

# 431 Quiz 1 for Fall 2025

Deadline: Wednesday 2025-10-22 at Noon

Thomas E. Love

2025-10-15

## Table of contents

<b>Instructions for Students</b>	<b>3</b>
0.1 The Google Form Answer Sheet . . . . .	3
0.2 The Data Sets . . . . .	3
0.3 Getting Help . . . . .	4
0.4 When Should I Ask For Help? . . . . .	4
0.5 Scoring the Quiz . . . . .	4
0.6 How long should this take? . . . . .	5
0.7 Writing Code into the Google Form . . . . .	6
0.8 R Packages and Love-431.R script . . . . .	6
<b>1 Question 1 (Q01, worth 4 points)</b>	<b>7</b>
Question 1 (continued) . . . . .	8
<b>2 Question 2 (Q02, worth 4 points)</b>	<b>9</b>
<b>3 Question 3 (Q03, worth 4 points)</b>	<b>9</b>
<b>4 Question 4 (Q04, worth 4 points)</b>	<b>10</b>
<b>5 Question 5 (Q05, worth 4 points)</b>	<b>11</b>
<b>6 Question 6 (Q06, worth 4 points)</b>	<b>12</b>
<b>7 Question 7 (Q07, worth 4 points)</b>	<b>12</b>
<b>8 Question 8 (Q08, worth 4 points)</b>	<b>13</b>
<b>9 Question 9 (Q09, worth 4 points)</b>	<b>14</b>

<b>10 Question 10 (Q10, worth 4 points)</b>	<b>15</b>
<b>11 Question 11 (Q11, worth 4 points)</b>	<b>16</b>
<b>12 Question 12 (Q12, worth 4 points)</b>	<b>16</b>
<b>13 Question 13 (Q13, worth 4 points)</b>	<b>18</b>
Output 1 (of 4) for Question 13 . . . . .	19
Output 2 (of 4) for Question 13 . . . . .	20
Output 3 (of 4) for Question 13 . . . . .	20
Output 4 (of 4) for Question 13 . . . . .	21
<b>14 Question 14 (Q14, worth 4 points)</b>	<b>21</b>
<b>15 Question 15 (Q15, worth 4 points)</b>	<b>21</b>
<b>16 Question 16 (Q16, worth 4 points)</b>	<b>22</b>
<b>17 Question 17 (Q17, worth 4 points)</b>	<b>22</b>
<b>18 Question 18 (Q18, worth 4 points)</b>	<b>24</b>
<b>19 Question 19 (Q19, worth 4 points)</b>	<b>24</b>
<b>20 Question 20 (Q20, worth 4 points)</b>	<b>25</b>
<b>21 Question 21 (Q21, worth 4 points)</b>	<b>25</b>
<b>22 Question 22 (Q22, worth 4 points)</b>	<b>26</b>
<b>23 Question 23 (Q23, worth 4 points)</b>	<b>27</b>
<b>24 Question 24 (Q24, worth 4 points)</b>	<b>28</b>
<b>25 Question 25 (Q25, worth 4 points)</b>	<b>28</b>
<b>26 Question 26 (Q26, worth 4 points)</b>	<b>29</b>
<b>27 Question 27 (Q27, worth 4 points)</b>	<b>30</b>
<b>This is the end of the Quiz. Congratulations!</b>	<b>30</b>
<b>Session Information</b>	<b>31</b>

## Instructions for Students

There are **27** questions on this Quiz and this PDF is **33** pages long. Be sure you have all **33** pages<sup>1</sup>. It is to your advantage to answer all **27** 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 post the questions online, use some sort of AI to help you, or discuss the questions on this quiz with anyone other than Dr. 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-quiz-1-answer-sheet-2025>

All of your answers must be submitted through the Google Form by noon on Wednesday 2025-10-22, without exception. The form will close at 12:30 PM on that date, and no extensions will be available, so do not wait until late in the morning on Wednesday 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

I have provided **four** simulated data sets (called **q05.Rds**, **q09.csv**, **q13.xlsx** and **q18.csv**) that are mentioned in the Quiz. In addition, Question 27 uses data from [our 431-data page](#) that you've actually had for some time. These can either be downloaded from our 431-data page, or found in our Shared Drive in the Quiz 1 folder's data sub-folder.

---

<sup>1</sup>The last three pages are just session information.

### 0.3 Getting Help

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

- Ask your quiz question via email to **Thomas dot Love at case dot edu**.

During the Quiz period (3 PM 2025-10-16 through 12:30 PM 2025-10-22) we will not answer questions about the Quiz in TA office hours or in person. Instead, I will respond to questions sent before 9 AM on 2025-10-22 in a timely fashion. Some cautions:

- Specific questions are more likely to get helpful answers.
- I will not review your code or your English for you.
- I will not tell you if your answer is correct, or if it is complete.
- I 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 Dr. Love a question via email.
- 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.

 If you are stuck, ASK A QUESTION

You should **NEVER** spend more than 20 minutes on any question without asking me for help. Just email me at **THOMAS DOT LOVE AT CASE DOT EDU**<sup>2</sup>.

### 0.5 Scoring the Quiz

**Each of the 27 questions is worth 4 points** yielding a total of **108 points**. The questions are not in any particular order, and range in difficulty from “things I expect everyone to get right” to “things that are deliberately tricky”.


---


<sup>2</sup>If you ignore this advice, that's your decision, but you'll have to live with the consequences. If you're taking more than 20 minutes to answer any of these questions, then you're very likely to be misunderstanding something.

## 0.6 How long should this take?

The Quiz is meant to take 5 hours. I expect most students will take 4-6 hours, and some will take as little as 2 or as many as 9. Questions 3, 5, 7 and 14 require you to write a longer response (one to three sentences.) Some other questions may take a little more time (not necessarily because they are harder, but because you have to do some analysis.)


On the other hand...

 email this comic to a friend!


 list all comics


 print this comic

 previous


 next


How long your Prof. thinks it should take to do something		How long it'll actually take you to do it
	↓	↓
"Trivial"	=	There goes your week.
"Easy enough"	=	Pull your hair out for a month.
"About a week"	=	Actually, this is pretty easy. He/she doesn't know there's technology that will do this for you now. Take the week off!
"Should keep you occupied for the rest of the term"	=	He/she will forget they asked you to do this by the end of the term. Don't even bother.
"This might make a good thesis topic"	=	Say hello to your thesis topic.
"Hmmm..."	=	Uh oh.


 -10

 -5

 +5


 +10

 first

 last

WWW.PHDCOMICS.COM

all images © jorge cham

 Emergency Button

JORGE CHAM © 2008

Source: <https://phdcomics.com/comics/archive.php?comid=1093>

## 0.7 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 in your responses.

You should not include the `library` command at any time for your responses on the Quiz that ask for code. Assume in all questions that all of the packages listed below have been loaded in R.

## 0.8 R Packages and Love-431.R script

Here is a list of packages in R. This doesn't mean you need to use all of these packages, and it also doesn't mean that I actually used all of them in building the Quiz and its answer sketch. It means only that I didn't use any other packages, other than **xfun** for session information, so you can do the entire Quiz without using any other packages.

I provided this listing to you [at this link](#), so you can easily copy and paste it into your own RStudio session.

```
library(car)
library(DescTools)
library(Epi)
library(ggdist)
library(ggpubr)
library(glue)
library(infer)
library(janitor)
library(knitr)
library(MKinfer)
library(mosaic)
library(naniar)
library(patchwork)
library(readxl)
library(rstanarm)
library(easystats)
library(tidyverse)

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

theme_set(theme_bw())
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 response.

## 1 Question 1 (Q01, worth 4 points)

A random sample of subjects were taken from a particular health system’s list of adult patients with hypertension who received primary care in that system over a six-month time period. An analyst ingested the data into the `q01` tibble, which I have **not** provided to you.

The researchers gathered the insurance used (there were four levels of **insurance**) by that subject at their most recent primary care visit, then identified (from American Community Survey data) the percentage of residents over the age of 25 in that subject’s home neighborhood (specifically, their census block) who had graduated from high school, and this is found in the `hsgraduates` variable.

Consider the following output provided by the analyst to you.

```
glimpse(q01)
```

```
Rows: 318
```

```
Columns: 3
```

```
$ subject      <chr> "A-0001", "A-0003", "A-0005", "A-0015", "A-0016", "A-0021"~  
$ insurance     <fct> Medicare, Uninsured, Medicaid, Medicare, Medicare, Medicar~  
$ hsgraduates   <dbl> 81.8, 63.2, 86.6, 83.3, 86.9, 87.2, 80.0, 77.1, 70.4, 63.6~
```

```
q01 |> tabyl(insurance) |> adorn_pct_formatting()
```

insurance	n	percent
Medicare	152	47.8%
Commercial	49	15.4%
Medicaid	101	31.8%
Uninsured	16	5.0%

## Question 1 (continued)

```
fit01 <- lm(hsgraduates ~ insurance, data = q01)
model_parameters(fit01, ci = 0.95)
```

Parameter	Coefficient	SE	95% CI	t(314)	p
(Intercept)	78.40	0.76	[76.90, 79.91]	102.53	< .001
insurance [Commercial]	3.14	1.55	[ 0.09, 6.18]	2.03	0.044
insurance [Medicaid]	-1.38	1.21	[-3.76, 1.00]	-1.14	0.256
insurance [Uninsured]	-4.29	2.48	[-9.16, 0.59]	-1.73	0.084

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

```
anova(fit01)
```

### Analysis of Variance Table

Response: hsgraduates

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
insurance	3	959.4	319.79	3.5981	0.01391 *
Residuals	314	27907.5	88.88		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Which of the following statements appropriately describe these results?

**CHECK ALL of the statements below that appropriately describe the results.**

- Less than 10% of the variation in high school graduation rates is explained by a linear model using `insurance` type.
- Our point estimate for the difference in mean high school graduation rate between Commercial and Medicaid patients is 3.14 percentage points.
- Our point estimate for the mean high school graduation rate in Medicare patients is 78.4%.
- From highest to lowest sample mean high school graduation rate, Medicare is the highest, then Commercial, Medicaid, and Uninsured.
- None of these statements describe the results shown above.



## 2 Question 2 (Q02, worth 4 points)

The analyst discussed in Question 1 ran the code below.

```
PostHocTest(aov(fit01), method = "bonferroni", conf.level = 0.95)
```

The table below shows Bonferroni confidence intervals (with a family-wise 95% confidence level) for differences in mean high school graduation rates across the six possible combinations of insurance types. I have lightly edited the output.

insurance comp.	difference	lower bound	upper bound
Commercial-Medicare	3.137	-0.975	7.249
Medicaid-Medicare	-1.379	-4.592	1.835
Uninsured-Medicare	-4.289	-10.868	2.290
Medicaid-Commercial	-4.516	-8.874	-0.158
Uninsured-Commercial	-7.426	-14.634	-0.219
Uninsured-Medicaid	-2.910	-9.645	3.825

According to this output, which of the following pairwise comparisons are **inconsistent** with the population means of the two specified insurance types being equal to each other?

**CHECK ALL of the comparisons that are not consistent with equal population means.**

- a. Commercial and Medicare
- b. Medicaid and Medicare
- c. Uninsured and Medicare
- d. Medicaid and Commercial
- e. Uninsured and Commercial
- f. Uninsured and Medicaid
- g. None of these.

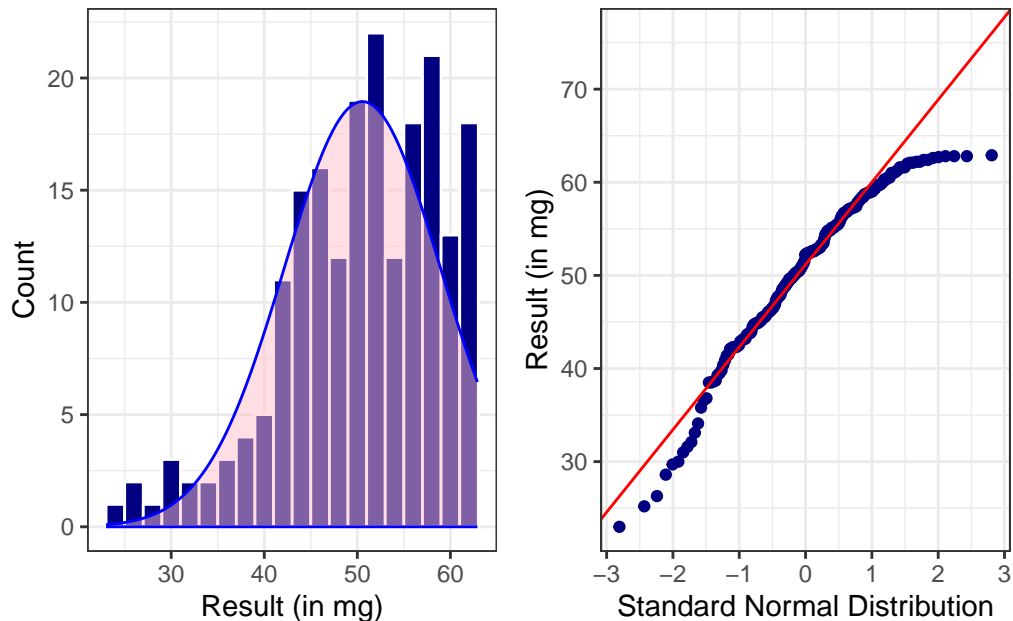
## 3 Question 3 (Q03, worth 4 points)

In the table shown in Question 2, the width of the confidence interval for the comparison of Medicaid to Commercial is about 8.7 percentage points. This is much smaller than the width of the confidence interval for the comparison of Uninsured to Commercial (that width is about 14.4 percentage points.) In one or two **complete English sentences**, tell us why this happens.

**Hint: To answer Question 3, consider all output we've seen from this analyst.**

#### 4 Question 4 (Q04, worth 4 points)

Consider the set of plots below, which use the `q04` tibble, which I have **not** provided, to describe a particular `result`, which is measured in milligrams, for a random sample of 200 subjects from a population of interest.



Which one of the following statements best describes this output?

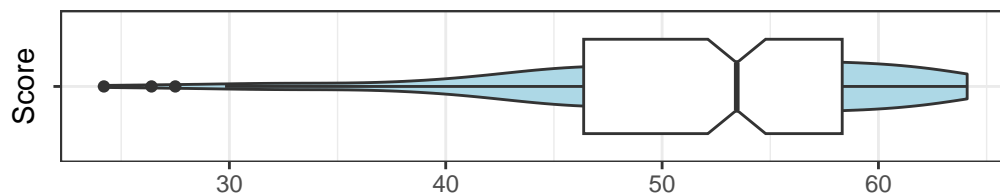
- a. The `result` data are left skewed.
- b. The `result` data are right skewed.
- c. The `result` data are essentially symmetric, but with more outliers than we would expect if the data followed a Normal distribution.
- d. The `result` data are essentially symmetric, but with fewer outliers than we would expect if the data followed a Normal distribution.
- e. The `result` data are well approximated by a Normal distribution.

## 5 Question 5 (Q05, worth 4 points)

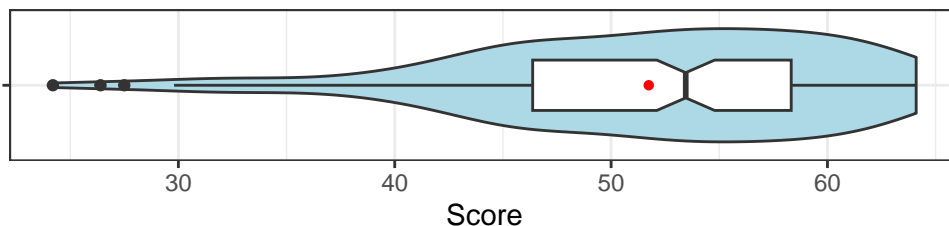
A new analyst wants to produce a boxplot using data stored in the `q05` tibble, containing `score` values for each of 200 subjects who completed a survey. They want their boxplot to include both a violin plot and a red point to indicate the sample mean score. Consider the Figure for Question 5. At the moment, they have produced the plot on the top. They want instead to produce the plot on the bottom, and ask for your help.

Figure for Question 5

Current (BAD) Plot of Q05 scores



Desired (GOOD) Plot of Q05 scores



The code that the analyst used to build the Current (BAD) plot follows. The data are available to you in the `q05.Rds` (R data set) file I have provided.

```
ggplot(q05, aes(x = score, y = "")) +  
  geom_violin(fill = "lightblue", width = 0.4) +  
  stat_summary(fun = mean, geom = "point", shape = 16, col = "red") +  
  geom_boxplot(notch = TRUE) +  
  labs(x = "", y = "Score", title = "Current (BAD) Plot of Q05 scores")
```

There are four things that must be changed in this code in order to produce the desired plot of Q05 scores. One of those four things is to change the first two words in the title specification in the `labs()` command from “Current (BAD)” to “Desired (GOOD)”. Another change that needs to be made is to move the `geom_boxplot()` statement after the `geom_violin()` line and, in particular, before the `stat_summary()` line.

Implementing those two changes, we would then have:

```
ggplot(q05, aes(x = score, y = "")) +  
  geom_violin(fill = "lightblue", width = 0.4) +  
  geom_boxplot(notch = TRUE) +  
  stat_summary(fun = mean, geom = "point", shape = 16, col = "red") +  
  labs(x = "", y = "Score", title = "Desired (GOOD) Plot of Q05 scores")
```

In two or three complete English sentences, specify how to make **the other two** necessary changes to this code in order to produce the desired plot shown at the bottom of the Figure for Question 5.

## 6 Question 6 (Q06, worth 4 points)

Questions 6 and 7 use the same information.

Ohio defines a student as chronically absent from school when they miss at least 10% of class time in a school year, or roughly two days per month<sup>3</sup>. Excused and unexcused absences, as well as suspensions all count toward the student's total absences. Ohio's average rate of chronic absence in non-charter schools was 25.1% in the 2024-25 school year. For a random sample of students enrolled in Ohio charter schools in 2024-25, suppose that 72 students met the definition of "chronically absent" while another 136 did not.

Use a **99%** confidence level and the Agresti-Coull estimation procedure to specify an appropriate point and confidence interval estimate for the rate of chronic absence in Ohio charter schools in 2024-25. Express your rate estimates as proportions, between 0 and 1, rounding to three decimal places. No complete sentences are required for Question 6.

## 7 Question 7 (Q07, worth 4 points)

In two or three complete sentences, what evidence does the sample we presented and calculation you did in Question 6 provide as to whether Ohio's charter school enrollees are either more likely or less likely to be chronically absent than students in Ohio's non-charter schools in 2024-25?

---

<sup>3</sup>Questions 6 and 7 are motivated by Laura Hancock's 2025-10-07 post to [Cleveland.com](#). There is no reason to read the article when responding to the Quiz.

## 8 Question 8 (Q08, worth 4 points)

Which of the following arguments regarding the value of a bootstrap approach to statistical inference are part of Spiegelhalter's discussion of bootstrapping and uncertainty intervals in his chapter about "How sure can we be about what is going on?" in *The Art of Statistics*. (CHECK ALL THAT APPLY.)

- a. Bootstrapping makes good use of computer power.
- b. Bootstrapping requires challenging assumptions about Normality.
- c. Bootstrapping does not require complex probability theory.
- d. Bootstrapping can be used to understand sampling distributions of estimates.
- e. Bootstrapping can be applied in both simple and complex estimation situations.
- f. None of these statements are part of Spiegelhalter's discussion.

## 9 Question 9 (Q09, worth 4 points)

The simulated data in this question are based on a study of systematic voice training combined with swallowing function exercises for the prevention of swallowing problems in adults who have had a stroke<sup>4</sup>. The **q09.csv** file provided to you contains the data we will use in Questions 9-12 of the Quiz.

The study used a combined intervention to try to improve (among other things) quality of life, using a measure of physiological quality of life estimated from responses to a survey. Higher quality of life is associated with higher scores on the measure. The intervention included systematic voice training combined with swallowing function exercise in a sample of 61 stroke patients, of whom 56 provided complete data at baseline (stored in **phys\_base**) and again one month after the intervention was complete (stored in **phys\_post**). Brief summaries of the data across those 56 subjects are shown below.

```
q09 <- read_csv("data/q09.csv", show_col_types = FALSE) |>
  mutate(subject = as.character(subject)) |>
  janitor::clean_names()

q09 |> describe_distribution()
```

Variable	Mean	SD	IQR	Range	Skewness	Kurtosis	n	n_Missing
phys_base	59.83	4.18	6.95	[48.50, 69.30]	-0.04	-0.16	56	0
phys_post	70.82	2.72	4.12	[64.20, 75.80]	-0.31	-0.63	56	0

I created a **diff** variable in the **q09** tibble, then ran the following summary.

```
q09 |> reframe(lovedist(diff)) |> kable(digits = 2)
```

n	miss	mean	sd	med	mad	min	q25	q75	max
56	0	10.99	4.92	11.3	4.74	-0.7	7.53	14	21.8

I used a single line of R code to create the **diff** variable within the **q09** tibble. What was that line of code?

---

<sup>4</sup>There is no reason to read [the full paper](#) when working on the Quiz, but if you want to see it, there you go.

## 10 Question 10 (Q10, worth 4 points)

Still working with the q09 data from Question 9, I obtained the following model results...

```
set.seed(431010)
fit10 <- stan_glm(diff ~ 1, data = q09, refresh = 0)
model_parameters(fit10, ci = 0.90)
```

Parameter	Median	90% CI	pd	Rhat	ESS	Prior
(Intercept)	11.00	[9.88, 12.11]	100%	1.001	2556	Normal (10.99 +- 12.29)

Uncertainty intervals (equal-tailed) computed using a MCMC distribution approximation.

Which one of the following statements best describes the findings presented above?

- a. The median of the paired (1 month post - baseline) differences is estimated to be 11 points on the physiological quality of life scale using the Bayesian model `fit10`.
- b. Our model `fit10`'s best estimate of the true mean difference is an increase from baseline to one month after the intervention of 10.99 points in the physiological quality of life measure.
- c. The mean of the paired (1 month post - baseline) differences is estimated to be 12.29 points on the physiological quality of life scale using the Bayesian model `fit10`.
- d. Our model `fit10`'s best estimate of the true mean difference is an increase from baseline to one month after the intervention of 11 points in the physiological quality of life measure.
- e. The intercept of the regression model shown in `fit10` is not interesting, because we have failed to estimate the model's slope.
- f. None of these responses is correct.

## 11 Question 11 (Q11, worth 4 points)

Suppose you were asked to carefully interpret the meaning of the 90% CI provided in the output for Question 10. Which of the ideas represented by the following statements would need to be incorporated into your response, in order to meet that request?

(CHECK ALL OF THE ELEMENTS THAT SHOULD BE INCLUDED)

- a. Our data are consistent with a change in means for the population of interest of between 9.88 and 12.11.
- b. Our key outcome of interest here is the true mean change from baseline to one month after the intervention.
- c. The appropriate units for this confidence interval estimate are points on the physiological quality of life measure.
- d. We can make our statement about this interval estimate with 90% confidence.
- e. We must be willing to assume that our sample of 56 stroke patients with complete data is representative (ideally a random sample) of the population in which we are interested.
- f. We must be willing to assume that the Bayesian linear model (`fit10`) with weakly informative priors is an appropriate model in this situation.
- g. None of these statements would be appropriate.

## 12 Question 12 (Q12, worth 4 points)

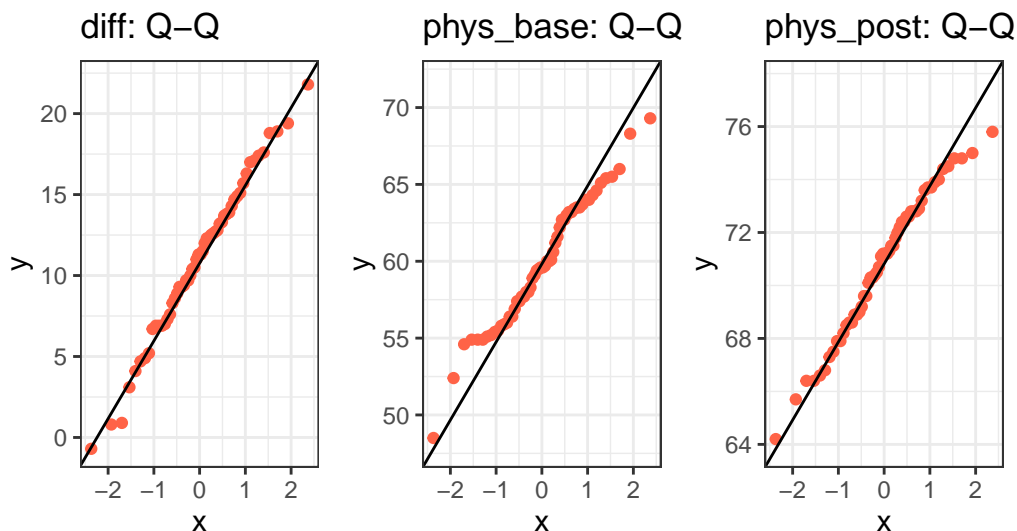
Still working with the `q09` data from Questions 9-11, I ran the following code...

```
p1 <- ggplot(q09, aes(sample = diff)) +  
  geom_qq(col = "tomato") + geom_qq_line(col = "black") +  
  labs(title = "diff: Q-Q")  
  
p2 <- ggplot(q09, aes(sample = phys_base)) +  
  geom_qq(col = "tomato") + geom_qq_line(col = "black") +  
  labs(title = "phys_base: Q-Q")  
  
p3 <- ggplot(q09, aes(sample = phys_post)) +  
  geom_qq(col = "tomato") + geom_qq_line(col = "black") +  
  labs(title = "phys_post: Q-Q")  
  
p1 + p2 + p3 +  
  plot_annotation(title = "Figure for Question 12",  
                  subtitle = "Data from q09 file")
```



## Figure for Question 12

Data from q09 file



Does the Figure for Question 12 provide any reason to be concerned about the accuracy of your findings in Question 10 or 11? Why or why not?

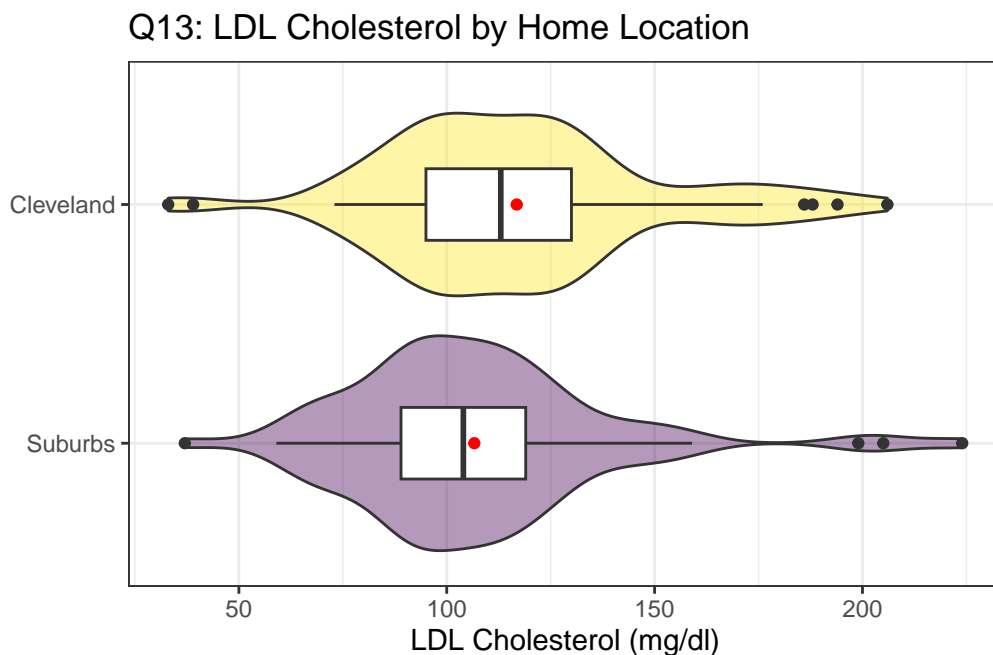
- a. Yes, because the plot of the **diff** information shows problems with assuming Normality.
- b. Yes, because the plots of **phys\_base** and **phys\_post** show problems with assuming Normality.
- c. Yes, because we have a small sample size.
- d. No, because the plot of the **diff** information show no serious problems with assuming Normality.
- e. No, because the plots of **phys\_base** and **phys\_post** show no serious problems with assuming Normality.
- f. No, because we are fitting a Bayesian model, so Normality of our sample is not important.
- g. No, because we have a large sample size.
- h. None of the above statements are accurate.

## 13 Question 13 (Q13, worth 4 points)

The `q13.xlsx` file I provided will be used in Questions 13-15. The file describes a random sample of 154 subjects from the electronic records of adult patients at System ABC with a history of statin prescriptions in the past two years. The data include a **subject** id code, an LDL cholesterol level (in mg/dl), and also indicate whether the subject lives within the city of Cleveland (`cleve` = 1), or in the Cuyahoga County suburbs (`cleve` = 0) outside Cleveland. I ingested the data into a tibble called `q13`, cleaned up the names, thanks to the **janitor** package, and converted the `cleve` information into a factor with categories Cleveland for 1, and Suburbs for 0.

Here is a plot of LDL cholesterol values (in mg/dl) by home location.

```
ggplot(q13, aes(x = cleve, y = ldl)) +  
  geom_violin(aes(fill = cleve)) + geom_boxplot(width = 0.3) +  
  stat_summary(fun = mean, geom = "point", col = "red") +  
  scale_fill_viridis_d(alpha = 0.4) + guides(fill = "none") +  
  labs(title = "Q13: LDL Cholesterol by Home Location",  
       x = "", y = "LDL Cholesterol (mg/dl)") + coord_flip()
```



Here is a numerical summary of some of the data.

```
q13 |> group_by(cleve) |> reframe(loveidist(ldl)) |> kable(digits = 2)
```

cleve	n	miss	mean	sd	med	mad	min	q25	q75	max
Suburbs	77	0	106.64	31.14	104	22.24	37	89	119	224
Cleveland	77	0	116.83	32.47	113	26.69	33	95	130	206

Our goal is to estimate (using 95% confidence) the Cleveland - Suburbs difference in means of LDL cholesterol for the population from which these data are sampled. To help, I created four pieces of output, which follow.

### Output 1 (of 4) for Question 13

```
set.seed(134)
boot.t.test(ldl ~ cleve, var.equal = TRUE, data = q13)
```

Bootstrap Two Sample t-test

data: ldl by cleve  
number of bootstrap samples: 9999  
bootstrap p-value = 0.0454  
bootstrap difference of means (SE) = -10.1088 (5.146439)  
95 percent bootstrap percentile confidence interval:  
-20.1168831 0.2077922

Results without bootstrap:  
t = -1.9884, df = 152, p-value = 0.04857  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-20.32458952 -0.06502087  
sample estimates:  
mean in group Suburbs mean in group Cleveland  
106.6364 116.8312

## Output 2 (of 4) for Question 13

```
t.test(ldl ~ cleve, var.equal = FALSE, data = q13)
```

Welch Two Sample t-test

data: ldl by cleve

t = -1.9884, df = 151.73, p-value = 0.04857

alternative hypothesis: true difference in means between group Suburbs and group Cleveland is

95 percent confidence interval:

-20.32473175 -0.06487864

sample estimates:

mean in group Suburbs mean in group Cleveland

106.6364

116.8312

## Output 3 (of 4) for Question 13

```
set.seed(123)
fit13c <- stan_glm(ldl ~ cleve, data = q13, refresh = 0)
fit13c |> model_parameters(ci = 0.95)
```

Parameter	Median	95% CI	pd	Rhat	ESS	Prior
(Intercept)	106.57	[99.62, 114.01]	100%	1.000	3300	Normal (111.73 +- 80.30)
cleveCleveland	10.19	[ 0.11, 20.20]	97.65%	0.999	3576	Normal (0.00 +- 160.07)

Uncertainty intervals (equal-tailed) computed using a MCMC distribution approximation.

### Output 4 (of 4) for Question 13

```
wilcox.test(ldl ~ cleve, conf.int = TRUE, data = q13)
```

Wilcoxon rank sum test with continuity correction

```
data: ldl by cleve
W = 2285, p-value = 0.01413
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -18.999994 -2.000024
sample estimates:
difference in location
      -10.00002
```

According to the output and summaries provided, which one of the following options shows the best available point and 95% confidence interval estimate for the mean (Cleveland - Suburbs) difference?

- a. 10.19 (0.11, 20.20) from the Bayesian linear model
- b. 10.11 (-0.21, 20.12) from the Bootstrap Two Sample t-test
- c. 10.19 (0.07, 20.32) from the Results without Bootstrap
- d. 10.19 (0.06, 20.32) from the Welch Two Sample t-test
- e. 10.00 (2.00, 19.00) from the Wilcoxon rank sum test

### 14 Question 14 (Q14, worth 4 points)

In a complete English sentence or two, explain why you selected the method you did for Question 13.

### 15 Question 15 (Q15, worth 4 points)

Now, fit a **99%** confidence interval for the mean “Cleveland” minus “Suburbs” difference to the `q13` data using the method you selected in Question 13, and specify the result (both lower and upper limits of the CI) rounded to two decimal places. Be sure to include appropriate units in your response.

**Note:** If you select a method requiring a random seed, use the same seed for that method that I used in preparing the Question 13 output.

## 16 Question 16 (Q16, worth 4 points)

In the introduction and early chapters of *The Art of Statistics*, David Spiegelhalter describes several common features of a strong visualization of data, including some identified by Alberto Cairo. Which of the following features are described as being characteristic of high-quality work in this regard? (CHECK ALL THAT APPLY.)

- a. The graph contains reliable information.
- b. The graph helps to raise more questions and encourage the reader to explore.
- c. The graph is accompanied by a meaningful title, and clear labels and captions.
- d. The graph is honest, clear, and contains deep insights.
- e. The graph is of the raw data only.
- f. The graph should never connect data points gathered at different times.
- g. The graph's appearance is presented in an attractive way.
- h. The graph's design is chosen so that relevant patterns become noticeable.
- i. None of these features are described as characteristic of high-quality work.

## 17 Question 17 (Q17, worth 4 points)

I collected **age** in years, and **dbp** (diastolic blood pressure, in mm Hg) for a sample of 100 women. Here are the last two observations.

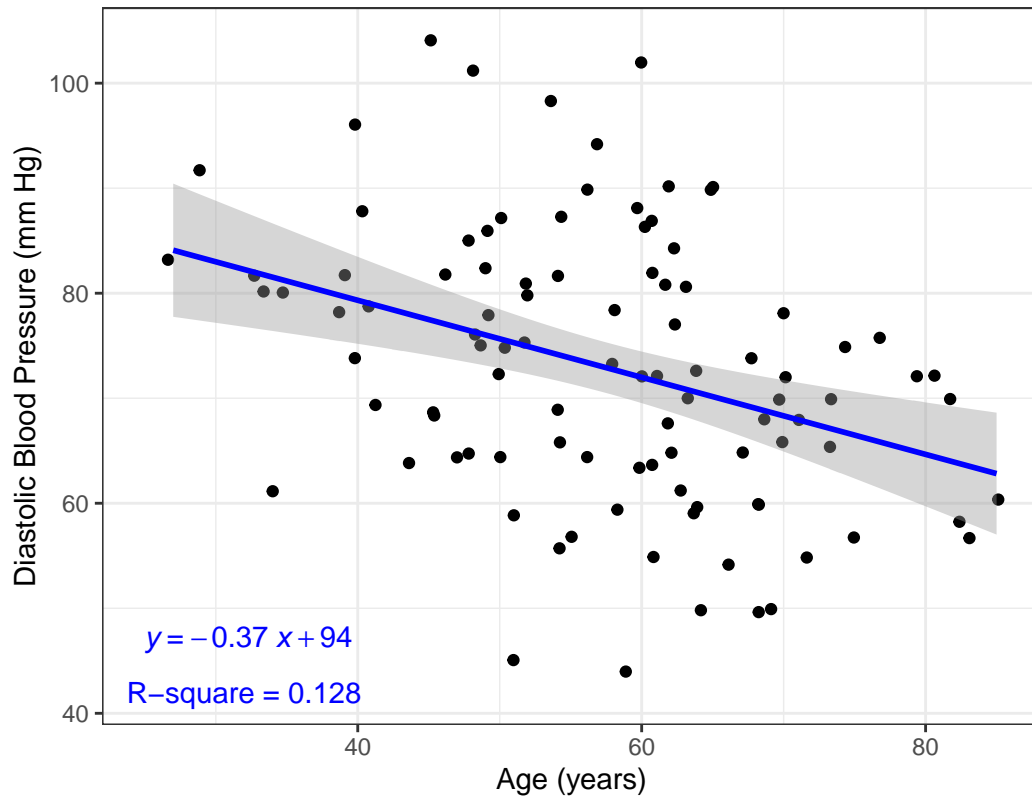
```
tail(q17, 2)
```

```
# A tibble: 2 x 3
  patid   age  dbp
  <chr> <dbl> <dbl>
1 W_099    70    66
2 W_100    56    64
```

The Scatterplot for Question 17 (next page) shows the association of **age** and **dbp**. An ordinary least squares linear regression model for these data yields the smooth and the summaries shown in blue in that Scatterplot.

```
ggplot(data = q17, aes(x = age, y = dbp)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = TRUE, formula = y ~ x, col = "blue") +
  stat_regline_equation(label.x = 25, label.y = 47, col = "blue") +
  annotate("text", x = 32, y = 42, label = "R-square = 0.128", col = "blue") +
  labs(title = "Scatterplot for Question 17",
       y = "Diastolic Blood Pressure (mm Hg)", x = "Age (years)" )
```

Scatterplot for Question 17



After rounding to two decimal places, what is the Pearson correlation between dbp and age in this sample of 100 women?

- a. 0.01
- b. -0.01
- c. 0.13
- d. -0.13
- e. 0.36
- f. -0.36
- g. 0.37
- h. -0.37
- i. 0.94
- j. -0.94
- k. There is insufficient information provided to answer the question.
- l. None of these answers are correct.

## 18 Question 18 (Q18, worth 4 points)

The data in the `q18.csv` file provided to you describe subjects who, at the start of the study, were between 25 and 90 years of age, and showed normal kidney function (as determined by their most recent serum creatinine<sup>5</sup> falling between 0.6 and 1.1 mg/dl if female, or between 0.7 and 1.3 mg/dl if male.) For each subject, we collected diastolic blood pressure at the start of the study (`dbp_now`, in mm Hg), and we will use this to predict that subject's systolic blood pressure (`sbp_3mo`, also in mm Hg) three months later<sup>6</sup>.

Starting with the data in `q18.csv`, first restrict yourself (after cleaning names) to the subset of the original data including only women who received care in **practice A**. (The `sex` variable is either Male or Female, and there are three **practice** locations in the data.)

Using only those women, fit a linear model (using ordinary least squares) to predict systolic blood pressure 3 months after the study began, using diastolic blood pressure at the start of the study and provide the following information:

- the number of women in practice A that you included in your model, and
- the regression equation you obtained, specifically including the point estimates of both the intercept term and the slope, each rounded to two decimal places.

**Note:** Question 18 does not ask for complete sentences.

## 19 Question 19 (Q19, worth 4 points)

Fit a complete set of five diagnostic plots for the model you fit in Question 18, using `set.seed(43119)`. In light of those plots, which of the following statements are true? (CHECK ALL of the true statements.)

- The posterior predictive check shows notable under-prediction near the middle of the systolic blood pressure distribution.
- The posterior predictive check shows notable over-prediction near the middle of the systolic blood pressure distribution.
- The linearity plot suggests some concern about a non-linear relationship.
- The homogeneity of variance plot suggests some concern about a non-linear relationship.
- The influential observations plot indicates multiple subjects outside the 0.5 contour line.
- The Normal Q-Q plot shows large skew to the left in the residuals.
- None of the above statements are true.

---

<sup>5</sup>Subjects were excluded if they did not have a serum creatinine value in the past year.

<sup>6</sup>I concede that this may a somewhat weird thing to do.



## 20 Question 20 (Q20, worth 4 points)

For this question, restrict yourself to all subjects who received care in practices B or C from the original q18 data. Use a Box-Cox transformation to consider possible transformations of the outcome (`sbp_3mo`) we've been interested in, when predicted using current DBP (`dbp_now`).

Which of the following best describes what the Box-Cox approach suggests for this sample<sup>7</sup>?

- a. Take the inverse of the outcome.
- b. Take the square of the outcome.
- c. Take the square root of the outcome.
- d. Use the raw outcome, without any transformation.
- e. Take the natural logarithm of the outcome.
- f. None of these options match what the Box-Cox approach suggests.

## 21 Question 21 (Q21, worth 4 points)

Using an OLS model incorporating the transformation you chose in Question 20, what is the predicted systolic blood pressure (3 months later) for a subject whose current diastolic blood pressure is equal to the median `dbp_now` across all 745 subjects included in the model for Question 20? Round your response to one decimal place. Be sure that your answer includes the units of measurement.

Hint: A complete sentence is not required in Question 21. Only a point estimate rounded to one decimal place (with appropriate units) is required.

---

<sup>7</sup>This sample should include 745 subjects, incidentally.

## 22 Question 22 (Q22, worth 4 points)

The `starwars` data set is part of the `dplyr` package loaded by the tidyverse. You can learn more about it at <https://dplyr.tidyverse.org/reference/starwars.html> if you like. In that data, we find information on Star Wars characters, specifying 14 different variables, including each character's `homeworld` and `gender`. After loading the packages used in developing this Quiz (specified in the Instructions for Students), try running the following code:

```
starwars |>
  select(name, hair_color, birth_year, homeworld, gender) |>
  kable()
```

Note that I'm not showing all of the rows in the table here, to save some space.

name	hair_color	birth_year	homeworld	gender
Luke Skywalker	blond	19.0	Tatooine	masculine
C-3PO	NA	112.0	Tatooine	masculine
R2-D2	NA	33.0	Naboo	masculine
Darth Vader	none	41.9	Tatooine	masculine
Leia Organa	brown	19.0	Alderaan	feminine
Owen Lars	brown, grey	52.0	Tatooine	masculine
Beru Whitesun Lars	brown	47.0	Tatooine	feminine
R5-D4	NA	NA	Tatooine	masculine
Biggs Darklighter	black	24.0	Tatooine	masculine
Obi-Wan Kenobi	auburn, white	57.0	Stewjon	masculine

Your job is to modify the code I've provided to produce a new table which includes only those characters in the `starwars` data set that have:

1. brown hair (do not include people with both brown and grey hair, for example)
2. a birth year of 15 or higher (years are measured here before the Battle of Yavin)
3. a known, non-missing, homeworld.

Now, review your new table, and answer these questions:

- a. How many characters of feminine gender appear in your new table?
- b. How many characters of masculine gender appear in your new table?

## 23 Question 23 (Q23, worth 4 points)

In this question, we consider data describing the age at onset (in years) for 17 women with a diagnosis of multiple sclerosis. The oldest age at onset was 44 years. The stem-and-leaf display shows the data for the first 17 subjects.

The decimal point is 1 digit(s) to the right of the |

```
1 | 46788889
2 | 0367
3 | 239
4 | 24
```

If the next subject added to the data is 28 years of age, which of the following values will decrease, as a result?

- I. The mean
  - II. The standard deviation
  - III. The median
- 
- a. I only
  - b. II only
  - c. III only
  - d. I and II
  - e. I and III
  - f. II and III
  - g. All three statements
  - h. None of the three statements

## 24 Question 24 (Q24, worth 4 points)

In *The Art of Statistics*, Spiegelhalter includes a discussion of “the signal and the noise” in his chapter on modeling relationships using regression. Which TWO of the following statements best describes what is meant by this idea in Spiegelhalter’s context? (Select one option related to “Signal” and one option related to “Noise”, please.)

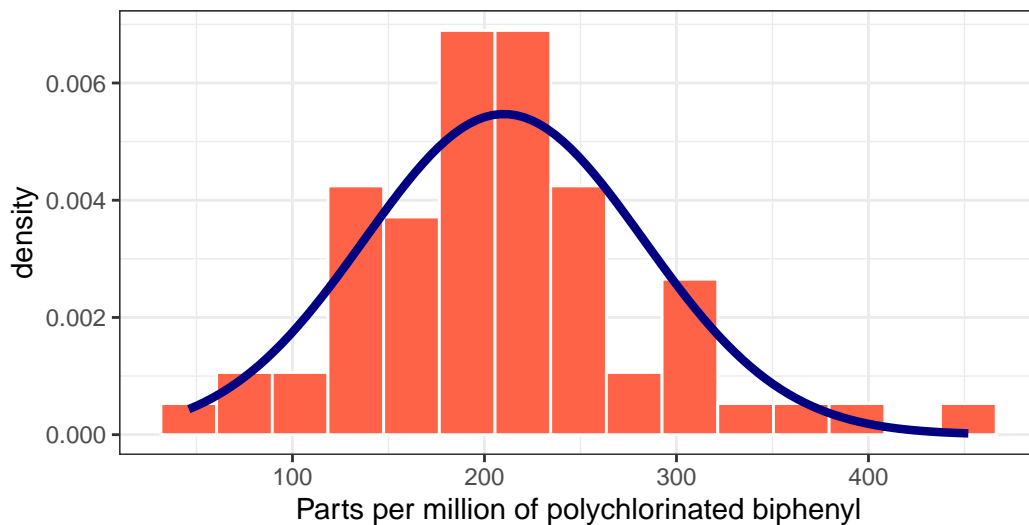
- a. “Signal” describes the actual broadcast of a radio station.
- b. “Signal” describes the essential words in a piece of writing.
- c. “Signal” describes the true pattern allowing for an accurate prediction.
- d. “Signal” describes the useful advice about a decision to be made.
- e. “Noise” describes irrelevant opinions about a decision to be made.
- f. “Noise” describes random variation that can lead to a wrong prediction.
- g. “Noise” describes unwanted interference or modification with a broadcast.
- h. “Noise” describes words that can be edited out of a piece of writing.

## 25 Question 25 (Q25, worth 4 points)

The data for this Question represent the concentration in parts per million of PCB (polychlorinated biphenyl, an industrial pollutant) for 65 Anacapa pelican eggs. The tibble is called `pelican` and the variable of interest is `ppm`. I have not provided these data.

### Question 25. Histogram of ppm of PCB

Data describe 65 Anacapa pelican eggs



Note: The colors used in this plot are called `tomato`, `navy` and `white` in R.

Here are eight lines of code. Note that I definitely used lines 1, 2 and 8 in my code to build the plot for Question 25. I also used some of the other lines (lines 3-7) but not all of them.

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

2 ggplot(pelican, aes(x = ppm)) +
3   geom_density(col = "navy", lwd = 1.5) +
4   geom_histogram(aes(y = after_stat(density)), bins=15, fill="tomato", col="white") +
5   geom_histogram(bins = 15, fill = "tomato", col = "white") +
6   stat_function(fun = dnorm,
                  args = list(mean = mean(pelican$ppm), sd = sd(pelican$ppm)),
                  col = "navy", lwd = 1.5) +
7   coord_flip() +
8   labs(title = "Question 25. Histogram of ppm of PCB",
         subtitle = "Data describe 65 Anacapa pelican eggs",
         x = "Parts per million of polychlorinated biphenyl")
```

Please select each of the line numbers that should be REMOVED from the code in order to create the Question 25 plot. (YOU MAY SELECT MORE THAN ONE OPTION.)

- a. Line 3
- b. Line 4
- c. Line 5
- d. Line 6
- e. Line 7

## 26 Question 26 (Q26, worth 4 points)

Suppose you are interested in how effectively shell thickness might be used to predict the concentration of environmental pollutants, in a setting like the study developed in Question 25. Which variable should go on the vertical (Y) axis of your scatterplot to display and model this association?

- a. the concentration in parts per million of PCB
- b. the thickness in micrometers of the egg's shell
- c. the egg identification number (1-65)
- d. It doesn't matter.
- e. It is impossible to tell from the information provided.

## 27 Question 27 (Q27, worth 4 points)

Consider the data from the 15 question survey done on [day 2 of class](#). Data from 2014 through 2025 are stored on our 431-data page [at this link](#).

Consider all subjects from 2024 or 2025 (this information is found in the **year** variable) who also have complete data on these two variables:

- **english**: Is English your most comfortable language?
- **glasses**: Do you wear corrective lenses (contacts or glasses)?

Develop an appropriate summary of the data, and then use it (**along with a Bayesian augmentation**) to obtain a point estimate and **90%** confidence interval for the relative risk of wearing corrective lenses comparing those for whom English is not their most comfortable language to those for whom English is their most comfortable language. Round each of your responses to three decimal places.

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

Note that we are just asking for the number in part a, and that your answer to part b should be of the form (0.123, 4.567). Question 27 doesn't require any complete sentences.

**This is the end of the Quiz. Congratulations!**

## Session Information

```
xfun::session_info()
```

R version 4.5.1 (2025-06-13 ucrt)  
Platform: x86\_64-w64-mingw32/x64  
Running under: Windows 11 x64 (build 26100)

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	base_4.5.1	base64enc_0.1-3
bayesplot_1.14.0	bayestestR_0.17.0	BH_1.87.0.1
bit_4.6.0	bit64_4.6.0-1	blob_1.2.4
boot_1.3-32	broom_1.0.10	bslib_0.9.0
cachem_1.1.0	callr_3.7.6	car_3.1-3
carData_3.0-5	cellranger_1.1.0	checkmate_2.3.3
class_7.3-23	cli_3.6.5	clipr_0.8.0
cmprsk_2.2-12	coda_0.19-4.1	codetools_0.2-20
colourpicker_1.3.0	commonmark_2.0.0	compiler_4.5.1
conflicted_1.2.0	correlation_0.8.8	corrplot_0.95
cowplot_1.2.0	cpp11_0.5.2	crayon_1.5.3
crosstalk_1.2.2	curl_7.0.0	data.table_1.17.8
datasets_4.5.1	datawizard_1.3.0	DBI_1.2.3
dbplyr_2.5.1	Deriv_4.2.0	desc_1.4.3
DescTools_0.99.60	digest_0.6.37	distributional_0.5.0
doBy_4.7.0	dplyr_1.1.4	DT_0.34.0
dtplyr_1.3.2	dygraphs_1.1.1.6	e1071_1.7-16
easystats_0.7.5	effectsize_1.0.1	emmeans_1.11.2-8
Epi_2.61	estimability_1.5.1	etm_1.1.2
evaluate_1.0.5	Exact_3.3	exactRankTests_0.8-35
expm_1.0-0	farver_2.1.2	fastmap_1.2.0
fontawesome_0.5.3	fontBitstreamVera_0.1.1	fontLiberation_0.1.0
fontquiver_0.2.1	forcats_1.0.1	foreach_1.5.2
Formula_1.2-5	fs_1.6.6	gargle_1.6.0

gdtools_0.4.4	generics_0.1.4	ggdist_3.3.3
ggformula_1.0.0	ggiraph_0.9.2	ggplot2_4.0.0
ggpubr_0.6.1	ggrepel_0.9.6	ggridges_0.5.7
ggsci_4.0.0	ggsignif_0.6.4	gld_2.6.8
glmnet_4.1-10	glue_1.8.0	gmp_0.7-5
googledrive_2.1.2	googlesheets4_1.1.2	graphics_4.5.1
grDevices_4.5.1	grid_4.5.1	gridExtra_2.3
gtable_0.3.6	gtools_3.9.5	haven_2.5.5
highr_0.11	hms_1.1.3	htmltools_0.5.8.1
htmlwidgets_1.6.4	httpuv_1.6.16	httr_1.4.7
ids_1.0.1	igraph_2.1.4	infer_1.0.9
inline_0.3.21	insight_1.4.2	isoband_0.2.7
iterators_1.0.14	janitor_2.2.1	jomo_2.7-6
jquerylib_0.1.4	jsonlite_2.0.0	knitr_1.50
labeling_0.4.3	labelled_2.15.0	later_1.4.4
lattice_0.22-7	lazyeval_0.2.2	lifecycle_1.0.4
litedown_0.7	lme4_1.1-37	lmom_3.2
loo_2.8.0	lubridate_1.9.4	magrittr_2.0.4
markdown_2.0	MASS_7.3-65	Matrix_1.7-4
MatrixModels_0.5.4	matrixStats_1.5.0	memoise_2.0.1
methods_4.5.1	mgcv_1.9-3	mice_3.18.0
miceadds_3.18-36	microbenchmark_1.5.0	mime_0.13
miniUI_0.1.2	minqa_1.2.8	mitml_0.4-5
mitools_2.4	MKdescr_0.9	MKinfer_1.2
modelbased_0.13.0	modelr_0.1.11	mosaic_1.9.2
mosaicCore_0.9.5	mosaicData_0.20.4	multcomp_1.4-28
mvtnorm_1.3-3	nanianr_1.1.0	nlme_3.1-168
nloptr_2.2.1	nnet_7.3-20	norm_1.0.11.1
numDeriv_2016.8-1.1	openssl_2.3.4	ordinal_2023.12.4.1
pan_1.9	parallel_4.5.1	parameters_0.28.2
patchwork_1.3.2	pbkrtest_0.5.5	performance_0.15.2
pillar_1.11.1	pkgbuild_1.4.8	pkgconfig_2.0.3
plyr_1.8.9	polynom_1.4-1	posterior_1.6.1
prettyunits_1.2.0	processx_3.8.6	progress_1.2.3
promises_1.3.3	proxy_0.4-27	ps_1.9.1
purrr_1.1.0	quadprog_1.5.8	quantreg_6.1
QuickJSR_1.8.1	R6_2.6.1	ragg_1.5.0
rappdirs_0.3.3	rbibutils_2.3	RColorBrewer_1.1-3
Rcpp_1.1.0	RcppArmadillo_15.0.2.2	RcppEigen_0.3.4.0.2
RcppParallel_5.1.11-1	Rdpack_2.6.4	readr_2.1.5
readxl_1.4.5	reformulas_0.4.1	rematch_2.0.0
rematch2_2.1.2	report_0.6.1	reprex_2.1.1
reshape2_1.4.4	rlang_1.1.6	rmarkdown_2.30



rootSolve_1.8.2.4	rpart_4.1.24	rstan_2.32.7
rstanarm_2.32.2	rstantools_2.5.0	rstatix_0.7.2
rstudioapi_0.17.1	rvest_1.0.5	S7_0.2.0
sandwich_3.1-1	sass_0.4.10	scales_1.4.0
see_0.12.0	selectr_0.4.2	shape_1.4.6.1
shiny_1.11.1	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_4.5.1	StanHeaders_2.32.10
stats_4.5.1	stats4_4.5.1	stringi_1.8.7
stringr_1.5.2	survival_3.8-3	sys_3.4.3
systemfonts_1.3.1	tensorA_0.36.2.1	textshaping_1.0.3
TH.data_1.1-4	threejs_0.3.4	tibble_3.3.0
tidyr_1.3.1	tidyselect_1.2.1	tidyverse_2.0.0
timechange_0.3.0	tinytex_0.57	tools_4.5.1
tzdb_0.5.0	ucminf_1.2.2	UpSetR_1.4.0
utf8_1.2.6	utils_4.5.1	uuid_1.2.1
V8_8.0.0	vctr_0.6.5	viridis_0.6.5
viridisLite_0.4.2	visdat_0.6.0	vroom_1.6.6
withr_3.0.2	xfun_0.53	xml2_1.4.0
xtable_1.8-4	xts_0.14.1	yaml_2.3.10
zoo_1.8-14		