431 Quiz 2 for Fall 2023

Deadline: Tuesday 2023-12-05 at 3 PM

Thomas E. Love

2023-11-30

Table of contents

Instructions for Students

There are **24** questions on this Quiz and this PDF is **39** pages long. Be sure you have all **39** pages. It is to your advantage to answer all **24** questions. Your score is based on the number of correct responses, so there's no chance a blank response will be correct, and a guess might be, so you should definitely answer all of the questions.

This is an open book, open notes quiz. You are welcome to consult the materials provided on the course website and that we've been reading in the class, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants. You will be required to complete a short affirmation that you have obeyed these rules as part of submitting the Quiz.

0.1 The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, located at…

- https://bit.ly/431-2023-quiz2-form

All of your answers must be submitted through the Google Form by 3 PM on Tuesday 2023-12-05, without exception. The form will close at 3:30 PM on that date, and no extensions will be made available, so do not wait until Tuesday afternoon to submit your work.

The Google Form contains places to provide your responses to each question, and a final affirmation where you'll type in your name to tell us that you followed the rules for the Quiz. You must complete that affirmation before you can submit your responses. When you submit your results (in the same way you submit a Minute Paper) you will receive an email copy of your submission, with a link that will allow you to edit your work.

If you wish to work on some of the quiz and then return later, you can do this by [1] completing the final question (the affirmation) which asks you to type in your full name, and then [2] submitting the quiz. You will then receive a link at your CWRU email which will allow you to return to the quiz without losing your progress.

0.2 The Data Sets

I have provided **five** data sets which may be of help. You will find the files in our Shared Drive, under the Quiz 2 materials folder. They are:

- `movies.csv` (7 columns, 201 rows) first mentioned in Question 8
- `nh_adult750.Rds` (16 columns, 750 rows) first mentioned in Question 18
- `nnyfs.Rds` (45 columns, 1518 rows) first mentioned in Question 5
- `optical.csv` (3 columns, 69 rows) first mentioned in Question 15
- `projA.xlsx` (4 columns, 42 rows) first mentioned in Question 11

0.3 Getting Help

If you need clarification on a Quiz question, you have exactly one way of getting help:

1. Ask your quiz question via email to **431-help at case dot edu**.

During the Quiz period (5 PM 2023-11-30 through 3 PM 2023-12-05) we will not answer questions about the Quiz through Campuswire or in TA office hours. Instead, we will only answer them through the email address listed above. We promise to respond to all questions received before 9 AM on 2023-12-05 in a timely fashion.

A few cautions:

- Specific questions are more likely to get helpful answers.
- We will not review your code or your English for you.
- We will not tell you if your answer is correct, or if it is complete.
- We will email all students if we find an error in the Quiz that needs fixing.

0.4 When Should I Ask for Help?

We recommend the following process.

- If you encounter a tough question, skip it, and build up your confidence by tackling other questions.
- When you return to the tough question, spend no more than 10-15 minutes on it. If you still don't have it, take a break (not just to do other questions) but an actual break.
- When you return to the question, it may be much clearer to you. If so, great. If not, spend 5-10 minutes on it, at most, and if you are still stuck, ask us for help.
- This is not to say that you cannot ask us sooner than this, but you should **never, ever** spend more than 20 minutes on any question (other than the two essays) without asking for help.

0.5 Scoring and Timing

Questions 1 and 24 are essays, worth 6 points each (and will take longer than the other questions), while the other 22 questions are worth 4 points each, adding to a total of 100 points. Questions 2-23 are not in any particular order, and range in difficulty from "things I expect everyone to get right" to "things that are deliberately tricky".

The Quiz is meant to take **6** hours. This is based on the assumption that questions 2-23 will take an average of 10 minutes each, and that questions 1 and 24 will take about 40 minutes each, with an extra hour at the end to check things over and revisit the trickier questions. I expect most students will take **4-8** hours, and some will take as little as **2-3** or as many as **9-10**.

0.6 Writing Code into the Google Form

Occasionally, we ask you to provide a single line of code. If not otherwise specified, a single line of code in response can contain **at most** two pipes, although you may or may not need the pipe in any particular setting. Note that I exclusively used the `|>` pipe, and not the `%>%` pipe, in developing this Quiz.

Moreover, you need not include the `library` command at any time for any of your code. Assume in all questions that all of the packages listed below have been loaded in R.

0.7 R Packages and Love-boost.R script

This doesn't mean I actually used all of these packages. It means that this list includes everything I did use in creating the Quiz and its answers, plus some other things.

```
library(boot); library(broom); library(car); library(corrr)
library(Epi); library(googlesheets4); library(GGally); library(ggdist)
library(ggrepel); library(ggridges); library(glue); library(gt)
library(haven); library(Hmisc); library(janitor); library(kableExtra)
library(knitr); library(MASS); library(mice); library(mitml)
library(mosaic); library(naniar); library(patchwork); library(pwr)
library(readxl); library(simputation); library(vcd); library(xfun)
library(tidyverse)

source("data/Love-boost.R")

theme_set(theme_bw())
gs4_deauth()

opts_chunk$set(comment = NA)
options(tidyverse.quiet = TRUE)
options(dplyr.summarise.inform = FALSE)
```

The version of `Love-boost.R` that I am using is posted at https://raw.githubusercontent.com/THOMASELOVE/431-data/main/data-and-code/Love-boost.R and on our Shared Drive in the Quiz 2 Materials folder.

The chunk of code listed above (including the link to Love-boost.R on our 431-data page) is available to you at https://github.com/THOMASELOVE/431-quizzes-2023/tree/main/quiz2 so that you can copy and paste it into your own document.

1 Question 1 (6 points)

Here is a ChatGPT response to the question "Is statistical significance important?" obtained earlier this year. Note that I've numbered the five paragraphs.

1. Statistical significance is an important concept in statistical analysis and can be useful in making decisions based on data. Statistical significance is a measure of whether an observed effect in a sample is likely to be due to chance or if it reflects a real difference or relationship in the population from which the sample was drawn.
2. If an effect is statistically significant, it means that there is strong evidence that the effect is not due to chance alone. This can be helpful in making decisions such as whether to approve a new drug, adopt a new policy, or change a marketing strategy.
3. However, it's important to keep in mind that statistical significance is just one piece of information and should be considered in conjunction with other factors such as effect size, practical significance, and the context in which the data was collected.
4. Additionally, statistical significance does not necessarily mean that an effect is important or meaningful in a real-world context. For example, a study might find a statistically significant difference between two groups, but the difference might be so small that it has little practical significance.
5. So, while statistical significance can be an important concept in statistical analysis, it should be interpreted and used with caution and in the context of other relevant factors.

Select one of the five paragraphs listed in the ChatGPT response above, and write a short critique **of that paragraph** in light of the 2019 Editorial "Moving to a World beyond 'p < 0.05'" by Wasserstein, Schirm, and Lazar.

Your goals in your essay (critique) should be to

- improve the paragraph produced by ChatGPT, **and**
- provide meaningful context for your suggested improvements based on what you've learned from the Editorial.

Your critique (essay) should consist of 5-12 complete English sentences, and include at least one specific quote from the Editorial as part of your response. It is appropriate to quote materials from section 1-7 of the Editorial, including the "Authors' Suggestions" section, so long as you use quotation marks and identify the section of the Editorial your quote comes from. As a benchmark, we want 80% (or more) of the words in your response to be yours alone, as opposed to quotes from ChatGPT or the Editorial.

2 Question 2

Two types of artificial knee are to be compared for range of motion, measured in degrees. Theoretically, either could give a greater range. A journal article on the first type of knee gave a sample mean of 112 degrees, with a standard deviation of 13 degrees, and an article on the second type gave a sample mean of 118 degrees with a standard deviation of 11 degrees.

We want to perform a new randomized trial to decide whether a 6 degree difference is statistically detectable using a 5% significance level, and maintaining at least 90% power. We are willing to assume that the population standard deviation is somewhere between the sample standard deviations reported in the two articles. What is the minimum number of subjects receiving each type of knee (in a balanced design) we must record?

Hint: Your answer should be a number, representing the number of subjects receiving a Type 1 knee (which will be equal to the number receiving a Type 2 knee).

3 Question 3

Five different regression models were fit to the same outcome, using different predictors, but fit to the same 245 observations, resulting in the following summary statistics. There were no missing values in the data.

Model	R^2	Adjusted R^2	$\hat{\sigma}$	AIC	Sample Size	df
`mod_1`	0.6542	0.6307	22.3	-462	245	4
`mod_2`	0.6492	0.6227	20.8	-458	245	5
`mod_3`	0.6812	0.6247	21.9	-483	245	4
`mod_4`	0.6602	0.6507	22.1	-498	245	6
`mod_5`	0.6342	0.6197	21.0	-452	245	5

Your job is to identify the best of the five models in terms of each of the four criteria presented in the table.

Columns:

- `mod_1`, `mod_2`, `mod_3`, `mod_4`, `mod_5`

Rows:

1. Highest proportion of variation in the outcome explained
2. Best value of adjusted R^2
3. Smallest residual standard deviation
4. Best value of AIC

4 Question 4

Suppose you have calculated a 90% confidence interval for the population proportion of success, using data from a sample of 200 subjects, each of whom has a result classified as either a success or failure. To calculate this original interval, you used the Wald method and the `binom.test` function from R's `mosaic` package.

Now, you are considering several potential changes in your approach to estimating the confidence interval. Which of the following changes will definitely make the new interval wider than your original one? (CHECK ALL CHANGES THAT WOULD DEFINITELY WIDEN THE INTERVAL.)

a. Using `ci.method = "Score"` in R to create the interval, without a continuity correction.
b. Switching to a 95% confidence level.
c. Setting the value of `p` to 0.25 in the call to `mosaic::binom.test`.
d. Switching the α level to 0.20.
e. Using a new sample of 150 subjects, which have the same sample proportion as the original sample.
f. Adding two successes and two failures to the data.
g. None of these changes will make the resulting confidence interval wider.

5 Question 5

This question uses the `nnyfs.Rds` data set, which is found on our 431-data page, as well as in the material I've provided for this Quiz. Note that we've also discussed the variables included in this file as part of Chapter 10 of the Course Notes, as well as in other parts of the course.

To begin, remove all subjects with missing information on the `plank_time` variable, which should yield a remaining group of 1384 subjects. Use only those 1384 subjects in answering this question. The variables used in this question include:

Variable	Description
`plank_time`	# of seconds plank position is held
`age_child`	child's age at screening (years)
`arm_circ`	arm circumference (cm)
`asthma_ever`	Have you ever been told you have asthma?
`bmi`	body-mass index (kg/m^2)
`energy`	energy consumed yesterday (kcal)
`fat`	total fat consumed yesterday (g)
`height`	standing height (cm)
`phys_health`	general health condition (Excellent - Poor)
`protein`	total protein consumed yesterday (g)
`race_eth`	race/hispanic ethnicity (4 levels)
`sex`	sex (Female or Male)
`sugar`	total sugar consumed yesterday (g)
`waist`	waist circumference (cm)
`water`	total plain water drank yesterday (g)
`weight`	weight in kg

Which of the following five sets of predictors for `plank_time` using a regression model has a substantial problem with collinearity? (CHECK EACH OF THE CORRECT RESPONSES.)

a. the child's age, sex, waist circumference, and height
b. the child's age, sex, body-mass index and race/ethnicity
c. the child's race/ethnicity and yesterday's consumption of energy, protein, and sugar
d. the child's sex, arm circumference, waist circumference and plain water consumption
e. the child's race/ethnicity, fat consumption, physical health, and whether they've been told they have asthma
f. None of the five models listed above has a substantial problem

6 Question 6

In Question 6, we will again use the `nnyfs.Rds` data (as we did in Question 5), but now we will leave in subjects with missing values of `plank_time` and instead build a new sample to include only the 1481 subjects who have a value of "1_Excellent", "2_VeryGood", or "3_Good" for their physical health (`phys_health`.)

The sample means of `age_child` and the sample sizes for each `phys_health` group used in Question 6 are shown in the table below.

group	phys_health	n	mean(age_child)
Excellent	1_Excellent	742	8.932615
Very Good	2_VeryGood	424	8.893868
Good	3_Good	315	9.473016

Note that I label the groups Excellent, Very Good and Good in the responses below. I'm not trying to be tricky.

- The Excellent group is made up of the subjects with `physhealth` = `1_Excellent`,
- The Very Good group is the set of subjects with `physhealth` = `2_VeryGood`, and
- The Good group is the set of subjects with `physhealth` = `3_Good`.

Now suppose we want to compare the mean age at screening (`age_child`) across the three remaining levels of physical health using a linear model. Perform a Tukey HSD comparison of the three group means using a 90% family-wide confidence level and use it to decide which pairs of physical health groups are detectably different in terms of mean age at screening.

a. None of the groups is detectably older than any of the others
b. Good is detectably older than Very Good, but there are no other detectable differences
c. Good is detectably older than Excellent, but there are no other detectable differences
d. Excellent is detectably older than Very Good, but there are no other detectable differences
e. Good is detectably older than both Very Good and Excellent, but there are no other detectable differences
f. Good and Excellent are each detectably older than Very Good, but there are no other detectable differences
g. Good is detectably older than Excellent, and Excellent is detectably older than Very Good, but there are no other detectable differences
h. Good is detectably older than Excellent, and both are detectably older than Very Good.
i. None of these options is correct

7 Question 7

Return once again to the original `nnyfs.Rds` data, and this time we will consider all subjects with complete data on these two variables:

- `physical_last_week`: Did the subject have any physical activity outside of school in the past week?
- `med_use`: Did the subject take a prescription medication in the past month?

Develop an appropriate summary of the data, and then use it to obtain a point estimate and 90% confidence interval for the relative risk of taking a prescription medication in the past month comparing those who did engage in some physical activity last week to those who did not engage in such an activity. Round each of your responses to three decimal places.

a. The point estimate is …
b. The 90% confidence interval is …

Note that your answer to part b should be of the form (0.123, 4.567)

8 Question 8

I have generated four figures (shown on the next two pages) to compare the films in our favorite movies data, a piece of which I've gathered in the `movies.csv` file made available to you. The `bw_rating` is the number (0-3) of Bechdel-Wallace criteria met by the movie, and the `us_pct` is the % of 10-star public ratings in IMDB as of 2023-09 (among US reviewers.)

Note that the Bechdel-Wallace criteria are:

- movie has to have at least two women in it
- who talk to each other
- about something besides a man

See https://bechdeltest.com/ for more details.

Here are some numerical summaries of the data...

```
movies <- read_csv("data/movies.csv", show_col_types = FALSE)

movies_q08 <- movies |>
  select(film_id, film, us_pct, bw_rating) |>
  drop_na() |>
  mutate(bw_rating = factor(bw_rating))

favstats(us_pct ~ bw_rating, data = movies_q08) |>
  gt()
```

bw_rating	min	Q1	median	Q3	max	mean	sd	n	missing
0	17	20.75	23	26.0	62	25.56250	10.301901	16	0
1	12	22.00	24	26.0	62	25.22222	7.810652	54	0
2	17	22.50	25	29.5	53	28.36364	10.763575	11	0
3	11	21.00	24	29.0	59	26.08036	7.960645	112	0

Figures 1 - 4 for Question 8 are shown on the next two pages.

Figure 1 for Question 8

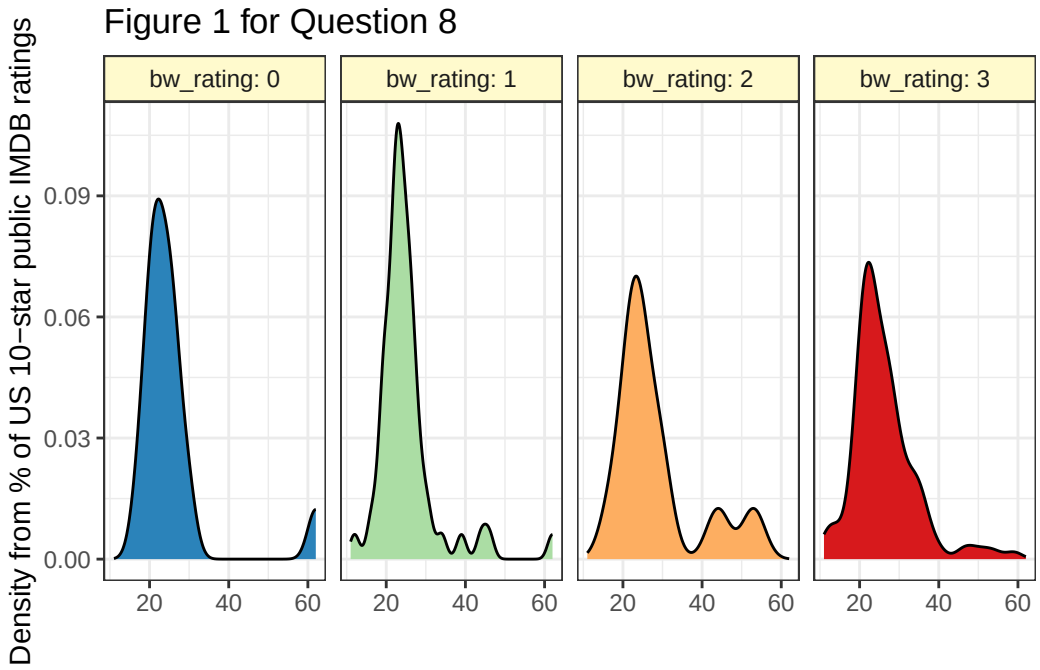
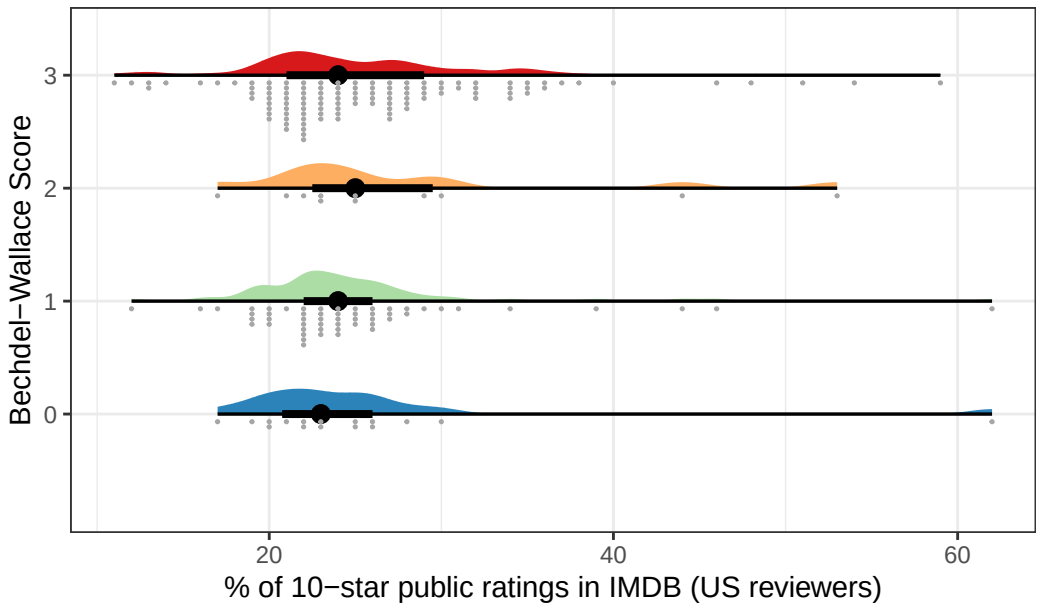


Figure 2 for Question 8



Question 8 continues on the next two pages.

Figure 3 for Question 8

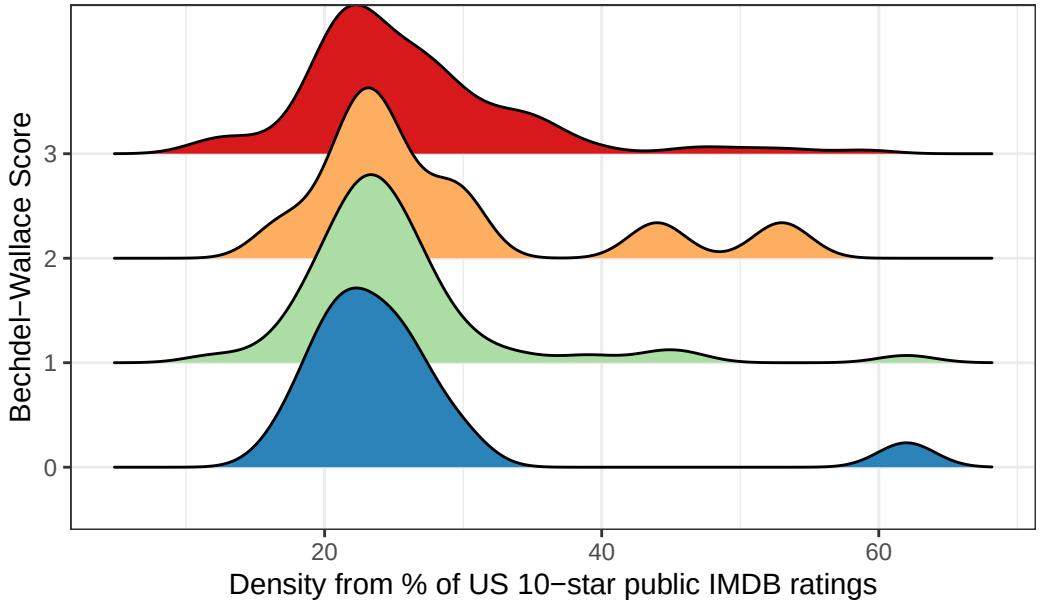
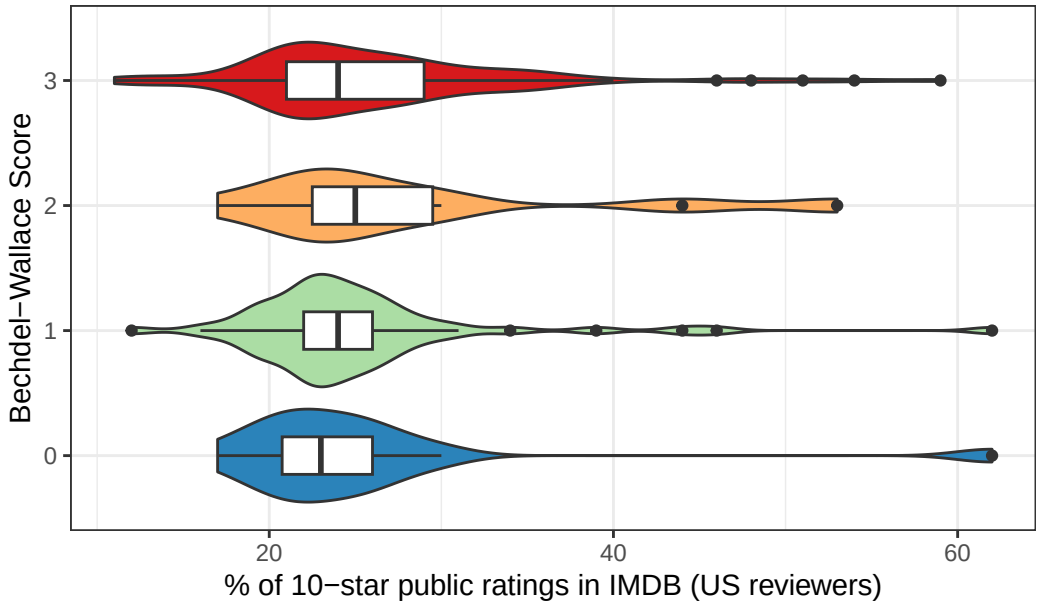


Figure 4 for Question 8



Question 8 continues on the next page.

Question 8 (continued)

Identify the four figures by the type of plot they display.

Columns:

a. Boxplot and Violin
b. Faceted Density Plot
c. Raindrop Plot
d. Ridgeline Plot
e. None of these

Rows:

1. Figure 1
2. Figure 2
3. Figure 3
4. Figure 4

9 Question 9

Here is the result of running an ANOVA and a Kruskal-Wallis comparison of the `us_pct` results for the four `bw_rating` groups shown in Question 8.

```
anova(lm(us_pct ~ bw_rating, data = movies_q08))
```

```
Analysis of Variance Table

Response: us_pct
           Df  Sum Sq Mean Sq F value Pr(>F)
bw_rating   3    96.9  32.297  0.4689 0.7043
Residuals 189 13018.1  68.879
```

```
kruskal.test(us_pct ~ bw_rating, data = movies_q08)
```

```
    Kruskal-Wallis rank sum test

data:  us_pct by bw_rating
Kruskal-Wallis chi-squared = 1.7689, df = 3, p-value = 0.6217
```

Which of the following statements are true? (CHECK ALL OF THE STATEMENTS THAT ARE TRUE.)

a. A linear model predicting `us_pct` using `bw_rating` accounts for $32.297/68.879 = 46.9\%$ of the variation.
b. The ANOVA and the Kruskal-Wallis results give different conclusions at typical significance levels.
c. A Bonferroni-Holm comparison would find `bw_rating` group 1's mean `us_pct` to be detectably larger than that of `bw_rating` group 3 with 90% confidence.
d. None of the other three statements are true.

10 Question 10

Consider these four scenarios. Which of these involve paired samples, and which involve independent samples?

10.1 Scenario 1 for Question 10

The southern white rhinoceros (*Ceratotherium simum simum*) is the most abundant rhino species in the world. Your outcome of interest is a measure of size. Suppose that you want to compare those living in an area of southern Africa subject to serious problems from poaching (you have data on 40 rhinos living near Watering Hole A) and those living in an area of southern Africa more than 1,000 km away with a less serious poaching problem (you have data on 40 rhinos living near Watering Hole B). Your interest is to understand how exposure to poaching is associated with average rhino size.

10.2 Scenario 2 for Question 10

A dose of one of two soporific drugs (dextro or laevo) were administered to 30 patients with trouble sleeping, with the drug selected at random for each patient. A soporific drug tends to induce drowsiness or sleep. The number of additional hours of sleep each drug provided (as compared to the night before, when no soporific drugs were used) was recorded. A week later, the 30 patients returned to the sleep lab and received the other drug (the one they didn't receive initially) and again, the number of additional hours of sleep each drug provided was recorded. You want to compare the mean improvement under dextro to the mean improvement under laevo.

10.3 Scenario 3 for Question 10

You are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the subject's ear). Sixteen readings (measured in degrees Fahrenheit) - eight per ear - were taken on a healthy subject at intervals of one minute, alternating ears. Your goal is to understand the difference in the means of the two ears.

10.4 Scenario 4 for Question 10

We include all asthma patients satisfying our inclusion criteria presenting for care over a period of time, and record the number of acute care visits for each patient during year 1. Then we provide them a standardized course of asthma training and record each patient's number of acute care visits during year 2. We want to understand whether the training increases or decreases the average number of acute care visits.

11 Question 11

The data available in the `projA.xlsx` file provided with the quiz contains, among other things, a variable called `lines`, which is the number of Quarto lines of code included in the final submission for 42 projects submitted in Fall 2023. We're going to create three confidence intervals based on this sample, which has sample mean = 768.4 lines and sample median = 703 lines of code.

a. Option A is to set a seed to 431, then obtain a 95% confidence interval for the population *mean* number of lines of code, using the bootstrap percentile estimate provided by the Hmisc package's `smean.cl.boot()` function, with 2000 bootstrap replications.

b. Option B is to obtain a 95% confidence interval for the population *mean* number of lines of code, using an ordinary least squares linear model and the `tidy()` function.

c. Option C is to again set a seed to 431, and then obtain a 95% confidence interval for the population *median* number of lines of code, this time using a bootstrap percentile interval obtained through the `boot.ci()` function.

Columns: Option A, Option B, Option C

Rows:

1. Which is the best choice if the data were randomly sampled from a population that is described well by the Normal distribution?
2. Which produces the confidence interval with the smallest width?
3. Which produces the largest upper bound for its confidence interval?
4. Which is symmetric around its point estimate?

12 Question 12

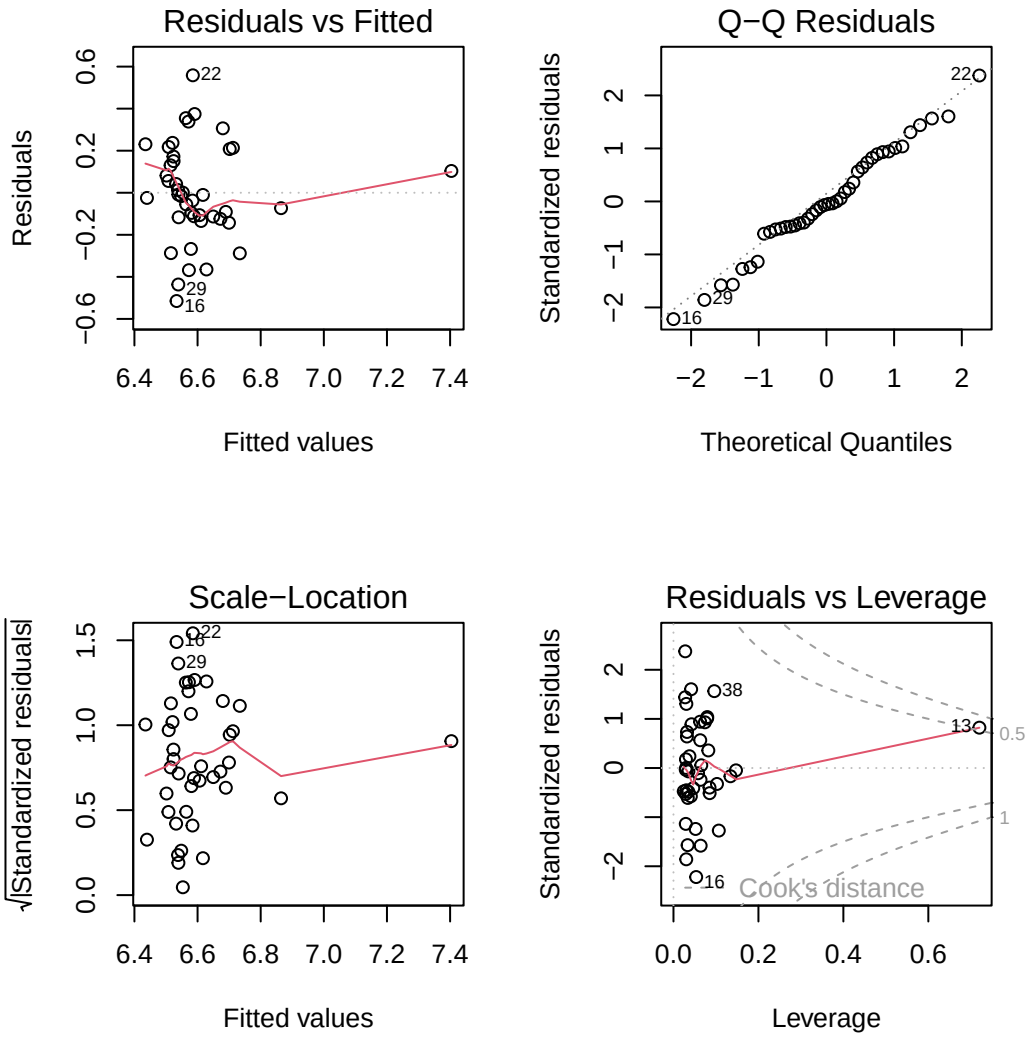
Consider again the data from Question 11, but here we are developing a model to predict the number of `lines` of code using two variables: the number of `counties` sampled by the student, and the number of `characters` used in writing the Project A reflection statement.

Which of the following transformations of `lines` does a Box-Cox approach suggest will be the best option for building such a model?

a. No transformation
b. Squaring the `lines`
c. The inverse of `lines`
d. The natural logarithm of `lines`
e. The square root of `lines`
f. The squared inverse of `lines`

13 Question 13

The four residual plots and the `outlierTest()` output shown below were generated using a regression model for the data used in Questions 11 and 12.




```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
22 2.535356           0.015469      0.64971
```

Question 13 (continued)

What is the principal problem with regression assumptions identified by the output we have provided?

a. A problem with the assumption of independence.
b. A problem with the assumption of linearity.
c. A problem with the assumption of Normality.
d. A problem with the assumption of constant variance.
e. A problem with collinearity.
f. A highly influential point.
g. There are no substantial problems with assumptions.

14 Question 14

We are comparing the length of stay in the emergency room and hospital (combined) across four hospitals for patients who present for care with a single broken bone in the arm (either a break in the upper arm bone - the *humerus* or in one of the two forearm bones - the *ulna* and the *radius*) as a result of a fall onto an outstretched hand.

The data on type of break and length of stay (split into three non-overlapping categories) were gathered over several recent years and aggregated across the four hospitals to produce the summary below.

Length of Stay	$<$ 6 hours	6-18 hours	$>$ 18 hours	ALL
Ulna	25	25	15	65
Radius	10	30	25	65
Humerus	10	15	20	45
ALL	45	70	60	175

```
Table14 <- matrix(c(25, 25, 15, 10, 30, 25, 10, 15, 20),
                  ncol= 3, nrow = 3, byrow = TRUE)
rownames(Table14) <- c("Ulna", "Radius", "Humerus")
colnames(Table14) <- c("LT6", "6-18", "GT18")
Table14
```

```
        LT6 6-18 GT18
Ulna     25   25   15
Radius   10   30   25
Humerus  10   15   20
```

Question 14 (continued)

```
chisq.test(Table14)
```

```
    Pearson's Chi-squared test

data:  Table14
X-squared = 12.239, df = 4, p-value = 0.01566
```

```
chisq.test(Table14)$expected
```

```
             LT6 6-18     GT18
Ulna    16.71429   26 22.28571
Radius  16.71429   26 22.28571
Humerus 11.57143   18 15.42857
```

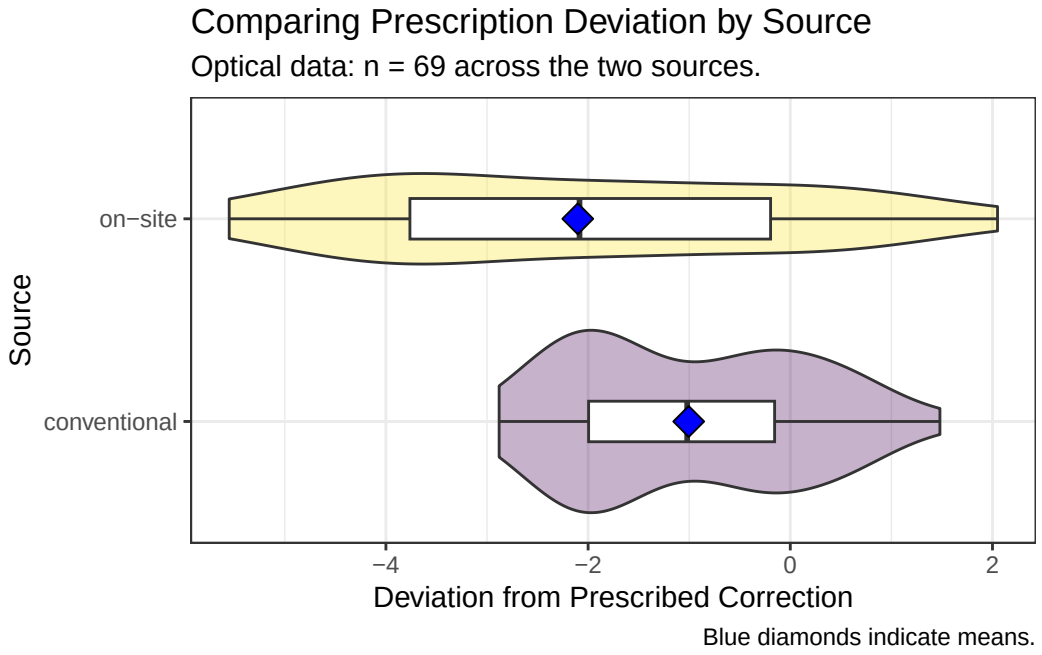
Which of the following statements is the most appropriate conclusion based on the data and output provided here?

a. There is a serious problem with assumptions here, which invalidates the test I would usually perform in these circumstances.
b. The length of stay and type of break are essentially independent, since the p value for the appropriate test is well above 0.10.
c. The length of stay and type of break are essentially independent, since the p value for the appropriate test is well below 0.10.
d. The length of stay and type of break are closely associated, since the p value for the appropriate test is well above 0.10.
e. The length of stay and type of break are closely associated, since the p value for the appropriate test is well below 0.10.
f. None of the statements above provides an appropriate conclusion.

15 Question 15

Recently, a number of opticians established on-side laboratories for preparing prescription eyeglasses. These labs provide more rapid service than conventional off-premises labs. Conventional opticians have questioned the accuracy of on-site labs. As a test, eyeglasses prescribed for nearsightedness were prepared by both types of labs. The glasses were then evaluated by very accurate devices that determine the percentage deviation from the prescribed correction. A minus sign indicates that the actual correction is less than prescribed; a plus sign, more than prescribed. The `optical.csv` data have been provided to you.

source	min	Q1	median	Q3	max	mean	sd	n	missing
conventional	−2.88	−2.00	−1.02	−0.15	1.48	−1.00	1.15	43	0
on-site	−5.55	−3.76	−2.08	−0.20	2.05	−2.10	2.12	26	0



Four analyses (labeled Analysis 1 - Analysis 4) using these data are presented on the next page.

15.1 Question 15 Analysis 1

```
tidy(lm(deviation ~ source, data = optical),
     conf.int = T, conf.level = 0.90) |>
  select(term, estimate, conf.low, conf.high) |> kbl(dig = 2)
```

term	estimate	conf.low	conf.high
(Intercept)	-1.0	-1.41	-0.60
sourceon-site	-1.1	-1.75	-0.44

15.2 Question 15 Analysis 2

```
tidy(wilcox.test(deviation ~ source, data = optical,
                 exact = F, conf.int = T, conf.level = 0.90)) |>
  select(estimate, conf.low, conf.high) |> kbl(digits = 2)
```

estimate	conf.low	conf.high
1.29	0.32	1.99

15.3 Question 15 Analysis 3

```
tidy(t.test(deviation ~ source, data = optical, conf.level = 0.90)) |>
  select(estimate1, estimate2, estimate, conf.low, conf.high, method) |>
  kbl(dig = 2)
```

estimate1	estimate2	estimate	conf.low	conf.high	method
-1	-2.1	1.1	0.34	1.86	Welch Two Sample t-test

15.4 Question 15 Analysis 4

```
set.seed(2023015)
bootdif(y = optical$deviation, g = optical$source,
        conf.level = 0.90, B.reps = 2000)
```

```
Mean Difference            0.05            0.95
     -1.0972719      -1.8186825      -0.3699978
```

Question 15 (continued)

Part 15a (2 points): Which of these 90% confidence intervals is our best choice for estimating the mean of the population on-site minus conventional difference in percentage deviation, in light of the information provided?

a. (0.60, 1.41)
b. (0.44, 1.75)
c. (0.32, 1.99)
d. (0.34, 1.86)
e. (0.37, 1.82)
f. None of the above.

Part 15b (2 points): In a sentence or two, specify the main reason why you chose the confidence interval you selected in Question 15a.

16 Question 16

You have completed a multiple regression analysis for an outcome, where you considered a set of six predictors. Your process included the following steps.

1. Create a single imputed data set to take care of missing values across the 2200 observations you have for the seven key variables (your outcome, which had no missing data, along with your six predictors, which each had some missing observations.)
2. Split the imputed data set into a model training sample of 1400 observations and a model test sample of the other 800 observations.
3. Fit four potential models in the training sample, and obtain the R^2 value and the adjusted R^2 value for each potential model.
4. Use the models you developed in step 3 to predict the outcome for the 800 observations in the test sample and calculate the square of the correlation between the predictions you made and the observed value of the outcome.
5. Select one of your four potential models, and then return to the initial data set and multiply impute 100 times, using the mice package. Run your selected set of predictors for the outcome, and pool the results across the 100 iterations to obtain appropriate estimates of each coefficient in the model and its standard error.
6. Then obtain the pooled R^2 value and pooled adjusted R^2 value for your selected model after completing step 5.

Question 16 (continued)

Note that for the set of predictors you eventually chose, you obtained the following results:

Step	Description	R^2 Value
3	training sample raw R^2	0.355
3	training sample adjusted R^2	0.344
4	squared correlation in test sample	0.333
6	pooled R^2 value	0.322
6	pooled adjusted R^2 value	0.311

Which of the five R^2 values summarized above would be the best choice to describe the proportion of variation in your outcome that your selected model can explain in new data, under the assumption that missing data is MAR, and accounting for missingness using multiple imputation?

a. 0.355
b. 0.344
c. 0.333
d. 0.322
e. 0.311
f. None of these.

17 Question 17

Below (and continuing on the next page) are several pieces of output, describing two different fits of a model to the same outcome and predictors. Note that you **do not** have access to the `q17` data.

```
m17a <- lm(y ~ x1 + x2 + x3, data = q17)

tidy(m17a, conf.int = T, conf.level = 0.9) |>
  select(term, estimate, conf.low, conf.high) |> gt() |>
  fmt_number(decimals = 3, columns = -c(term))
```

term	estimate	conf.low	conf.high
(Intercept)	62.654	58.818	66.489
x1	0.712	0.567	0.857
x2	−1.248	−1.431	−1.066
x3	5.013	3.556	6.470

```
glance(m17a) |> select(sigma, AIC, BIC, nobs) |> gt()
```

sigma	AIC	BIC	nobs
6.133585	1299.047	1315.538	200

```
m17b <- rlm(y ~ x1 + x2 + x3, data = q17)

tidy(m17b, conf.int = T, conf.level = 0.9) |>
  select(term, estimate, conf.low, conf.high) |> gt() |>
  fmt_number(decimals = 3, columns = -c(term))
```

term	estimate	conf.low	conf.high
(Intercept)	42.893	39.736	46.050
x1	1.436	1.316	1.555
x2	−1.030	−1.180	−0.880
x3	5.159	3.960	6.358

Question 17 (continued)

```
glance(m17b) |> select(sigma, AIC, BIC, nobs) |> gt()
```

sigma	AIC	BIC	nobs
5.297472	1375.286	1391.778	200

In looking at the residual plots for `m17a` (not shown here), there appear to be several substantial outliers.

1. In a couple of sentences, what does the output provided suggest about the impact of these outliers on the modeling for this outcome?
2. Which model (`m17a` or `m17b`) produces a larger prediction of the outcome for a new subject with `x1` = 25, `x2` = 10 and `x3` = 1?

18 Question 18

Questions 18 and 19 use data from the `nh_adult750.Rds` data that is found on our 431-data page, as well as in the material I've provided for this Quiz. The variables we are studying here are:

Variable	Description
ID	Same as SEQN, just a subject identifying code
Pulse	Pulse rate in beats per minute
HealthGen	Self-reported overall health (Excellent - Poor)
SleepTrouble	Do you have trouble sleeping? (Yes or No)
BMI	Body-mass index

Consider the following output.

```
nh_adult750 <-
  read_rds("https://github.com/THOMASELOVE/431-data/raw/main/data-and-code/nh_adult750.Rds

q18 <- nh_adult750 |>
  select(ID, Pulse, HealthGen, SleepTrouble, BMI)

miss_var_summary(q18)
```

```
# A tibble: 5 x 3
  variable     n_miss pct_miss
  <chr>         <int>    <dbl>
1 HealthGen        99   13.2
2 Pulse            32    4.27
3 BMI               5    0.667
4 ID                0    0
5 SleepTrouble      0    0
```

```
mcar_test(q18)
```

```
# A tibble: 1 x 4
  statistic    df p.value missing.patterns
      <dbl> <dbl>   <dbl>            <int>
1      16.6    23   0.830                8
```

In a sentence or two, what does the `mcar_test()` result tell us?

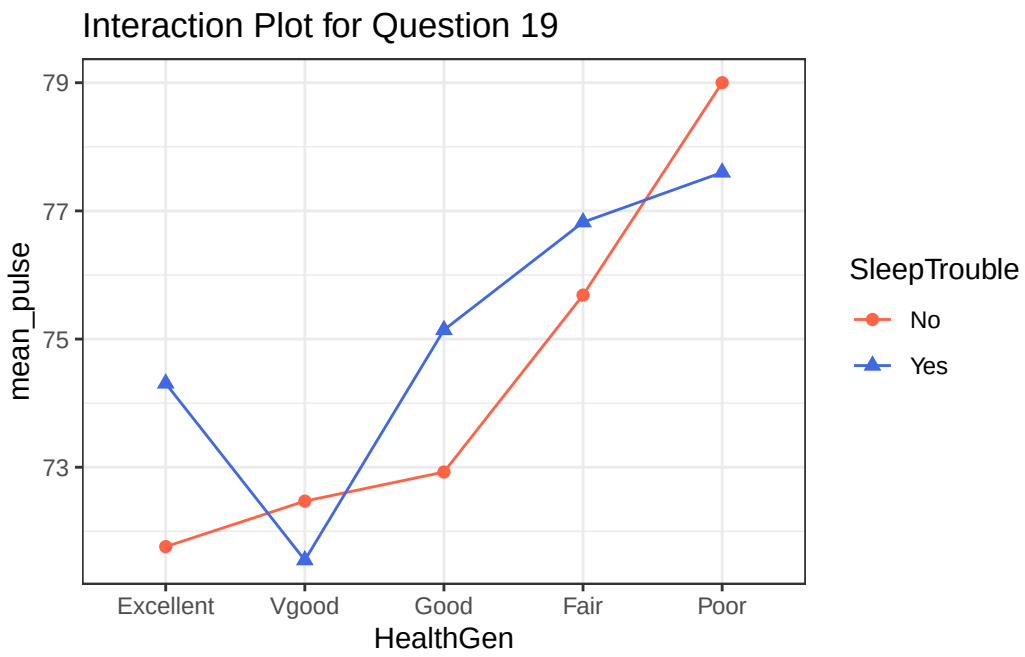
19 Question 19

Consider the output below, which builds on the work we did in Question 18.

```
q19_imp <- q18 |>
  impute_cart(HealthGen ~ SleepTrouble) |>
  impute_pmm(BMI ~ HealthGen + SleepTrouble) |>
  impute_median(Pulse ~ SleepTrouble + HealthGen)

q19_summary <- q19_imp |>
  group_by(HealthGen, SleepTrouble) |>
  summarise(mean_pulse = mean(Pulse, na.rm = TRUE))

ggplot(q19_summary, aes(x = HealthGen, y = mean_pulse)) +
  geom_line(aes(group = SleepTrouble, color = SleepTrouble)) +
  geom_point(aes(pch = SleepTrouble, color = SleepTrouble),
             size = 2) +
  scale_color_manual(values = c("tomato", "royalblue")) +
  labs(title = "Interaction Plot for Question 19")
```



Question 19 continues on the next page.

Question 19 (continued)

```r
m19a <- lm(Pulse ~ HealthGen * SleepTrouble, data = q19_imp)
m19b <- lm(Pulse ~ HealthGen + SleepTrouble, data = q19_imp)

anova(m19a)
```

```
Analysis of Variance Table

Response: Pulse
                       Df Sum Sq Mean Sq F value  Pr(>F)
HealthGen               4   1408  351.91  2.7284 0.02832 *
SleepTrouble            1    226  225.83  1.7509 0.18618
HealthGen:SleepTrouble  4    255   63.81  0.4947 0.73963
Residuals             740  95445  128.98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m19b)
```

```
Analysis of Variance Table

Response: Pulse
              Df Sum Sq Mean Sq F value  Pr(>F)
HealthGen      4   1408  351.91  2.7358 0.02797 *
SleepTrouble   1    226  225.83  1.7556 0.18558
Residuals    744  95700  128.63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indicate whether each of the four statements below are true, false or we cannot tell from the information provided.

Rows (Statements) are:

1. The plot shows no interaction of `HealthGen` and `SleepTrouble` on the mean of `Pulse`.
2. The interaction term explains less than 1% of the variation in `Pulse` rates.
3. These analyses assume that missingness is completely at random.
4. A simple regression of `Pulse` on `SleepTrouble` would have a p value below 0.10.

Columns: a. True b. False c. We cannot tell.

20 Question 20

In the first 79 games of Wordle that Dr. Love played on his phone, he managed to win (guess the word in 6 or fewer letters) 77 times. Since then, he has won 115 games in a row, as of 2023-11-30.

Assume that his rate of winning through the 194 games he has played on his phone is a random sample from the process of interest, even though there may be problems with the assumption of independence.

Write a single line of code to use the "plus4" method to estimate the probability (and a 95% confidence interval around that probability) that Dr. Love will win the Wordle when he plays it on his phone on 2023-12-05.

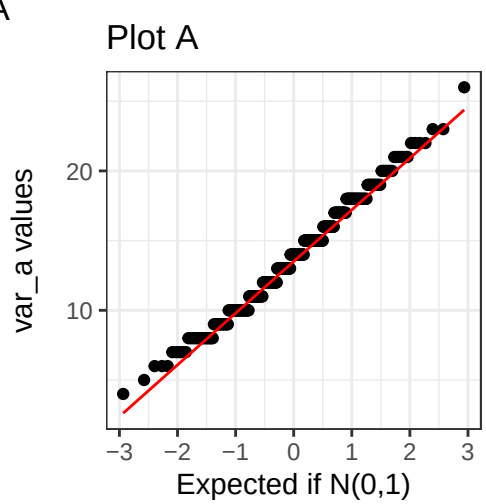
21 Question 21

Which of the four Normal Q-Q plots (A, B, C or D) shown on the next page displays data from a substantially right-skewed distribution? (CHECK ALL PLOTS THAT SHOW MEANINGFULLY RIGHT-SKEWED DATA.)

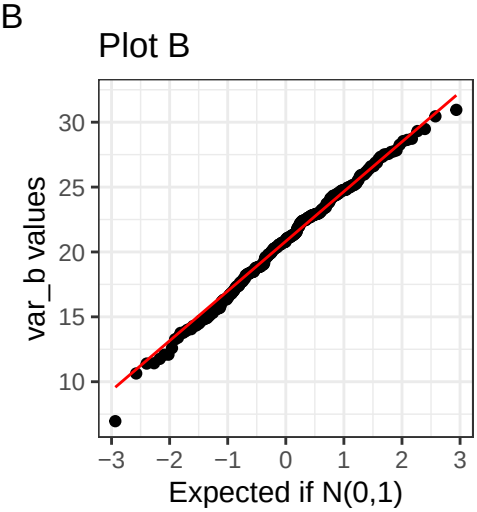
a. Plot A
b. Plot B
c. Plot C
d. Plot D
e. None of these plots

Plots for Question 21

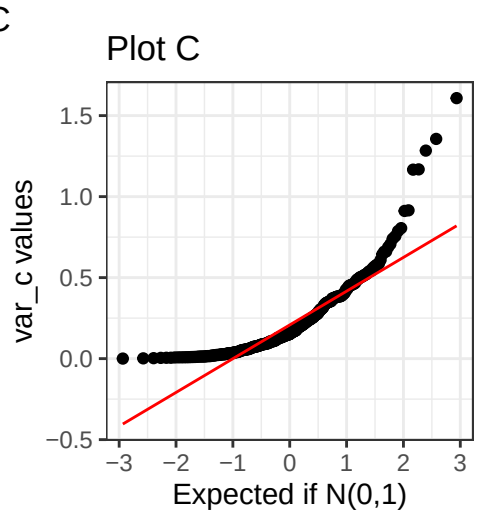
A



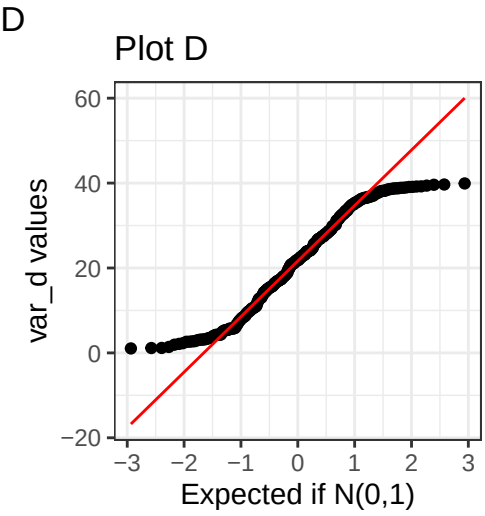
B



C



D



22 Question 22

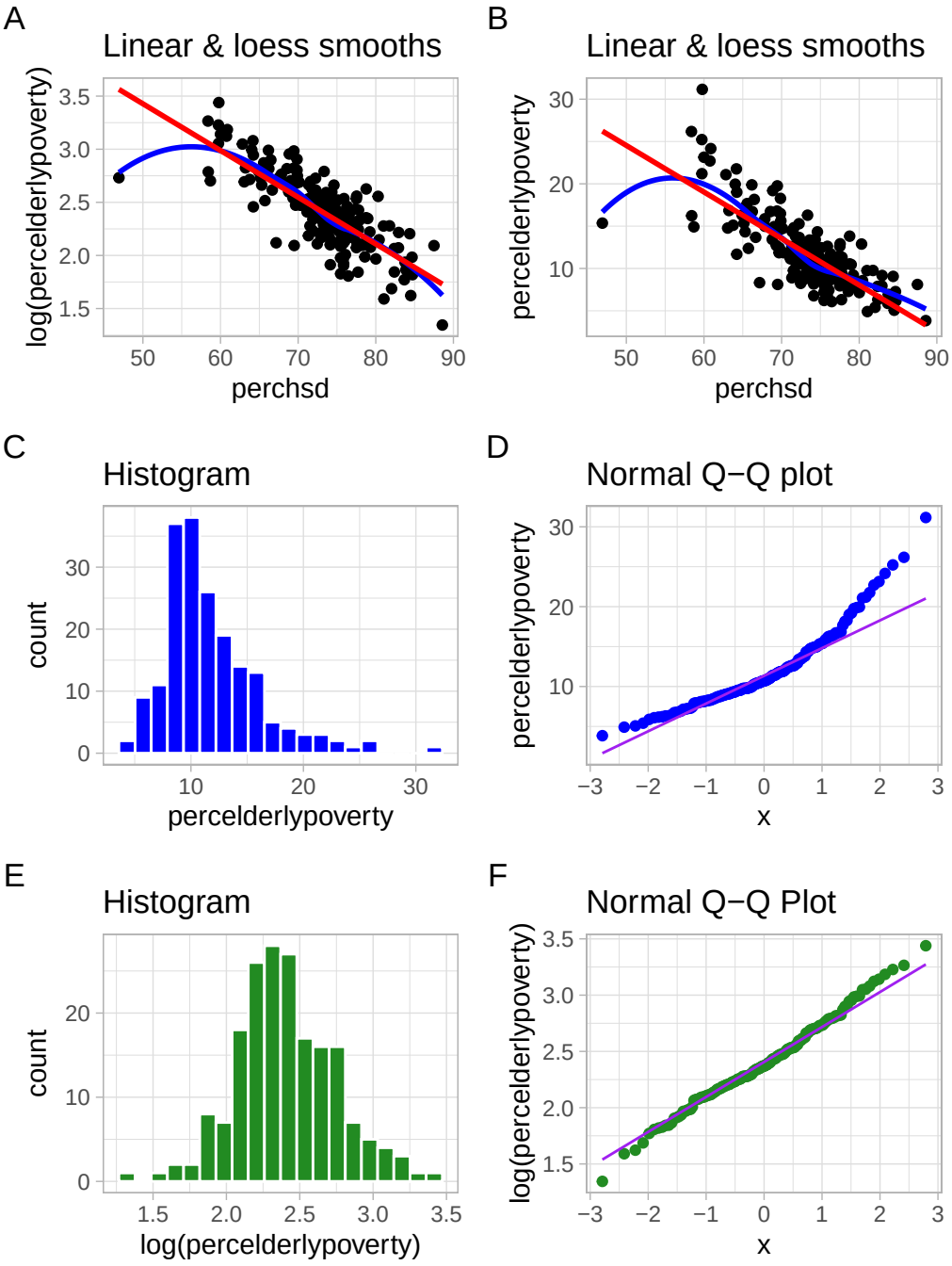
Suppose you are using a subset of the `midwest` data from the `ggplot2` package. You are trying to determine for this subset whether or not a transformation of the outcome (specifically, taking the natural logarithm of the outcome) is necessary to fit a linear regression model to describe the relationship between `perchsd` (the predictor, specifically the percent with a high school diploma) and `percelderlypoverty` (the outcome, specifically the percent of adults ages 65 and over below the poverty level). Which of the Plots shown in the Figure for Question 22 would be of the most help in assessing whether using this transformation would improve the assumption of linearity?

a. Plot A
b. Plot B
c. Plots C and D
d. Plots E and F
e. They would all be equally useful.

The Figure for Question 22 is shown on the next page.

Question 22 (continued)

Figure for Question 22



23 Question 23

Suppose we are considering six potential sets of predictors in regression models for the same outcome, and obtain the following results in our test sample. We will assume all six models show comparable performance and adherence to regression assumptions in the model development sample.

Model	RMSPE	Mean APE	Maximum APE	Validated R^2
1	10	8	15	0.56
2	12	9	13	0.59
3	14	10	14	0.57
4	9	12	10	0.60
5	11	10	16	0.61
6	7	9	14	0.63

Which of these models are dominated across these summaries by other models, so that we should no longer consider them in light of this analysis? (CHECK ALL OF THE MODELS THAT ARE DOMINATED BY OTHER, BETTER MODELS.)

a. Model 1
b. Model 2
c. Model 3
d. Model 4
e. Model 5
f. Model 6
g. None of these models are dominated.

24 Question 24 (6 points)

Write a clear and well-composed essay of 150 to 300 words describing an important idea from David Spiegelhalter's *The Art of Statistics* about doing statistical science well that Dr. Love **didn't cover** in Classes 1-23. Your essay should state the idea in your own words, and should indicate why you feel it is important.

If you quote Spiegelhalter (and we prefer that you do), specify the Chapter containing your quote. If Dr. Love discussed your idea in class, you'll lose 1 of 6 available points. If your essay is unclear, or if you miss Spiegelhalter's point, that will have a bigger impact on your score. Each Chapter in Spiegelhalter includes a summary of key points. Feel free to use these summaries to help spark ideas, but do not quote the summaries.

These instructions are 137 words long.

This is the end of the Quiz.

Be sure to complete the Affirmation at the end of the Answer Sheet, and that you have submitted your Answer Sheet, and received your copy in your CWRU email by the deadline.

Session Information

```
session_info()
```

```
R version 4.3.1 (2023-06-16 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22621)


Locale:
  LC_COLLATE=English_United States.utf8
  LC_CTYPE=English_United States.utf8
  LC_MONETARY=English_United States.utf8
  LC_NUMERIC=C
  LC_TIME=English_United States.utf8

time zone: America/New_York
tzcode source: internal

Package version:
  abind_1.4-5             askpass_1.2.0           backports_1.4.1
  base64enc_0.1-3         bigD_0.2.0              bit_4.0.5
  bit64_4.0.5             bitops_1.0.7            blob_1.2.4
  boot_1.3-28.1           brio_1.1.3              broom_1.0.5
  bslib_0.5.1             ca_0.71.1               cachem_1.0.8
  callr_3.7.3             car_3.1-2               carData_3.0-5
  cellranger_1.1.0        checkmate_2.3.0         class_7.3.22
  cli_3.6.1               clipr_0.8.0             cluster_2.1.4
  cmprsk_2.2-11           codetools_0.2-19        colorspace_2.1-0
  commonmark_1.9.0        compiler_4.3.1          conflicted_1.2.0
  corrr_0.4.4             cpp11_0.4.6             crayon_1.5.2
  curl_5.1.0              data.table_1.14.8       DBI_1.1.3
  dbplyr_2.4.0            DEoptimR_1.1.3          desc_1.4.2
  diffobj_0.3.5           digest_0.6.33           distributional_0.3.2
  doRNG_1.8.6             dplyr_1.1.3             dtplyr_1.3.1
```

e1071_1.7.13 ellipsis_0.3.2 Epi_2.47.1
etm_1.1.1 evaluate_0.22 fansi_1.0.5
farver_2.1.1 fastmap_1.1.1 fontawesome_0.5.2
forcats_1.0.0 foreach_1.5.2 foreign_0.8-84
Formula_1.2-5 fs_1.6.3 gargle_1.5.2
gclus_1.3.2 generics_0.1.3 GGally_2.1.2
ggdist_3.3.0 ggforce_0.4.1 ggformula_0.10.4
ggplot2_3.4.4 ggrepel_0.9.4 ggridges_0.5.4
ggstance_0.3.6 glmnet_4.1-8 glue_1.6.2
googledrive_2.1.1 googlesheets4_1.1.1 gower_1.0.1
graphics_4.3.1 grDevices_4.3.1 grid_4.3.1
gridExtra_2.3 gt_0.10.0 gtable_0.3.4
haven_2.5.3 highr_0.10 Hmisc_5.1-1
hms_1.1.3 htmlTable_2.4.1 htmltools_0.5.6.1
htmlwidgets_1.6.2 httr_1.4.7 ids_1.0.1
isoband_0.2.7 iterators_1.0.14 itertools_0.1.3
janitor_2.2.0 jomo_2.7-6 jquerylib_0.1.4
jsonlite_1.8.7 juicyjuice_0.1.0 kableExtra_1.3.4
knitr_1.44 labeling_0.4.3 labelled_2.12.0
laeken_0.5.2 lattice_0.21-8 lifecycle_1.0.3
lme4_1.1-34 lmtest_0.9-40 lubridate_1.9.3
magrittr_2.0.3 markdown_1.11 MASS_7.3-60
Matrix_1.6-1.1 MatrixModels_0.5.2 memoise_2.0.1
methods_4.3.1 mgcv_1.8-42 mice_3.16.0
mime_0.12 minqa_1.2.6 missForest_1.5
mitml_0.4-5 modelr_0.1.11 mosaic_1.8.4.2
mosaicCore_0.9.2.1 mosaicData_0.20.3 munsell_0.5.0
naniar_1.0.0 nlme_3.1-162 nloptr_2.0.3
nnet_7.3-19 norm_1.0-11.1 numDeriv_2016.8-1.1
openssl_2.1.1 ordinal_2022.11.16 pan_1.9
parallel_4.3.1 patchwork_1.1.3 pbkrtest_0.5.2
permute_0.9.7 pillar_1.9.0 pkgbuild_1.4.2
pkgconfig_2.0.3 pkgload_1.3.3 plyr_1.8.9
polyclip_1.10-6 praise_1.0.0 prettyunits_1.2.0
processx_3.8.2 progress_1.2.2 proxy_0.4.27
ps_1.7.5 purrr_1.0.2 pwr_1.3-0
qap_0.1.2 quadprog_1.5-8 quantreg_5.97
R6_2.5.1 ragg_1.2.6 randomForest_4.7.1.1
ranger_0.15.1 rappdirs_0.3.3 RColorBrewer_1.1-3
Rcpp_1.0.11 RcppArmadillo_0.12.6.4.0 RcppEigen_0.3.3.9.3
reactable_0.4.4 reactR_0.5.0 readr_2.1.4
readxl_1.4.3 registry_0.5.1 rematch_2.0.0
rematch2_2.1.2 reprex_2.0.2 reshape_0.8.9

rlang_1.1.1 rmarkdown_2.25 rngtools_1.5.2
robustbase_0.99.0 rpart_4.1.19 rprojroot_2.0.3
rstudioapi_0.15.0 rvest_1.0.3 sass_0.4.7
scales_1.2.1 selectr_0.4.2 seriation_1.5.1
shape_1.4.6 simputation_0.2.8 snakecase_0.11.1
sp_2.1.1 SparseM_1.81 splines_4.3.1
stats_4.3.1 stringi_1.7.12 stringr_1.5.0
survival_3.5-5 svglite_2.1.2 sys_3.4.2
systemfonts_1.0.5 testthat_3.2.0 textshaping_0.3.7
tibble_3.2.1 tidyr_1.3.0 tidyselect_1.2.0
tidyverse_2.0.0 timechange_0.2.0 tinytex_0.48
tools_4.3.1 TSP_1.2.4 tweenr_2.0.2
tzdb_0.4.0 ucminf_1.2.0 UpSetR_1.4.0
utf8_1.2.3 utils_4.3.1 uuid_1.1.1
V8_4.4.0 vcd_1.4-11 vctrs_0.6.4
vegan_2.6.4 VIM_6.2.2 viridis_0.6.4
viridisLite_0.4.2 visdat_0.6.0 vroom_1.6.4
waldo_0.5.1 webshot_0.5.5 withr_2.5.1
xfun_0.40 xml2_1.3.5 yaml_2.3.7
zoo_1.8-12