431 Quiz 2 for Fall 2024

Deadline: Wednesday 2024-12-04 at Noon

Thomas E. Love

2024-11-22

Table of contents

Instructions for Students

There are **28** questions on this Quiz and this PDF is **31** pages long. Be sure you have all **31** pages. It is to your advantage to answer all **28** 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 (not even 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/431-2024-quiz2-form

All of your answers must be submitted through the Google Form by the deadline (noon on December 4), without exception. The form will close at **12:15 PM** on that date, and no extensions will be made available, so do not wait until late in the morning to submit your work. We will only accept responses through the Google Form.

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 as often as you like without losing your progress.

0.2 The Data Sets

Along with the other materials, I have provided **6** data sets (called **nh_adult750.Rds**, **nnyfs.Rds**, **projA23.xlsx**, **surveyA.xlsx**, **beds1.csv**, and **beds2.csv**) that are mentioned in the Quiz. They may be helpful to you.

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 **Thomas dot Love at case dot edu**.

During the Quiz period (5 PM 2024-11-21 through 12:15 PM 2024-12-04) we will not answer questions about the Quiz through Campuswire or in TA office hours or in person. Instead, Dr. Love will respond to questions sent to the email address listed above. We promise to respond to all questions received before 10 AM on 2024-12-04 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 without asking for help.

0.5 Scoring and Timing

Two of the questions (questions 1 and 2) are worth **6** points each, while **10** of the questions are worth **4** points, and the remaining 16 are worth **3** points, adding to a total of 100 points. Available points are specified at the start of each question. The questions are not in any particular order, **except** that the two "essay" questions come first (Questions 1 and 2). Questions 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. I expect most students will take 4-8 hours, and some will take as little as 2 or as many as 10. It is not a good idea to spend a long time on any one question.

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, but you may use either.

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-431.R script

This doesn't mean you need to use all of these packages, or indeed, that I used all of them in building the Quiz and its answer sketch.

```
library(janitor)
library(broom); library(car)
library(Epi); library(GGally)
library(glue); library(gt)
library(infer); library(knitr)
library(mice); library(MKinfer)
library(naniar); library(patchwork)
library(readxl); library(rstanarm)
library(xfun)
library(easystats)
library(tidyverse)

source("data/Love-431.R")

theme_set(theme_lucid())
knitr::opts_chunk$set(comment = NA)
```

> **Tip**
>
> Statistics is a "getting the details right" business. Read each question carefully, to be sure you get the details right in your answer.

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.

As you'll see in the Google Form Answer Sheet, you have the option to either type in your Question 1 response directly into the Google Form, or upload a Word, PDF or Google Doc to the form containing your response to Question 1.

2 Question 2 (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 1-24. 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.

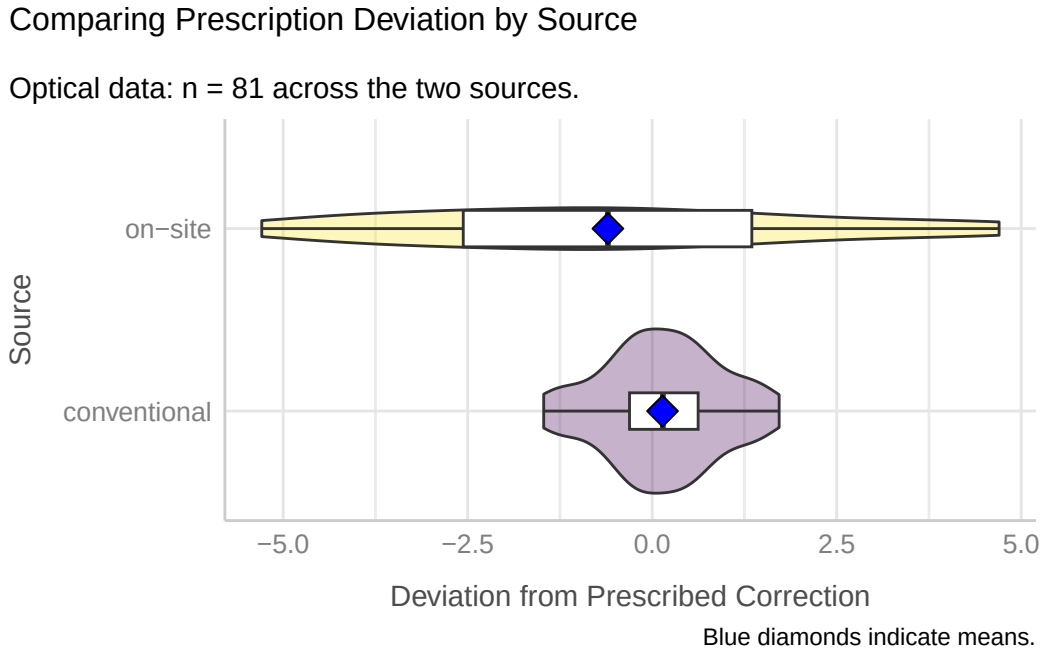
As you'll see in the Google Form Answer Sheet, you have the option to either type in your Question 1 response directly into the Google Form, or upload a Word, PDF or Google Doc to the form containing your response to Question 1.

3 Question 3 (4 points)

Recently, a number of opticians established on-site 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. In Question 3, your task is to use the output provided below, and on the next two pages, to specify the most appropriate point estimate and 90% confidence interval (from those in Analyses A-D) for the **conventional minus on-site** difference in percentage deviation. Then, in a sentence or two, specify the main reason why you chose the confidence interval from the Analysis (A-D) that you settled on.

```
optical |> group_by(source) |> reframe(lovedist(deviation))
```

```
# A tibble: 2 x 11
  source           n  miss  mean    sd    med   mad   min    q25   q75   max
  <chr>        <int> <int> <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
1 conventional    52     0  0.14 0.767  0.145 0.704 -1.47 -0.308 0.622  1.72
2 on-site         29     0 -0.6  2.78  -0.6   2.91  -5.29 -2.56  1.35   4.7
```



Analyses A-D using these data are shown on the next two pages.

3.1 Question 3 Analysis A

```
fit3a <- lm(deviation ~ source, data = optical)
model_parameters(fit3a, ci = 0.90)
```

```
Parameter        | Coefficient |   SE |         90% CI | t(79) |     p
-------------------------------------------------------------------------
(Intercept)      |        0.14 | 0.24 | [-0.27,  0.55] |  0.57 | 0.569
source [on-site] |       -0.74 | 0.41 | [-1.42, -0.06] | -1.81 | 0.074

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a Wald t-distribution approximation.
```

3.2 Question 3 Analysis B

```
fit3b <- t.test(deviation ~ source, var.equal = FALSE,
                conf.int = TRUE, conf.level = 0.90, data = optical)
model_parameters(fit3b, ci = 0.90)
```

```
Welch Two Sample t-test

Parameter |  Group | source = conventional | source = on-site | Difference |        90% CI |
-------------------------------------------------------------------------------------------
deviation | source |                  0.14 |            -0.60 |       0.74 | [-0.15, 1.63] |

Alternative hypothesis: true difference in means between group conventional and group on-site
```

3.3 Question 3 Analysis C

```
set.seed(4310031)
fit3c <- stan_glm(deviation ~ source, data = optical, refresh = 0)
model_parameters(fit3c, ci = 0.90)
```

```
Parameter     | Median |         90% CI |     pd |  Rhat |     ESS |                   Prior
-------------------------------------------------------------------------------------------
(Intercept)   |   0.14 | [-0.26,  0.54] | 71.23% | 1.001 | 3913.00 | Normal (-0.12 +- 4.47)
sourceon-site |  -0.72 | [-1.39, -0.08] | 96.60% | 1.001 | 3779.00 |  Normal (0.00 +- 9.27)
```

```
Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a MCMC distribution approximation.
```

3.4 Question 3 Analysis D

```r
set.seed(4310032)
optical |>
  specify(deviation ~ source) |>
  calculate(stat = "diff in means", order = c("conventional", "on-site"))
```

```
Response: deviation (numeric)
Explanatory: source (factor)
# A tibble: 1 x 1
   stat
  <dbl>
1  0.74
```

```r
optical |>
  specify(deviation ~ source) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("conventional", "on-site")) |>
  get_confidence_interval(level = 0.90, type = "percentile")
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1   -0.147     1.63
```

> **Note**
>
> This is the end of the output for Question 3.

4 Question 4 (3 points)

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	20	25	70
Radius	10	35	20	65
Humerus	10	15	20	45
ALL	45	70	65	180

```
Table4 <- matrix(c(25, 20, 25, 10, 35, 20, 10, 15, 20),
              ncol= 3, nrow = 3, byrow = TRUE)
rownames(Table4) <- c("Ulna", "Radius", "Humerus")
colnames(Table4) <- c("LT6", "6-18", "GT18")
Table4
```

```
        LT6 6-18 GT18
Ulna     25   20   25
Radius   10   35   20
Humerus  10   15   20
```

```
chisq.test(Table4)
```

```

    Pearson's Chi-squared test

data:  Table4
X-squared = 13.152, df = 4, p-value = 0.01056
```

```
chisq.test(Table4)$expected
```

```
          LT6     6-18     GT18
Ulna    17.50 27.22222 25.27778
Radius  16.25 25.27778 23.47222
Humerus 11.25 17.50000 16.25000
```

Which of the following statements is the most appropriate conclusion (assuming we want to use a 90% confidence level) 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.

5 Question 5 (3 points)

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 5 (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	0.455
3	training sample adjusted R^2	0.444
4	squared correlation in test sample	0.433
6	pooled R^2 value	0.422
6	pooled adjusted R^2 value	0.411

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. 0.455
b. 0.444
c. 0.433
d. 0.422
e. 0.411
f. None of these.

6 Question 6 (3 points)

Questions 6 and 7 use data from the `nh_adult750.Rds` data found 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.

```
q6 <- read_rds("data/nh_adult750.Rds") |>
  select(ID, Pulse, HealthGen, SleepTrouble, BMI)

miss_var_summary(q6)
```

```
# A tibble: 5 x 3
  variable     n_miss pct_miss
  <chr>         <int>    <num>
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(q6)
```

```
# 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?

7 Question 7 (4 points)

Consider the output on the next two pages, which builds on the work we did in Question 6.

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

Rows (Statements) are:

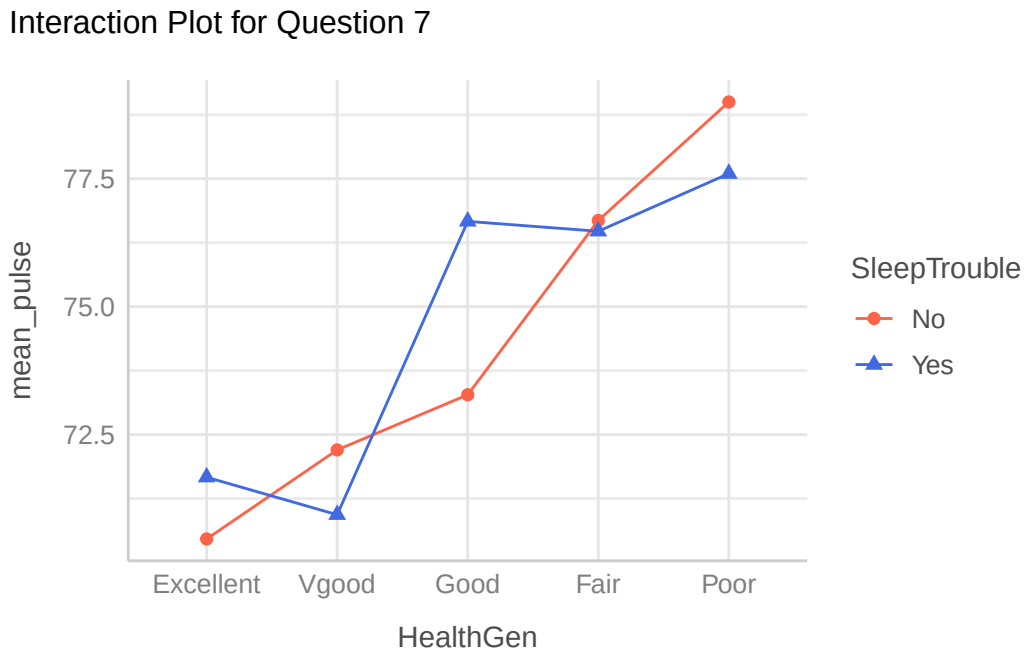
1. The plot shows no interaction of `HealthGen` and `SleepTrouble` on the mean of `Pulse`.
2. The interaction term explains less than 1% of the variation in `Pulse` rates.
3. These analyses assume that missingness is completely at random.
4. A simple regression of `Pulse` on `SleepTrouble` would have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

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

```r
q7 <- mice(q6, m = 1, seed = 4310007, print = FALSE) |>
  complete() |> tibble()

q7_summary <- q7 |>
  group_by(HealthGen, SleepTrouble) |>
  summarise(mean_pulse = mean(Pulse))

ggplot(q7_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 7")
```



```r
m7a <- lm(Pulse ~ HealthGen * SleepTrouble, data = q7)
anova(m7a)
```

```
Analysis of Variance Table

Response: Pulse
                        Df Sum Sq Mean Sq F value    Pr(>F)
HealthGen                4   2973  743.27  5.6975 0.0001611 ***
SleepTrouble             1    196  195.67  1.4999 0.2210835
HealthGen:SleepTrouble   4    575  143.69  1.1014 0.3547536
Residuals              740  96538  130.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m7b <- lm(Pulse ~ HealthGen + SleepTrouble, data = q7)
anova(m7b)
```

```
Analysis of Variance Table

Response: Pulse
              Df Sum Sq Mean Sq F value    Pr(>F)
HealthGen      4   2973  743.27  5.6944 0.0001619 ***
SleepTrouble   1    196  195.67  1.4990 0.2212071
Residuals    744  97113  130.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> **Note**
>
> This is the end of the output for Question 7.

8 Question 8 (3 points)

In a sentence or two, explain why the ANOVA for model `m7a` shown at the top of this page shows 4 in the `Df` column for the `HealthGen` variable, but only 1 in that column for the `SleepTrouble` variable.

9 Question 9 (4 points)

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

9.1 Scenario 1 for Question 9

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.

9.2 Scenario 2 for Question 9

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.

9.3 Scenario 3 for Question 9

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.

10 Question 10 (4 points)

Questions 10-15 use the `nnyfs.Rds` data set, which is found in the material I've provided for this Quiz.

For Question 10, 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 Question 10. The variables used in Question 10 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

11 Question 11 (4 points)

Return to the `nnyfs.Rds` data provided to you, and this time we will consider the 1514 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[1]. 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)

12 Question 12 (3 points)

Now return again to the original `nnyfs` data provided to you, and create a linear model, which we'll call `m12`, with ordinary least squares to predict `protein` consumption in grams using `phys_health` (general health condition) and the child's age (`age_child`) as predictors. The model should use all 1518 subjects in the `nnyfs` data. On fitting that model, you should obtain a point estimate of 13.02 for one of the coefficients related to the `phys_health` variable.

Provide a 2-3 sentence description of the meaning of the point estimate (13.02) in your model.

13 Question 13 (3 points)

In the model `m12` that you fit for Question 12, you should find that the coefficient of `age` is 1.84. Specify and provide a careful description of the meaning of the 90% confidence interval for that estimated coefficient of `age`.

[1] Your relative risk estimate should be built using the "did engage" group in its numerator, and should be built without any Bayesian augmentation.

14 Question 14 (4 points)

Now fit a model (which we'll call `m14`) that adds the `race_eth` (race-ethnicity, in four levels) variable to the predictors used in the model `m12` that you fit for Question 12.

Comparing the performance of models `m12` and `m14` on our sample of 1518 observations in the `nnyfs` data, for which of the following summaries do we get a better result with `m12` than we do with `m14`? (CHECK ALL THAT APPLY.)

a. Raw R-square
b. Adjusted R-square
c. RMSE
d. Sigma
e. AIC
f. Corrected AIC
g. BIC
h. None of these

15 Question 15 (3 points)

Run `check_model()` plots for your model `m14` fit in Question 14. Which of the following assumptions is problematic according to those plots? (SELECT ALL PROBLEMATIC ASSUMPTIONS)

a. Linearity
b. Constant Variance
c. Points with high influence
d. Collinearity
e. Non-Normality of residuals
f. None of these assumptions are problematic.

16 Question 16 (3 points)

The `Pottery` data are part of the `car` package in R. The data describe the chemical composition of ancient pottery found at four sites in Great Britain. I've built two data visualizations using these data, shown on the next page.

Visualization 1 describes a variable Dr. Love has labeled `var1`, and Visualization 2 describes a variable labeled `var2`. Note that the white dots in the boxplots within each visualization are placed on the sample mean.

Based on the output provided, **and** whatever other work you deem appropriate, identify each of the variables I've visualized.

Rows:

1. Variable `var1`.
2. Variable `var2`.

Columns:

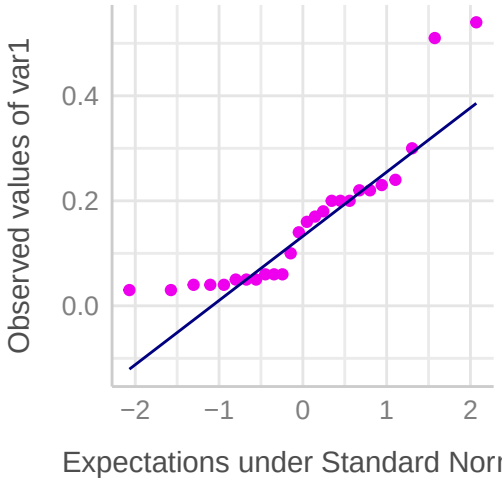
a. Calcium (`Ca`)
b. Iron (`Fe`)
c. Magnesium (`Mg`)
d. Sodium (`Na`)
e. Aluminum (`Al`).
f. We cannot determine.

```
Pottery |> head()
```

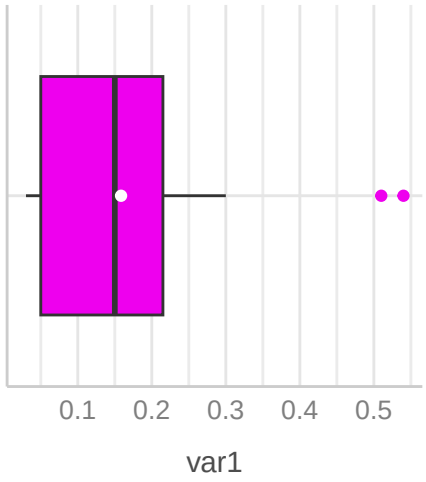
```
       Site   Al   Fe   Mg   Ca   Na
1 Llanedyrn 14.4 7.00 4.30 0.15 0.51
2 Llanedyrn 13.8 7.08 3.43 0.12 0.17
3 Llanedyrn 14.6 7.09 3.88 0.13 0.20
4 Llanedyrn 11.5 6.37 5.64 0.16 0.14
5 Llanedyrn 13.8 7.06 5.34 0.20 0.20
6 Llanedyrn 10.9 6.26 3.47 0.17 0.22
```

Visualization 1 for Question 16: var1

Normal Q−Q: var1

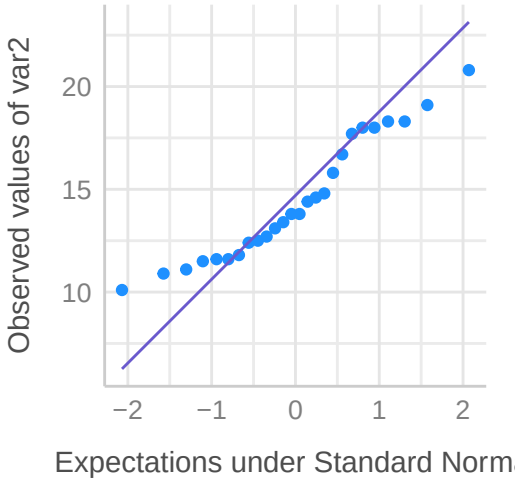


Boxplot with Mean: var1

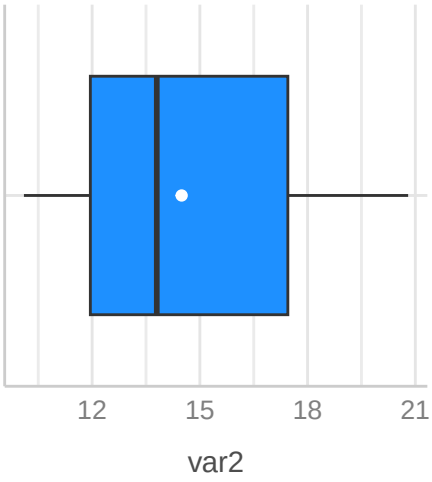


Visualization 2 for Question 16: var2

Normal Q−Q: var2



Boxplot with Mean: var2



Note

This is the end of the output for Question 16.

17 Question 17 (3 points)

Consider the responses to the item "How important do you think statistics will be in your future career?" as gathered in the `statfuture` variable within the `surveyA.xlsx` data set I have provided to you.

These data are drawn from the quick survey taken at the start of the semester by your class (and past classes) taking 431. The `year` variable (your `year` is 2024) indicates the year (Fall) in which the respondents took 431.

There were six years in the `surveyA.xlsx` data where the class had no missing responses on the `statfuture` question. Use **only** the data from the years where the class had **no** missing responses on the `statfuture` item to create your estimates for Question 17.

Rounding your responses to three decimal places, specify a point estimate and 90% confidence interval for the true value of the proportion of subjects who responded with the value 7 (meaning extremely important), using the SAIFS approach. You need only to provide the estimates, not a detailed interpretation of them, in this question.

18 Question 18 (3 points)

Is the sample proportion of "7" values from the 2024 survey (ignoring missing values completely) inside the interval that you reported in Question 17?

a. Yes, and it must be inside the interval, by definition
b. Although it doesn't have to be inside the interval, it is.
c. No, the sample proportion is lower than all values in the interval
d. No, the sample proportion is higher than all values in the interval
e. It is impossible to tell from the information provided.

19 Question 19 (4 points)

The data available in the `projA23.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. Use 1000 bootstrap replications.

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 and the `infer` package.

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 `model_parameters()` 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 using the bootstrap and the `infer` package.

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?

20 Question 20 (3 points)

Consider again the `projA23` data from Question 19, 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`

21 Question 21 (4 points)

Consider again the data in the `projA23.xlsx` file from Question 19, and form a Bayesian linear model to build the model for the number of `lines` of code with the transformation you specified in Question 20, using two variables: the number of `counties` sampled by the student, and the number of `characters` used in writing the Project A reflection statement.

Set the seed for your model to be `431021`, and use the default weakly informative priors. Call the model `m21`.

a. Specify the resulting regression equation you obtain for `m21`.
b. Use model `m21` to predict the value of `lines` for a new project where the student has sampled 650 counties and wrote a reflection with 1000 characters. A point estimate is sufficient for this question, and you can round to zero decimal places.

22 Question 22 (4 points)

I fit four models using linear regression to predict left ventricular ejection fraction (LVEF) for a sample of 120 adults who are suspected of having heart failure. A summary of some key results follows:

Model	Predictors used	R^2	Adjusted R^2	AIC	Residual SD
A	use of hydralazine	.135	.134	503	5.24
B	age, use of hydralazine	.365	.357	422	3.12
C	age, sex, use of hydralazine	.404	.343	458	2.92
D	age, use of hydralazine, angina, sex	.406	.354	456	2.91

Identify the model that …

Rows:

1. explains the largest fraction of the outcome's variation for these 120 adults
2. shows the smallest violations of linear regression modeling assumptions
3. displays the best performance using the Akaike information criterion
4. would be most improved by using a transformation of LVEF as the outcome

Columns:

a. Model A
b. Model B
c. Model C
d. Model D
e. It is impossible to tell.

23 Question 23 (4 points)

Consider the number of community hospital beds per 1000 population that are available in each of the 50 US states as well as the District of Columbia. The data for both 2010 and 2014 are provided for you in the `beds1.csv` data set (where the number of beds per 1000 population are saved under the variable name `beds_index` and the indicator of year under the name `year`), and again in the `beds2.csv` data set, using a different format, where the values for 2010 are saved under the variable name `beds_2010` and the values for 2014 are saved under the variable name `beds_2014`.

Use your understanding of the data to establish an appropriate 90% confidence interval for the true difference in the mean number of hospital beds comparing 2010 and 2014. Note that the appropriate point estimate for the mean difference (2010 - 2014) is 0.161 beds per 1000 residents.

Which of the following approaches show an appropriate way to obtain the desired confidence interval? (CHECK ALL THAT APPLY.)

a. `t.test(beds_index ~ year, var.equal = TRUE, conf.level = 0.90, data = beds1)` which yields (-0.116, 0.438).
b. `t.test(beds_index ~ year, var.equal = TRUE, alpha = 0.10, data = beds1)` which yields (-0.170, 0.492).
c. `set.seed(43121); boot.t.test(beds_index ~ year, conf.level = 0.90, data = beds1)` which yields (-0.114, 0.435).
d. `t.test(beds2$beds_2010 - beds2$beds_2014, conf.level = 0.90)`, which yields (0.126, 0.196).
e. `t.test(beds2$beds_2010 - beds2$beds_2014, sig.level = 0.10)`, which yields (0.119, 0.203).
f. `set.seed(43121); boot.t.test(beds2$beds_2010 - beds2$beds_2014, conf.level = 0.90)`, which yields (0.127, 0.198)
g. None of these responses are appropriate.

24 Question 24 (3 points)

A study reports that the sensitivity of a particular type of mammogram as a screening test for detecting breast cancer is 0.85, while its specificity is 0.80.

In a population of 1,000,000 women in which 2,500 women actually have breast cancer, what is the probability that a woman has breast cancer given that her mammogram is positive? Round your answer to three decimal places.

25 Question 25 (3 points)

Suppose a data set (which I have *not* provided) called `mydat` contains 400 rows and five variables. Specifically, `out` is the outcome of interest, and we have four candidate predictors, labeled `pred_a`, `pred_b`, `pred_c` and `pred_d`. Consider the following code:

```
mydat |> reframe(lovedist(out))
```

```
# A tibble: 1 x 10
      n  miss  mean    sd   med   mad   min   q25   q75   max
  <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1   400     0  199.  114.  200.  145.     1   102   297   400
```

```
miss_var_summary(mydat)
```

```
# A tibble: 5 x 3
  variable n_miss pct_miss
  <chr>     <int>    <num>
1 pred_b       41    10.2
2 pred_a       31     7.75
3 pred_c       17     4.25
4 out           0     0
5 pred_d        0     0
```

Suppose you were willing to create a multiple imputation for your Bayesian model predicting `out` using the four predictors. What mechanism for missingness are you assuming to be true in making that decision? (CHECK ALL THAT APPLY.)

a. Missing At Random (MAR)
b. Missing Completely At Random (MCAR)
c. Missing Not At Random
d. No Collinearity
e. None of the Above

26 Question 26 (3 points)

If running the `miss_case_table()` function on the `mydat` tibble described in Question 25 would produce no cases with more than 1 missing value, how many of the 400 rows in the data set must contain at least one missing value?

27 Question 27 (3 points)

Suppose now that the `miss_case_table()` result for `mydat` (as discussed in Questions 25 and 26) instead produced the following result:

```
miss_case_table(mydat)
```

```
# A tibble: 3 x 3
  n_miss_in_case n_cases pct_cases
           <int>   <int>     <dbl>
1              0     337     84.2
2              1      37      9.25
3              2      26      6.5
```

What is a reasonable minimum number of imputations that we should incorporate in building our final model for the `outcome` in light of these results?

28 Question 28 (3 points)

Once a confidence interval for the difference between two means is calculated, several design changes may be used by a researcher to make a confidence interval wider or narrower. For the changes listed in each of the rows below, indicate the impact of that change the width of the confidence interval by selecting the correct column.

Rows:

1. Increase the level of confidence.
2. Increase the sample size.
3. Use a bootstrap instead of a OLS-based approach to estimate the CI.

Columns:

a. CI will become wider
b. CI will become narrower
c. CI width will not change
d. It is impossible to tell

This is the end of the Quiz. Congratulations!

Session Information

```
session_info()
```

```
R version 4.4.2 (2024-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

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

Package version:
  abind_1.4-8            arrangements_1.1.9     askpass_1.2.1
  backports_1.5.0        base64enc_0.1-3        bayesplot_1.11.1
  bayestestR_0.15.0      BH_1.84.0.0            bigD_0.3.0
  bit_4.5.0              bit64_4.5.2            bitops_1.0.9
  blob_1.2.4             boot_1.3-31            broom_1.0.7
  bslib_0.8.0            cachem_1.1.0           callr_3.7.6
  car_3.1-3              carData_3.0-5          cellranger_1.1.0
  checkmate_2.3.2        cli_3.6.3              clipr_0.8.0
  cmprsk_2.2-12          coda_0.19-4.1          codetools_0.2-20
  colorspace_2.1-1       colourpicker_1.3.0     commonmark_1.9.2
  compiler_4.4.2         conflicted_1.2.0       correlation_0.8.6
  cowplot_1.1.3          cpp11_0.5.0            crayon_1.5.3
  crosstalk_1.2.1        curl_6.0.1             data.table_1.16.2
  datasets_4.4.2         datawizard_0.13.0      DBI_1.2.3
  dbplyr_2.5.0           Deriv_4.1.6            desc_1.4.3
  digest_0.6.37          distributional_0.5.0   doBy_4.6.24
  dplyr_1.1.4            DT_0.33                dtplyr_1.3.1
  dygraphs_1.1.1.6       easystats_0.7.3        effectsize_0.8.9
  emmeans_1.10.5         Epi_2.56               estimability_1.5.1
  etm_1.1.1              evaluate_1.0.1         exactRankTests_0.8-35
  fansi_1.0.6            farver_2.1.2           fastmap_1.2.0
  fontawesome_0.5.3      forcats_1.0.0          foreach_1.5.2
```

Formula_1.2-5	fs_1.6.5	gargle_1.5.2
generics_0.1.3	GGally_2.2.1	ggplot2_3.5.1
ggridges_0.5.6	ggstats_0.7.0	glmnet_4.1-8
glue_1.8.0	gmp_0.7-5	googledrive_2.1.1
googlesheets4_1.1.1	graphics_4.4.2	grDevices_4.4.2
grid_4.4.2	gridExtra_2.3	gt_0.11.1
gtable_0.3.6	gtools_3.9.5	haven_2.5.4
highr_0.11	hms_1.1.3	htmltools_0.5.8.1
htmlwidgets_1.6.4	httpuv_1.6.15	httr_1.4.7
ids_1.0.1	igraph_2.1.1	infer_1.0.7
inline_0.3.20	insight_0.20.5	isoband_0.2.7
iterators_1.0.14	janitor_2.2.0	jomo_2.7-6
jquerylib_0.1.4	jsonlite_1.8.9	juicyjuice_0.1.0
knitr_1.49	labeling_0.4.3	later_1.3.2
lattice_0.22-6	lazyeval_0.2.2	lifecycle_1.0.4
lme4_1.1-35.5	loo_2.8.0	lubridate_1.9.3
magrittr_2.0.3	markdown_1.13	MASS_7.3-61
Matrix_1.7-1	MatrixModels_0.5.3	matrixStats_1.4.1
memoise_2.0.1	methods_4.4.2	mgcv_1.9-1
mice_3.16.0	miceadds_3.17-44	microbenchmark_1.5.0
mime_0.12	miniUI_0.1.1.1	minqa_1.2.8
mitml_0.4-5	mitools_2.4	MKdescr_0.8
MKinfer_1.2	modelbased_0.8.9	modelr_0.1.11
multcomp_1.4-26	munsell_0.5.1	mvtnorm_1.3-2
naniar_1.1.0	nlme_3.1-166	nloptr_2.1.1
nnet_7.3-19	norm_1.0-11.1	numDeriv_2016.8-1.1
openssl_2.2.2	ordinal_2023.12.4.1	pan_1.9
parallel_4.4.2	parameters_0.23.0	patchwork_1.3.0
pbkrtest_0.5.3	performance_0.12.4	pillar_1.9.0
pkgbuild_1.4.5	pkgconfig_2.0.3	plyr_1.8.9
posterior_1.6.0	prettyunits_1.2.0	processx_3.8.4
progress_1.2.3	promises_1.3.0	ps_1.8.1
purrr_1.0.2	quantreg_5.99	QuickJSR_1.4.0
R6_2.5.1	ragg_1.3.3	rappdirs_0.3.3
RColorBrewer_1.1-3	Rcpp_1.0.13-1	RcppArmadillo_14.0.2.1
RcppEigen_0.3.4.0.2	RcppParallel_5.1.9	reactable_0.4.4
reactR_0.6.1	readr_2.1.5	readxl_1.4.3
rematch_2.0.0	rematch2_2.1.2	report_0.5.9
reprex_2.1.1	reshape2_1.4.4	rlang_1.1.4
rmarkdown_2.29	rpart_4.1.23	rstan_2.32.6
rstanarm_2.32.1	rstantools_2.4.0	rstudioapi_0.17.1
rvest_1.0.4	sandwich_3.1-1	sass_0.4.9
scales_1.3.0	see_0.9.0	selectr_0.4.2

shape_1.4.6.1	shiny_1.9.1	shinyjs_2.1.0
shinystan_2.6.0	shinythemes_1.2.0	snakecase_0.11.1
sourcetools_0.1.7.1	SparseM_1.84.2	splines_4.4.2
StanHeaders_2.32.10	stats_4.4.2	stats4_4.4.2
stringi_1.8.4	stringr_1.5.1	survival_3.7-0
sys_3.4.3	systemfonts_1.1.0	tensorA_0.36.2.1
textshaping_0.4.0	TH.data_1.1-2	threejs_0.3.3
tibble_3.2.1	tidyr_1.3.1	tidyselect_1.2.1
tidyverse_2.0.0	timechange_0.3.0	tinytex_0.54
tools_4.4.2	tzdb_0.4.0	ucminf_1.2.2
UpSetR_1.4.0	utf8_1.2.4	utils_4.4.2
uuid_1.2.1	V8_6.0.0	vctrs_0.6.5
viridis_0.6.5	viridisLite_0.4.2	visdat_0.6.0
vroom_1.6.5	withr_3.0.2	xfun_0.49
xml2_1.3.6	xtable_1.8-4	xts_0.14.1
yaml_2.3.10	zoo_1.8-12	