviation of the above sampling distribution called?

30 U.S. Elections

Buzzfeed News performed a poll of 1024 individuals a week before the election to try to determine the percentage support for Candidate X and found that 47.8% supported Candidate X.

1. Who is the population?
2. What is the population parameter?
3. What is the sample?
4. What is the statistic?
5. Under what conditions are the results based on the sample generalizable to the population?

it's all **politics**

Poll: Support For Obama Among Young Americans Eroding

December 4, 2013 · 12:53 PM ET

ADAM WOLLNER

President Obama speaks at a town hall meeting at Binghamton University in Vestal, N.Y., in August.

Mike Groll/AP

After voting for him in large numbers in 2008 and 2012, young Americans are souring on President Obama.

According to a new Harvard University Institute of Politics poll, just 41 percent of millennials — adults ages 18-29 — approve of Obama's job performance, his lowest-ever standing among the group and an 11-point drop from April.

Obama's signature health care law is also unpopular among millennials. Fifty-seven percent of those surveyed said they disapprove of Obamacare, compared with 38 percent who said they approve.

A majority of respondents also said they disapprove of the way Obama is handling the economy, Syria, Iran and the budget deficit.

The results reflect a similar downward trend among the public at large. Recent polls ranging from Gallup to CNN show Obama's approval rating hovering around 40 percent, while disapproval of the health care law is in the mid-to-high 50s.

"Millennials are starting to look a lot more like their older brothers and sisters, parents and grandparents," IOP polling director John Della Volpe said in a conference call with reporters Wednesday.

The online survey of 2,089 adults was conducted from Oct. 30 to Nov. 11, just weeks after the federal government shutdown ended and the problems surrounding the implementation of the Affordable Care Act began to take center stage. The poll's margin of error was plus or minus 2.1 percentage points.

Fifty-five percent of the survey's respondents said they voted for Obama in the last presidential election, while 33 percent said they voted for Republican Mitt Romney. If the election were held again, Obama would still come out on top, but by a tighter 46 to 35 percent vote; 13 percent said they would vote for someone else.

According to the Pew Research Center, 66 percent of 18- to 29-year-olds voted for Obama in 2008, and 60 percent voted for his re-election in 2012.

Harvard's poll found millennials, like the rest of the public, aren't happy with Congress either. Just 19 percent of respondents said they approve of congressional Republicans, while 35 percent approve of their Democratic counterparts. Both figures are single-digit drops from April. Forty-five percent also said they would "recall and replace" their member of Congress if they had the option.

millennials president obama

31 Polling

The previous two pages are an NPR article reporting the results of a Harvard University Institute of Politics poll conducted in between October 30 and November 11, 2013 of millennials' (adults aged 18-29) approval of President Obama's job performance.

a) Who is the study population?

b) What is the name/terminology for the population parameter of interest described in the second paragraph?

c) What is the mathematical notation for the population parameter of interest described in the second paragraph?

d) Say we had access to infinite resources, what would be the best way to measure the population parameter of interest?

e) What is the best way to ensure that the sample of n young Americans is representative of the study population?

f) What is the name/terminology of the point estimate/sample statistic for this poll?

g) What is the mathematical notation for the point estimate/sample statistic for this poll?

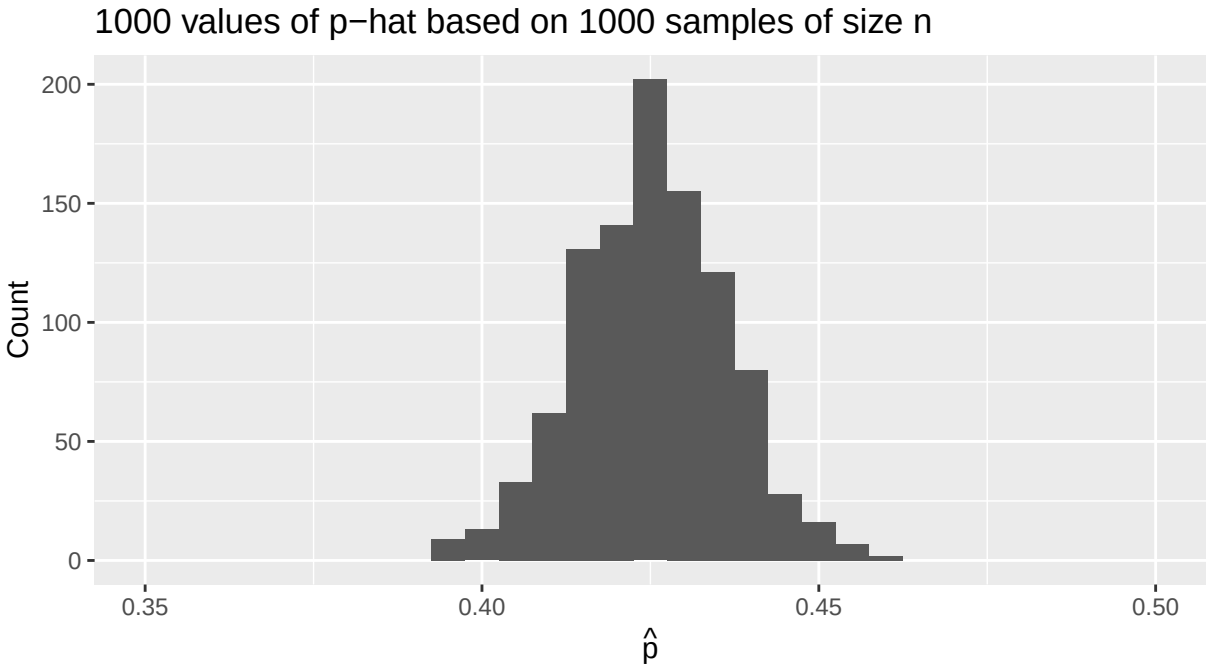
h) What is the numerical value of the point estimate/sample statistic for the poll referred to in the second paragraph?

32 Sampling

Consider the following **hypothetical** extension of the previous question on the poll of Obama's approval ratings among young Americans in 2013. Say the following 1000 polling companies conduct polls on the same dates as the Harvard poll, using the same sampling methodology and the same sample size $n = 2089$. You observe the following results:

- Poll 1: Gallup finds that out of their sample of size n, 844 young Americans approve of Obama.
- Poll 2: Ipsos finds that out of their sample of size n, 857 young Americans approve of Obama.
- ...
- Poll 1000: Monmouth University finds that out of their sample of size n, 871 young Americans approve of Obama.

Based on these 1000 polls based on samples of size n, you compute 1000 values of $\widehat{p}$. You plot these in a histogram:



a) The above normal-shaped distribution is called the "**X**" of the sample proportion $\widehat{p}$. It shows the results of a simulation illustrating how different values of $\widehat{p}$ vary from sample to sample due to sampling variation. What is "**X**"?

b) If the sampling in all cases is done in a representative fashion, what is the name of the value where this histogram centered? Do not give the numerical value, which one can observe is 0.425, rather the appropriate terminology. Hint: the answer is not simply the "mean" or "median."

c) What is the precise name of the term that quantifies the spread of the histogram above.

d) What would happen to this histogram if the sample size were $n = 5000$?

e) Assuming the sampling is done at random, would observing a poll with $\widehat{p} = 0.40$ be more likely when the sample size of the poll is 2089 or 5000?

f) Why are the polling results better when using a sample size of $n = 5000$ instead of $n = 2089$?

g) Bringing things back to real life and focusing on what we would do in practice, we would *not* take 1000 different samples of size $n = 2089$ and compute 1000 different values of $\widehat{p}$, but instead take only a single sample of $n = 2089$? How can we still study variation in the $\widehat{p}$ due to sampling variation using only a single sample? **Answer in two sentences or less**.

h) There is actually a mathematical formula for the term in part c). Which is the correct formula for this term: A) $\sqrt{p(1-p)}$ or B) $\sqrt{\frac{p(1-p)}{n}}$? How do you know this?

33 Putting it all together: "Formatting is off"

The office of the president of a small liberal arts college in New England wants to promote the public launch of a fundraising campaign to all alumni. In particular, they would like the email to include a quote by American novelist and essayist Marilynne Robinson, followed by a link to donate money. However, the president is concerned that the formatting of the email will affect the "click-through rate" of the link: the proportion of those receiving the email that follow through and click on the link. In particular, they are **very** concerned about any possible differences in click-through rates arising due to the formatting of the quote attribution. So the office creates the two versions of the same email where the only difference is the quote attribution:

Version 1:

President's Office <president@[redacted].edu>
Reply-To: President's Office <president@[redacted].edu>
To: [redacted]

In the U.S., education, especially at the higher levels, is based around powerful models of community. We choose our colleges in order to be formed by them and supported by them in the identities we have or aspire to. If the graft takes, we consider ourselves ever after to be members of that community. As one consequence, graduates tend to treat the students who come after them as kin and also as heirs. They take pride in the successes of people in classes forty years ahead of or behind their own. They have a familial desire to enhance the experience of generations of students who are, in fact, strangers to them, except in the degree that the ethos and curriculum of the places does indeed form its students over generations.
Marilynne Robinson

Version 2:

President's Office <president@[redacted].edu>
Reply-To: President's Office <president@[redacted].edu>
To: [redacted]

In the U.S., education, especially at the higher levels, is based around powerful models of community. We choose our colleges in order to be formed by them and supported by them in the identities we have or aspire to. If the graft takes, we consider ourselves ever after to be members of that community. As one consequence, graduates tend to treat the students who come after them as kin and also as heirs. They take pride in the successes of people in classes forty years ahead of or behind their own. They have a familial desire to enhance the experience of generations of students who are, in fact, strangers to them, except in the degree that the ethos and curriculum of the places does indeed form its students over generations.

--Marilynne Robinson

Here is the sequence of events:

- They randomly select 25,138 alumni from the alumni database to send emails to.
- From these 25,138 alumni, they randomly choose 12,460 alumni and send them email Version 1. They send Version 2 to the remaining 12,678 alumni.
- Of those alumni who received Version 1 10,578 followed through and clicked the link for a rate of 84.9%. Of those alumni who received Version 2 11,169 followed through and clicked the link for a rate of 88.1%.

a) What kind of study are we considering: an observational study or a randomized experiment? Why?

b) In this scenario, can we establish the *causal* effect (and not just the *associated* effect) of the formatting on click-through rate? Why or why not?

c) Who is the study population in this scenario?

d) What is the statistical name of the population parameter of interest in this scenario?

e) What is the mathematical notation for the population parameter of interest in this scenario?

f) What is the statistical name for the point estimate (AKA sample statistic) of the population parameter of interest in this scenario?

g) What is the mathematical notation for the point estimate of interest in this scenario?

h) What is the numerical value of the point estimate of interest in this scenario?

i) The standard error of the point estimate in this question can roughly be approximated by the mathematical formula when constructing confidence intervals:

$$\text{SE}_{\widehat{p}_1-\widehat{p}_2} \approx \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

Construct a 95% confidence interval appropriate to answer the president's concerns.

j) We say that we are "95% confident" that this confidence interval captures the true value of the unknown population parameter. However, we say this as shorthand for the more involved statistical interpretation of a confidence interval. What is this precise statistical interpretation?

k) Write down the relevant hypothesis test using non-statistical language.

k) Write down the relevant hypothesis test using mathematical notation.

l) What is the statistical name of the relevant test statistic in this scenario?

m) What is the numerical value of the *observed* test statistic in this scenario?

m) What is being assumed throughout this hypothesis testing scenario?

m) The standard error of the point estimate in this question can roughly be approximated by the mathematical formula when conducting hypothesis tests:

$$\text{SE}_{\widehat{p}_1 - \widehat{p}_2} \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2}}$$

where $\widehat{p}$ is the *pooled sample proportion* where you pool all observations in both groups into a single group and compute a single proportion:

$$\widehat{p} \quad = \quad \frac{\text{\# of Version 1 link clicks} + \text{\# of Version 2 link clicks}}{n_1 + n_2}$$

For hypothesis testing, why is this pooling appropriate?

n) Draw the null distribution. Hint: while not always the case, the sampling distribution of the point estimate of interest is normally shaped.

j) Recall that the president is **very** concerned about any possible differences in click-through rates arising due to the formatting of the quote attribution. If conducting a hypothesis test, would you use a "liberal" α value or a "conservative" α value?

k) Based on your response above, conduct the hypothesis test.

l) Based on your analysis, what do you tell the President? Keep in mind the President is very busy (monitoring the formatting of emails for example), so they would prefer a shorter response.

34 Evals continued

Recall the evals data of teaching evaluations of professors. Let say instead that these 463 professors are a randomly chosen set of instructors from all of the University of Texas system and not just UT Austin. Consider the following *simple linear regression* using only one numerical explanatory variable:

```
score_model <- lm(score ~ age, data = evals)
get_regression_table(score_model)
```

term	estimate	std_error	lower_ci	upper_ci
intercept	4.46	0.13	4.21	4.7
age	-0.01	0.00	-0.01	0.0

a) Interpret the slope coefficient for age.

b) Using statistical language, interpret the standard error for the slope for age.

c) Using non-technical language, interpret the standard error for the slope for age.

35 Hypothesis Testing

A friend of yours claims to be psychic. You are skeptical. To test this you take a stack of $n = 100$ playing cards and have your friend try to identify the suit, either hearts/diamonds/clubs/spades, without looking. They get 45 out of 100 right. **Please start writing all your responses where indicated below.**

a) Define a hypothesis test in non-statistical terms where H_0: Your friend is not psychic.

b) Define the corresponding hypothesis test in statistical terms. Hint: it should involve p.

c) What is the standard error of $\widehat{p}$ used for hypothesis testing in general?

d) What is the standard error of $\widehat{p}$ used for this particular hypothesis test?

e) Draw the distribution of how different values of $\widehat{p}$ based on different samples of size $n = 100$ will behave from sample-to-sample. In particular focus on 1) its shape, 2) its center, and 3) where the middle 95% of values of $\widehat{p}$ will lie. Be as precise as possible, in other words do not use any "rules of thumb."

f) In the plot above, mark with a dashed vertical line where the observed value of $\widehat{p}$ lies.

g) Even though you don't have access to a computer, make a guess about the conclusion of the hypothesis test. State your conclusions both statistically and in terms of a statement on your friend being psychic.

h) BONUS Say you did have access to a computer right now. In order to compute the p-value above exactly, what values of `A`, `B`, and `C` must you input?

```
library(mosaic)
xpnorm(A, mean = B, sd = C)
```

36 Hypothesis Testing

The General Social Survey (GSS) provides information on educational attainment and gender. In a random sample in 2010 of 584 male Americans and a random sample of 118 female Americans, the GSS concluded that males obtained on average 13.66 years of education compared to 12.89 years for females, which leads us to an observed difference of 0.77 years. Layout the "There is Only One Test" framework for testing whether these collected samples provide evidence for the claim that male Americans obtain *less* education on average than female Americans, using simulation/computing-based constructions and not mathematical ones. You will not actually be able to conduct this test without a computer, but sketch out the procedure and make a guess as to the test's conclusion using $\alpha = 0.1$ as best you can.

37 Confidence Intervals

Recall we saw an example of an NPR poll of $n = 2089$ young Americans' approval of Obama back in 2013. Of these respondents, 856 said they approved of Obama's job performance.

a) What is the numerical value of $\widehat{p}$, the point estimate of the population proportion p of all young Amercians who approve of Obama's job performance?

b) Say CBS conducted a similar poll with $n = 2089$ and finds that 860 young Americans approve of Obama, leading to one point estimate $\widehat{p}$ of p. Say NBC conducted a similar poll with $n = 2089$ and finds that 844 young Americans approve of Obama, leading to another point estimate $\widehat{p}$ of p. Say Buzzfeed News conducted a similar poll with $n = 2089$ and finds that 871 young Americans approve of Obama, leading to yet another point estimate $\widehat{p}$ of p. What is the name of the value that quantifies this variability?

c) Construct a 95% confidence interval for the population proportion p of all young Americans who approved of Obama's job performance. Note the following mathematical formula approximating the standard error:

$$\text{SE}_{\widehat{p}} \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

d) Marc-Edouard Vlasic states "I read on NPR that back in 2013, as little as 43% of **all** young Americans approved of Obama." What assumption must be met for Marc-Edouard's statement to be valid?

e) What assumption about the sampling distribution of $\widehat{p}$ must be met for the confidence interval in part c) to be valid?

38 Inference for Regression

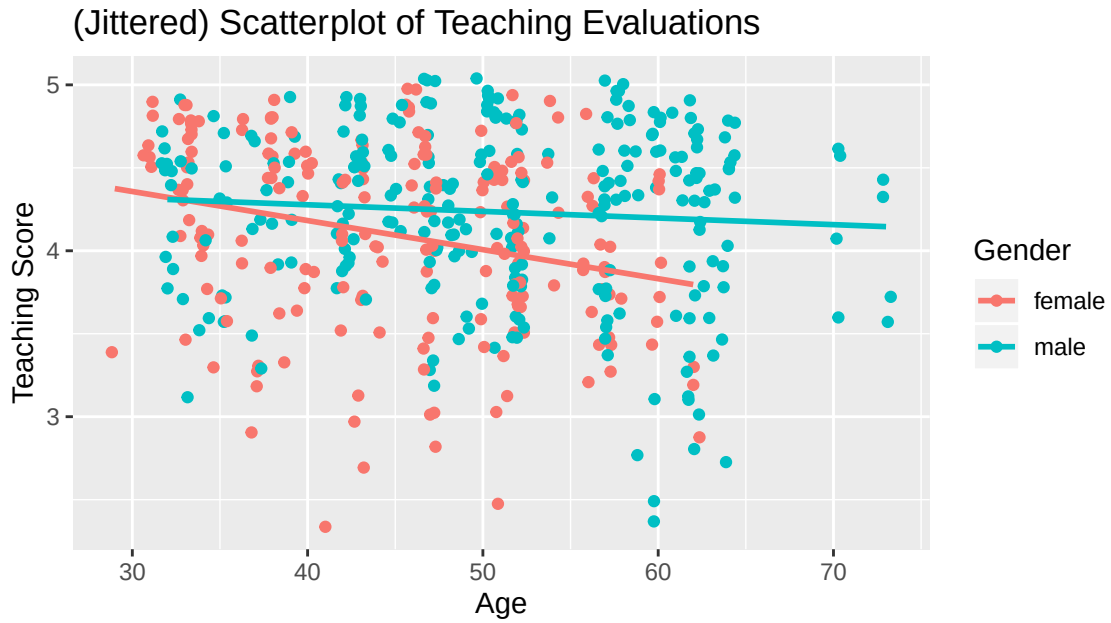
Recall our professor evaluations dataset based on the study from the University of Texas in Austin. In particular, we were interested in explaining a professor's teaching evaluation score using their gender and age as explanatory variables. Here is a random sample of 5 rows out of the $n = 463$ professors in dataset:

```
## # A tibble: 5 x 3
##   score gender   age
##   <dbl> <fct>  <int>
## 1   4.9 male      50
## 2   4.3 male      48
## 3   5   male      50
## 4   4.5 female    52
## 5   4.8 male      43
```

Recall we fit the following regression model *with an interaction term*:

$$\begin{aligned} \widehat{y} &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 \\ \widehat{\text{score}} &= b_0 + b_{\text{age}}\text{age} + b_{\text{male}}\mathbb{1}[\text{is male}] + b_{\text{age,male}}\text{age}\mathbb{1}[\text{is male}] \end{aligned}$$

Recall the visual representation of the our model. Hint: look at this closely.



Finally, recall the results of the regression with confidence intervals

```
evals_model <- lm(score ~ age * gender, data=evals)
get_regression_table(evals_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.88	0.20	23.8	0.00	4.48	5.29
age	-0.02	0.00	-3.9	0.00	-0.03	-0.01
gendermale	-0.45	0.26	-1.7	0.09	-0.97	0.08
age:gendermale	0.01	0.01	2.5	0.02	0.00	0.02

a) The table reports a p-value of 0 in the age row. Write down the corresponding hypothesis H_0 vs H_A in terms of the β_{age}, the true population associated effect of age on teaching score.

b) The p-value mentioned in part a) is 0. Report what this means for the hypothesis test corresponding to the two hypotheses above. Report this both in 1) statistical terms and 2) language that non-statisticians can understand.

c) Based on these results, among male professors at the University of Austin for every year increase in age, there is an associated X of on average Y units in teaching score. What are X and Y?

d) What conclusion is suggested by the 95% confidence interval for $\beta_{\text{age:gendermale}}$ of $(0.003, 0.024)$?

e) Say we relaxed the gender categorical variable to allow for the following three levels: female, male, and non-binary, and furthermore say some professors selected the new "non-binary" option. Describe precisely how the above plot would change.

f) BONUS 1 Describe precisely how the shape of the above regression table would change.

g) BONUS 2 The 95% confidence interval for $\beta_{\text{gendermale}}$ is $(-0.968, 0.076)$. Based on values in the table, write down your best guess of the formula that R uses to compute the left end point of -0.968. Your formula and the reported left endpoint of -0.968 should match up to 2 decimal places.

39 Regression

You run the code below to analyze departure delays from the 3 New York City airports, but for some weird reason, you only get the incomplete output below. Note AS corresponds to Alaska, F9 corresponds to Frontier, and AA corresponds to American.

```
library(dplyr)
library(nycflights13)
library(moderndive)

flights_subset <- flights %>%
  filter(carrier == "AS" | carrier == "F9" | carrier == "AA")

dep_delay_model <- lm(dep_delay ~ carrier, data = flights_subset)
get_regression_table(dep_delay_model, digits = 3)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	8.6	0.21	40.7	0.000	8.2	8.999
carrierAS	-2.8	1.43	-1.9	0.052	-5.6	0.025
carrierF9	11.6	1.46	8.0	0.000	8.8	NA

a) Interpret the 11.6 estimate value in the `carrierF9` row (third row, second column). Is its relationship of with the outcome variable meaningful?

b) Compute the missing right endpoint of the 95% confidence interval in the `carrierF9` row.

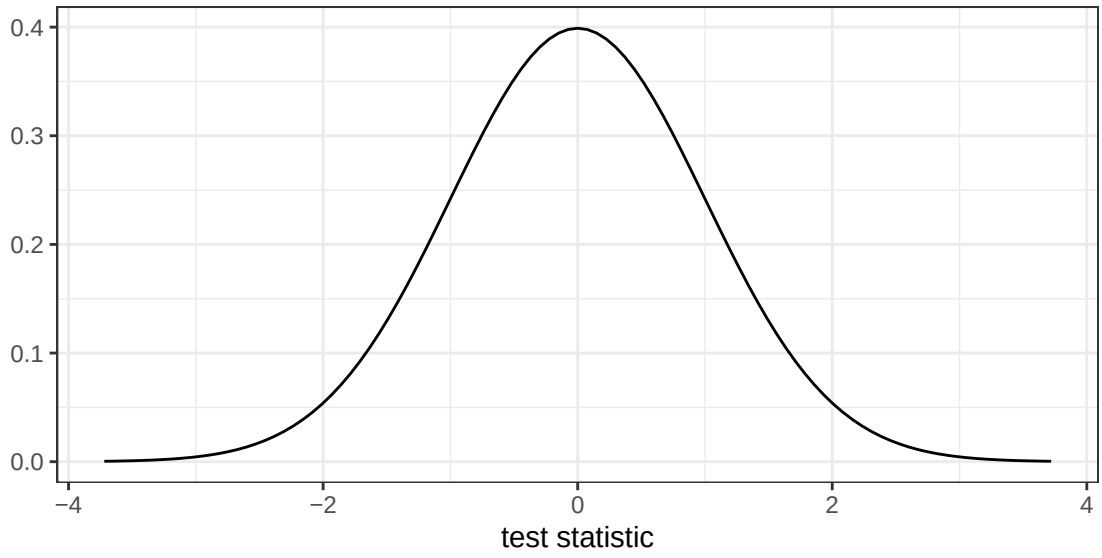
c) State the scientific conclusion reached based on the now complete 95% confidence interval.

e) Write down the hypothesis test corresponding to the `carrierAS` row using mathematical notation. Do not carry out the hypothesis test, simply state the two competing hypotheses.

f) Say you were given an α cutoff value of 0.01 for the hypothesis test above. Write down the conclusion of this hypothesis test both in statistical terms and using non-statistical language that an airline executive can understand.

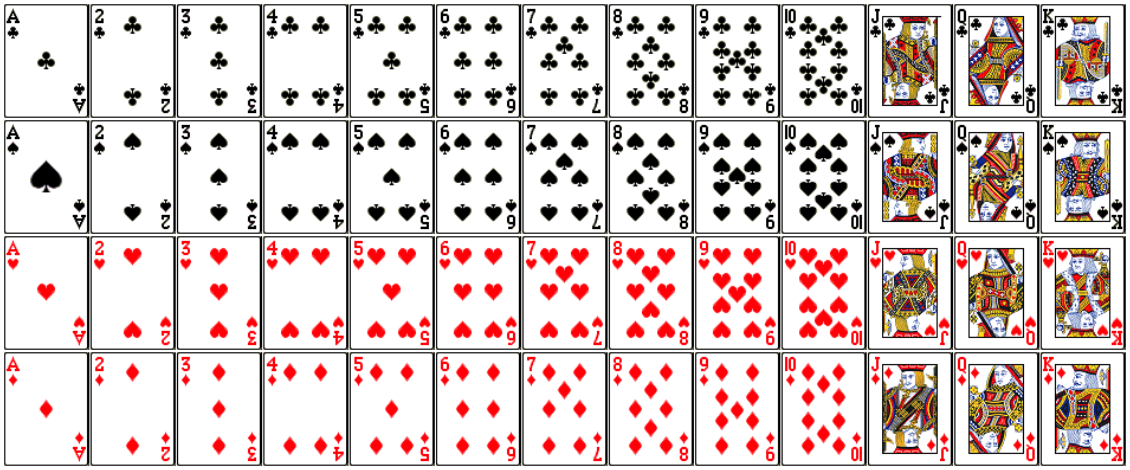
c) In the second row, fifth column there is a p-value missing. What is the hypothesis test corresponding to this missing p-value?

d) Sketch on the follow plot of the corresponding *null* distribution what the missing p-value in the second row, fifth column is:


0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
test statistic

40 Simulation

You are interested in studying the probabilities behind the game of poker. In poker, each player is dealt a *hand* of 5 cards chosen at random from a deck of 52 cards. Among the stronger cards in poker are the Ace cards (denoted below by A's).



a) Say you are playing poker by yourself. Using the tools from the `mosaic` package, write out the pseudocode of the simulation necessary to study the random number of Aces included in a hand of poker across many, many, many hands.

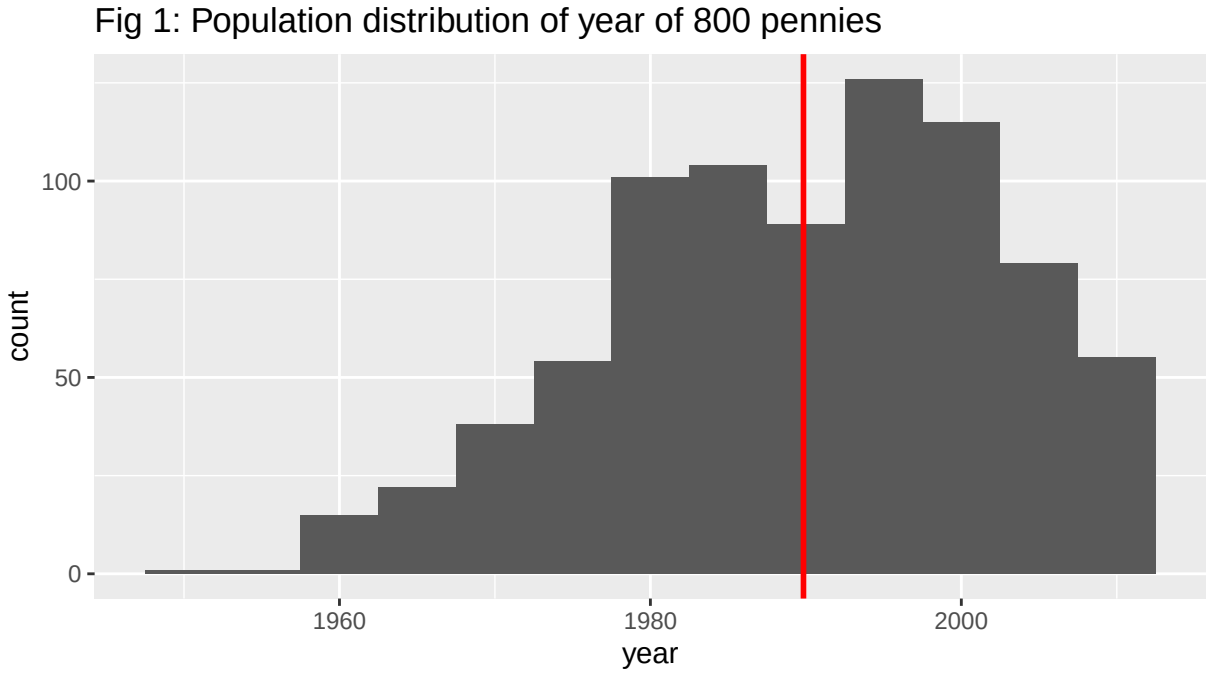
b) Write out the pseudocode to generate an appropriate graphic to show the **probability distribution** of the number of aces in a hand of poker based on the output of part a) above.

c) Draw as best you can what this graphic looks like.

41 Pennies

Recall the "sack" of $N = 800$ pennies in the dataframe `pennies` from which we virtually sampled $n = 50$ pennies from in Problem Set 10.

```
library(ggplot2)
library(dplyr)
library(moderndive)

mean(pennies$year)

## [1] 1990

sd(pennies$year)

## [1] 12

ggplot(pennies, aes(x = year)) +
  geom_histogram(binwidth = 5) +
  labs(x = "year", title = "Fig 1: Population distribution of year of 800 pennies") +
  geom_vline(xintercept = mean(pennies$year), col = "red", size = 1)
```



START WRITING YOUR RESPONSES WHERE INDICATED BELOW.

a) What are the (study) population, the name of the population parameter, the numerical value of the population parameter, the name of the point estimate, the formula for the true standard error, and the numerical value of the true standard error.

b) Say you (virtually) sample $n = 50$ pennies from the above population, compute the sample mean, and replace the pennies. Then everyone on your floor does the same. In what range of values do you expect 95%

of the resulting sample means to lie?

c) Say at the end of the day, you have 679 such sample means and you plotted a histogram of it. What is the name (not shape) of this distribution?

d) Why is the distribution of sample means in c) normal even though the population distribution of the $N = 800$ pennies is left-skewed as seen above?

e) In practice we would never perform such a simulation; rather we would just take a single sample of size n. What is the point of this simulation then?

f) Say all your floormates are germophobes and instead of randomly sampling pennies, they selectively only select the shinier and cleaner pennies. What impact will this have on the distribution in part c)?

g) Say instead the population of interest is not the sack of 800 pennies, but all pennies in circulation in the US. List the steps needed to construct a 99% confidence interval for the population parameter of interest based on $n = 50$.