Answer Sketch with Results for 431 Quiz 3

Thomas E. Love

Version: 2021-12-01 23:30:07

Packages Loaded

```
library(broom)
library(car)
library(Epi)
library(fivethirtyeight)
library(Hmisc)
library(janitor)
library(knitr)
library(magrittr)
library(naniar)
library(patchwork)
library(tidyverse)

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

theme_set(theme_bw())
```

1 Answer 01 is `c`.

In total, there are **ten** pairwise comparisons to be made at Watchmaker's Technical Institute: A[my] vs. B[eth], A vs. C[armen], A vs. D[onna], A vs. E[lena], B vs. C, B vs. D, B vs. E, C vs. D, C vs. E and D vs. E. So if we want to retain a 5% significance level with a Bonferroni correction, we'd have to run the two-sample t tests at a significance level 1/10 that size, or 0.005. Any setting larger than that for the Bonferroni-corrected α level will not work.

1.1 Results for Question 01 (3 points)

Question 01	Result
% responses that were correct	68
% of available Points Awarded	68

- No partial credit was available.
- The most common incorrect response was `e`, followed by `b`.

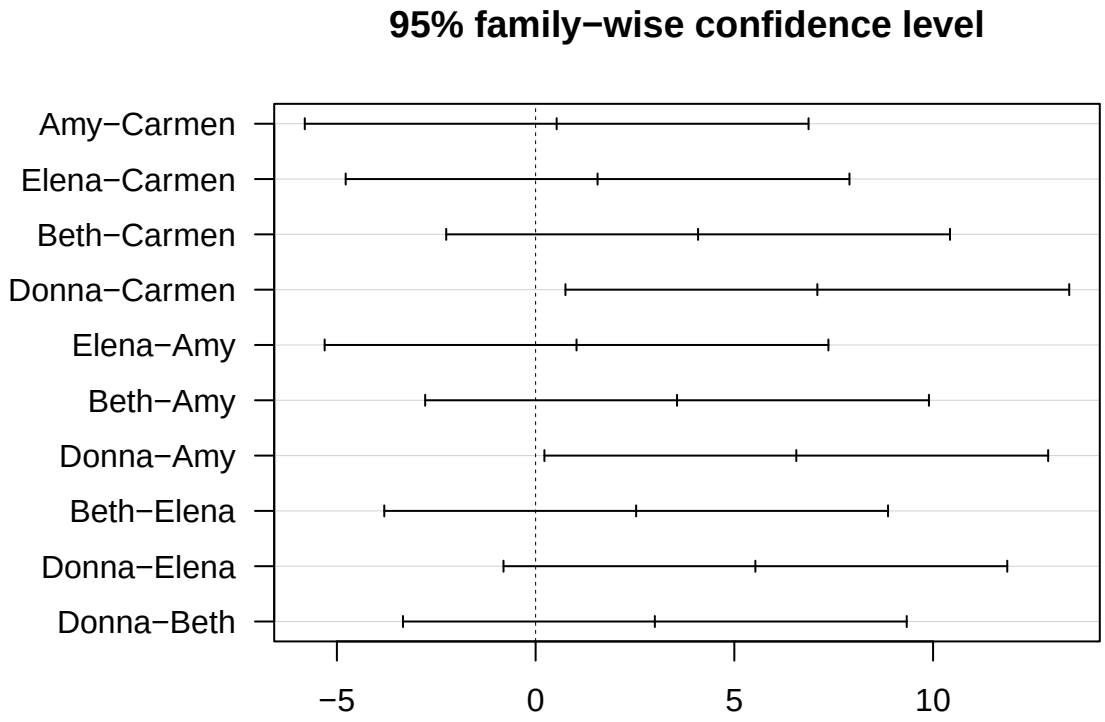
2 Answer 02 is e.

```
wti_exam <- read_csv("data/wti_exam.csv",
                     show_col_types = FALSE) %>%
    clean_names()

tidy(TukeyHSD(aov(exam2 ~ ta, data = wti_exam),
              ordered = TRUE)) %>%
    select(-term, -adj.p.value) %>%
    kable(digits = 2)
```

contrast	null.value	estimate	conf.low	conf.high
Amy-Carmen	0	0.53	-5.81	6.87
Elena-Carmen	0	1.56	-4.78	7.90
Beth-Carmen	0	4.09	-2.25	10.43
Donna-Carmen	0	7.09	0.75	13.43
Elena-Amy	0	1.03	-5.31	7.37
Beth-Amy	0	3.56	-2.78	9.90
Donna-Amy	0	6.56	0.22	12.90
Beth-Elena	0	2.53	-3.81	8.87
Donna-Elena	0	5.53	-0.81	11.87
Donna-Beth	0	3.00	-3.34	9.34

```
par(mar = c(3,8,3,3))
plot(TukeyHSD(aov(exam2 ~ ta, data = wti_exam),
              ordered = TRUE), las = 1)
```



There are ten comparisons being made, but all but two have Tukey HSD confidence intervals which include a mean difference of zero. The exceptions are the comparisons of Donna to Amy and Donna to Carmen In each case, Donna has higher scores, since the D-A and D-C differences are positive.

2.1 Results for Question 02 (3 points)

Question 02	Result
% responses that were correct	95
% of available Points Awarded	95

- No partial credit was available.
- The most common incorrect response was `c`, followed by `b`.

3 Answer 03 is `d`.

We have two especially poorly fitted points in the data, located in rows 28 and 116. These two outliers don't quite have enough leverage to be identified by Cook's distance as influential in our Residuals vs. Leverage plot, but they are very poorly fit, with standardized residuals above +7 (for row 116) and below -4 (for row 28), neither of which is reasonable given virtually any sample size, and certainly not when we have only 150 observations in the data. This is primarily a problem with the assumption of Normality. In this case, one plausible straightforward solution suggests itself: removing these two outliers (and studying them separately) while refitting the model might not change our conclusions appreciably, based on the current leverage and influence profile.

3.1 Results for Question 03 (3 points)

Question 03	Result
% responses that were correct	22
% of available Points Awarded	22

- No partial credit was available.
- The most common incorrect response was `e`, followed by `c`.

This was a very disappointing result, from my perspective. I meant this to be a very straightforward item, and it really didn't turn out that way.

Some of you decided the main problem was with the assumption of constant variance. That's just not a reasonable conclusion, as there is just no suggestion of a fan shape in the residuals vs. fitted values plot, nor is there any evidence of a meaningful trend in the scale-location plot.

An even more common incorrect response was that there was nothing wrong here, which is also not something I can see any defense for. When you have standardized residuals anywhere near this size, it is undoubtedly a major problem for the Normality assumption in your model.

4 Answer 04 is `a`.

Let's run the code and see what we get.

```
data(Pottery)
anova(lm(Na ~ Site, data = Pottery))
```

```
Analysis of Variance Table

Response: Na
          Df  Sum Sq  Mean Sq F value    Pr(>F)
Site       3 0.25825 0.086082  9.5026 0.0003209 ***
Residuals 22 0.19929 0.009059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA F test gives us a p value well below 0.01, so we can certainly conclude that the population mean Na level in at least one of the four sites is different than the others, at a 1% significance level. That is, in fact, what statement `a` says. The other statements (`b`, `c` and `d`) show various mistaken versions of what the ANOVA F test might do.

4.1 Results for Question 04 (3 points)

Question 04	Result
% responses that were correct	92
% of available Points Awarded	92

- No partial credit was available.
- The most common incorrect response was `b`, followed by `e`.

5 Answer 05 is `1-d` and `2-e`

Visualization 1 shows `Na` and Visualization 2 shows `Al`. What's most helpful for me to do the relevant detective work is to obtain a few numerical summaries of each variable.

```
data(Pottery)
mosaicCore::df_stats(~ Ca + Fe + Mg + Na + Al,
                     data = Pottery) %>%
    kable(digits = 2)
```

response	min	Q1	median	Q3	max	mean	sd	n	missing
Ca	0.01	0.06	0.16	0.22	0.31	0.15	0.10	26	0
Fe	0.92	1.70	5.46	6.59	7.09	4.47	2.41	26	0
Mg	0.53	0.67	3.83	4.50	7.23	3.14	2.18	26	0
Na	0.03	0.05	0.15	0.22	0.54	0.16	0.14	26	0
Al	10.10	11.95	13.80	17.45	20.80	14.49	2.99	26	0

There are lots of ways we can get to the correct response. Let's focus on the boxplots.

- Based again on the median shown in the boxplot in Visualization 1, we have two candidates to be `var1`: `Ca` and `Na`, with similar medians. But only `Na` has a maximum value as large as that shown in the boxplot in Visualization 1, so that must be it.
- `var2` describes `Al`, since that's the only element with a median between 12 and 15, as the boxplot in Visualization 2 requires.

5.1 Results for Question 05 (3 points)

Question 05	Result
% responses that were correct	95
% of available Points Awarded	95

- Each part of the question was worth 1.5 points.
- There was no particular pattern to the incorrect responses.

6 Answer 06 is `a-1`, `b-3`, `c-3`, and `d-3`

```
leopard <- read_csv("data/leopard.csv",
                    show_col_types = FALSE) %>%
    clean_names()

modelA <- lm(sqrt(behavior) ~
               x1 + x2 + x3 + x4 + x5 + x6, data = leopard)
modelB <- lm(sqrt(behavior) ~
               x1 + x2 + x3 + x4 + x5, data = leopard)
modelC <- lm(sqrt(behavior) ~
               x1 + x2 + x3 + x4, data = leopard)
modelD <- lm(sqrt(behavior) ~
               x1 + x2, data = leopard)

result06 <- bind_rows(
    glance(modelA),
    glance(modelB),
    glance(modelC),
    glance(modelD)) %>%
    mutate(model = c("A", "B", "C", "D")) %>%
    select(model, r.squared, adj.r.squared, AIC, BIC, nobs)

result06 %>% kable(digits = 3)
```

model	r.squared	adj.r.squared	AIC	BIC	nobs
A	0.881	0.878	565.048	594.678	300
B	0.880	0.878	564.343	590.270	300
C	0.880	0.879	562.487	584.710	300
D	0.500	0.497	987.089	1001.904	300

Model A (of course, as it contains all of the other models) has the highest Multiple R^2 (the raw R^2), but Model C has the best result for adjusted R^2, as well as both AIC and BIC.

Criterion	Best Model
a. Multiple R^2	1. Model A. is largest.
b. AIC	3. Model C is smallest.
c. Adjusted R^2	3. Model C is largest.
d. BIC	3. Model C is smallest.

6.1 Results for Question 06 (3 points)

Question 06	Result
% responses that were correct	92
% of available Points Awarded	92

- Each part of the question was worth 0.75 point.
- Everyone got part a right.
- There was no particular pattern to the incorrect responses to b, c or d.

7 Answer 07 is a, and only a.

Here are the results of running the specified code.

The first piece of code obtains the variance inflation factors in Model A. None of the values exceed 5, or even approach it, so we don't have any indication of severe collinearity. So that suggests that response a is TRUE, and b is FALSE.

```
round(car::vif(modelA),3)
```

```
   x1    x2    x3    x4    x5    x6
1.020 1.012 1.015 1.019 1.969 1.960
```

The second piece of code looks at whether a Box-Cox power transformation would be of help to us in Model A. Since the result is fairly close to 1, I'd say that the square root we're already using looks like a reasonable choice.

```
car::powerTransform(modelA)
```

```
Estimated transformation parameter
       Y1
0.8989649
```

The third piece of code compares Model A to Model B using a test of significance. It suggests that dropping the `x6` predictor removes only negligible predictive value from our model, so that Model B is not detectably worse at predicting `sqrt(behavior)` than Model A at typical choices of significance level. I wouldn't attach a lot of weight to this, but that's what the output suggests.

```
anova(modelA, modelB)
```

```
Analysis of Variance Table

Model 1: sqrt(behavior) ~ x1 + x2 + x3 + x4 + x5 + x6
Model 2: sqrt(behavior) ~ x1 + x2 + x3 + x4 + x5
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    293 109.52
2    294 109.99 -1  -0.47384 1.2677 0.2611
```

None of this output tells us anything about whether the independence assumption is reasonable in this case (the fact that the data are cross-sectional might be helpful in making that determination, but there's no evidence in the output) so c cannot be true. Statement d is actually FALSE on its merits. The residual variance for model A is actually (slightly) smaller than that of model B, as you can see from the output below, and again, there's no direct evidence of this in the output I asked you to create for Question 7.

```
bind_rows(
    glance(modelA),
    glance(modelB),
    glance(modelC),
    glance(modelD)) %>%
    mutate(model = c("A", "B", "C", "D"),
           residual_variance = sigma^2) %>%
    select(model, sigma, residual_variance) %>%
    kable(digits = 4)
```

model	sigma	residual_variance
A	0.6114	0.3738
B	0.6116	0.3741
C	0.6108	0.3730
D	1.2435	1.5462

7.1 Results for Question 07 (3 points)

Question 07	Result
% responses that were correct	83
% of available Points Awarded	87

- People who selected `a` but also one other response received 1.5 points.
- The most common incorrect response was to add `d` to `a`.

8 Answer 08 is `a` and `e`.

The `augment` and `predict` functions will do this job, as demonstrated below. The others will not, or don't actually exist.

```
newleopard <- tibble( behavior = c(49, 24),
                      x1 = c(105, 90), x2 = c(105, 90),
                      x3 = c(105, 90), x4 = c(105, 90),
                      x5 = c(105, 90), x6 = c(105, 90))

predict(modelD, newdata = newleopard)
```

```
       1        2
6.931753 4.980896
```

```
augment(modelD, newdata = newleopard)
```

```
# A tibble: 2 x 9
  behavior    x1    x2    x3    x4    x5    x6 .fitted  .resid
     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>
1       49   105   105   105   105   105   105    6.93  0.0682
2       24    90    90    90    90    90    90    4.98 -0.0819
```

8.1 Results for Question 08 (3 points)

Question 08	Result
% responses that were correct	62
% of available Points Awarded	75

- People who selected `a` or `e` alone received 1.5 points.
- People who selected both `a` and `e` and also something else received 1 point.
- I didn't see a clear pattern in the incorrect responses.

9 Answer 09 is `e`.

To back out of the square root (`sqrt`) and thus return to our original scale for `behavior`, we'd square the predicted values, which we can do using choice `e` to get the correct result (25, 81).

9.1 Results for Question 09 (3 points)

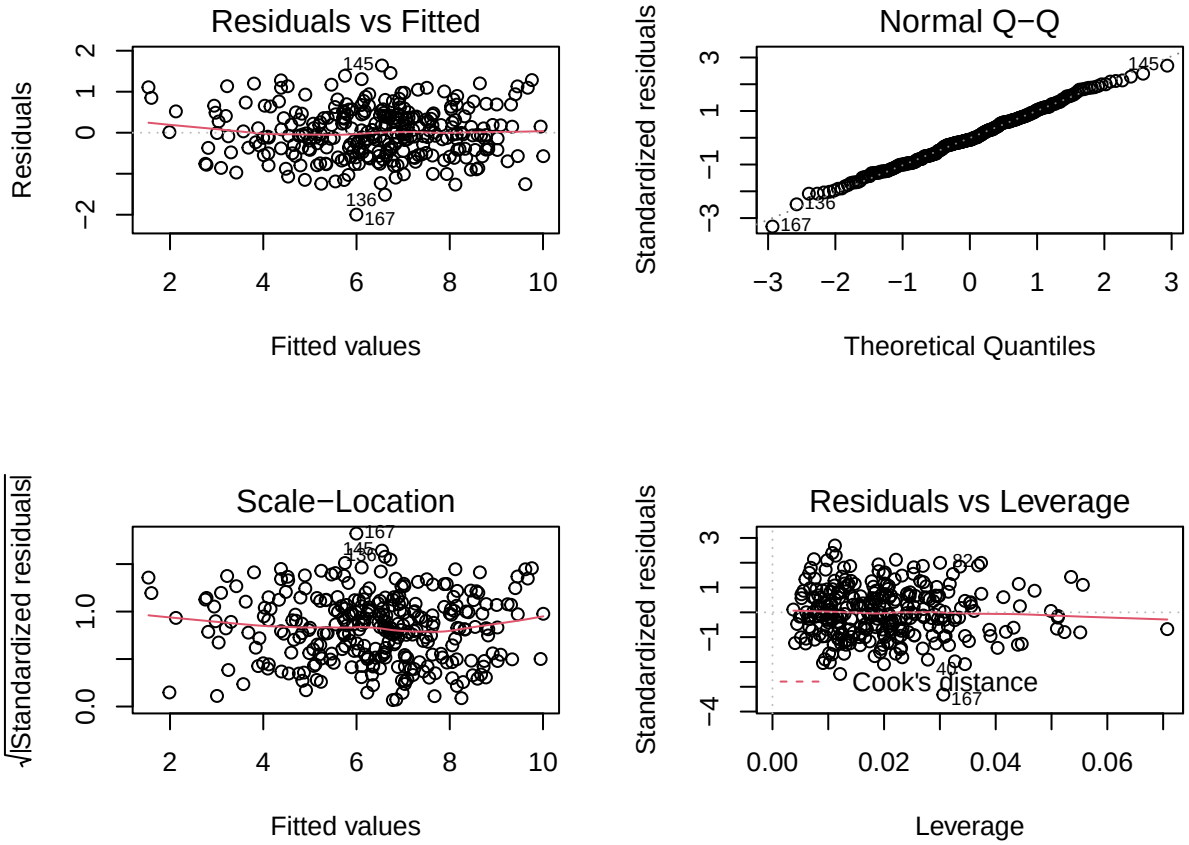
Question 09	Result
% responses that were correct	92
% of available Points Awarded	92

- No partial credit was available.
- The most common incorrect response was `f`.

10 Answer 10 is `d`.

Here are the residual plots for Model B.

```
par(mfrow = c(2,2)); plot(modelB); par(mfrow = c(1,1))
```



I see no serious curve in the plot of residuals vs. fitted values, so there's no clear problem with *linearity*. I see no serious fan shape in the residuals vs. fitted values, and no clear rise or fall in the scale-location plot, so there's no clear problem with the assumption of constant variance. The Normal Q-Q plot shows no sign of substantial problems with the assumption of Normality - neither enormous outliers nor big deviations from symmetry in the body of the plot. Nor are there any clearly influential points. I see no problems. Choice `d` is correct.

10.1 Results for Question 10 (3 points)

Question 10	Result
% responses that were correct	92
% of available Points Awarded	92

- No partial credit was available.
- The most common incorrect response was `e`.

11 Answer 11 is b.

Here's the table, which shows that it's the ages 30-44 group.

```
wc <- fivethirtyeight::weather_check %>%
  select(female, ck_weather, age) %>%
  rename(sex = female) %>%
  mutate(sex = fct_recode(factor(sex),
                          "Female" = "TRUE",
                          "Male" = "FALSE"),
         ck_weather = fct_recode(factor(ck_weather),
                         "Check" = "TRUE",
                         "No Check" = "FALSE")) %>%
  mutate(sex = fct_relevel(sex, "Female"),
         ck_weather = fct_relevel(ck_weather, "Check"))
```

```
wc %>% tabyl(sex, ck_weather, age)
```

```
$`18 - 29`
    sex Check No Check
 Female    79       30
   Male    41       26

$`30 - 44`
    sex Check No Check
 Female   105       28
   Male    56       15

$`45 - 59`
    sex Check No Check
 Female   115       29
   Male   119       15

$`60+`
    sex Check No Check
 Female   121       20
   Male   103       14

$NA_
    sex Check No Check
 Female     0        0
   Male     0        0
   <NA>     7        5
```

11.1 Results for Question 11 (3 points)

Question 11	Result
% responses that were correct	over 95
% of available Points Awarded	over 95

- No partial credit was available.

12 Answer 12 is 916.

```
miss_case_table(wc)
```

```
# A tibble: 2 x 3
  n_miss_in_case n_cases pct_cases
           <int>   <int>     <dbl>
1              0     916     98.7
2              2      12      1.29
```

12.1 Results for Question 12 (3 points)

Question 12	Result
% responses that were correct	over 95
% of available Points Awarded	over 95

- No partial credit was available.

13 Answer 13 estimate = -0.023; 90% CI (-0.066, 0.021)

```
wc_complete <- wc %>% filter(complete.cases(.))

twoby2(wc_complete %$% table(sex, ck_weather),
       conf.level = 0.90)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Check
Comparing : Female vs. Male

       Check No Check    P(Check) 90% conf. interval
Female   420      107      0.7970    0.7666   0.8243
Male     319       70      0.8201    0.7858   0.8499

                                  90% conf. interval
            Relative Risk:  0.9718    0.9215   1.0250
         Sample Odds Ratio:  0.8613    0.6505   1.1406
Conditional MLE Odds Ratio:  0.8615    0.6405   1.1556
    Probability difference: -0.0231   -0.0656   0.0207

             Exact P-value: 0.3982
        Asymptotic P-value: 0.3819
------------------------------------------------------
```

13.1 Results for Question 13 (3 points)

Question 13 Overall	Result
% responses that were correct	65

Question 13 Overall	Result
% of available Points Awarded	80

Point Estimate	Result
% responses that were correct	87
% of available Points Awarded	89

CI Endpoints	Result
% responses that were correct	65
% of available Points Awarded	75

- The point estimate was worth 1 point.
- The confidence interval endpoints were worth 2 points (one point per endpoint.)
- Students who typed in an extra decimal place (-0.0231) got full credit on the point estimate.
- Students who got the sign wrong received half credit for both the point estimate and each of their confidence interval endpoints.
- Rounding errors in the endpoints (-0.065, 0.020) got half-credit for each of the endpoints with that problem.
- Students who typed in an extra decimal place (-0.0656, 0.0207) got full credit on the confidence interval.

14 Answer 14 is c

This is an independent samples study. We want to compare means.

14.1 Results for Question 14 (3 points)

Question 14	Result
% responses that were correct	85
% of available Points Awarded	85

- No partial credit was available.
- The most common incorrect response was `b`, but these are definitely not paired samples.

15 Answer 15 is an essay.

I think the best response is the two-sample pooled t test, given the equal sample sizes in the two Holes and the plots of each Hole which suggest Normal models are reasonable in each case. Here is one version of my response:

> Based on a pooled t test to compare means of independent samples with a balanced design and Normal distributions within each sample, our point estimate for the Hole A minus Hole B difference in mean size is -19.65, with 90% confidence interval (-35.24, -4.06). Since 0 is not included in that interval, we have detectable evidence (with 90% confidence) that the rhinos living near Hole B are larger in size than those living near Hole A.

Since these are, in fact, independent samples, some rearrangement of the data is required. In particular, we need to make the data set longer, and set up appropriate variables to describe the location (Hole A or Hole B) and size for each of the 80 rhinos.

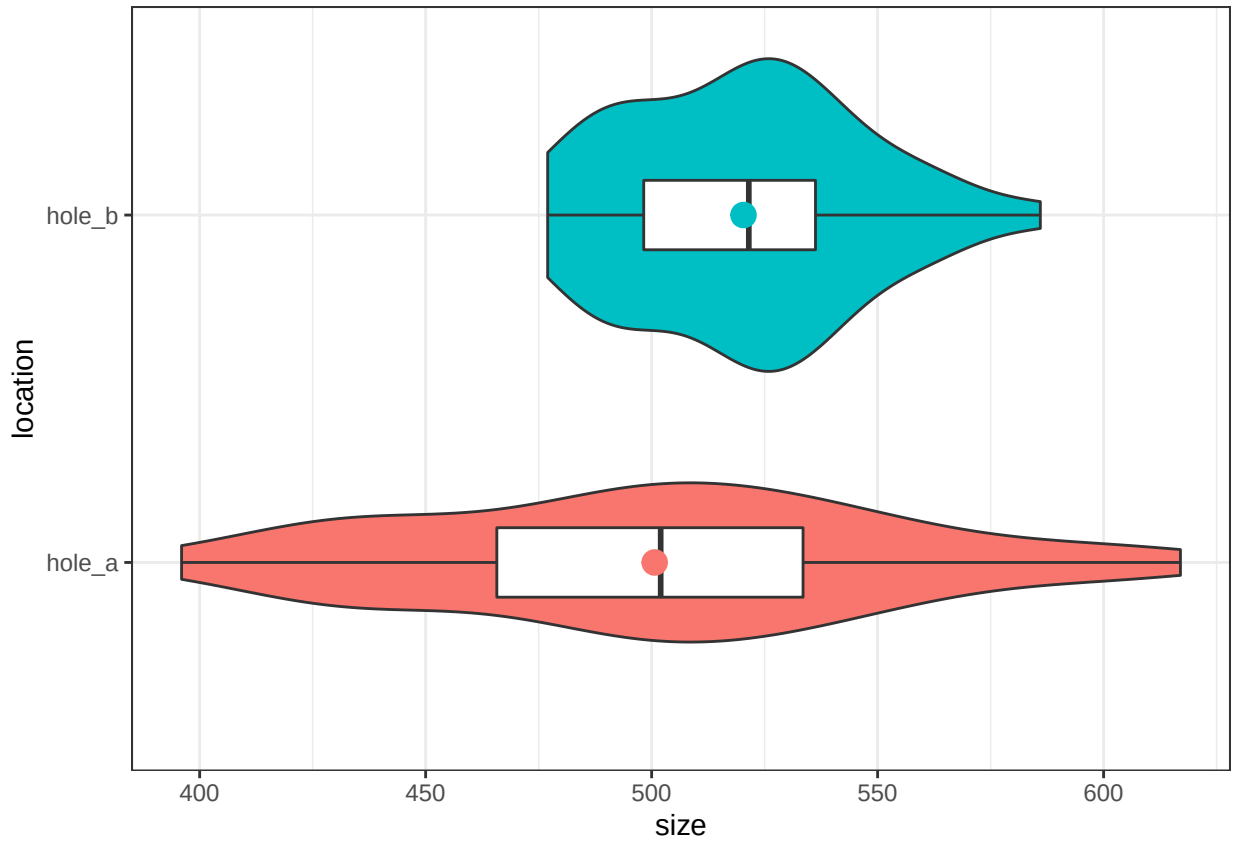
```
rhino <-
    read_csv("data/rhino.csv", show_col_types = FALSE) %>%
    clean_names()

rhino_long <- pivot_longer(rhino, c(hole_a, hole_b),
                           names_to = "location",
                           values_to = "size") %>%
    mutate(rhino = as.character(1:80))

rhino_long
```

```
# A tibble: 80 x 3
   location  size rhino
   <chr>    <dbl> <chr>
 1 hole_a     511 1
 2 hole_b     538 2
 3 hole_a     486 3
 4 hole_b     490 4
 5 hole_a     475 5
 6 hole_b     564 6
 7 hole_a     396 7
 8 hole_b     520 8
 9 hole_a     501 9
10 hole_b     524 10
# ... with 70 more rows
```

```
ggplot(rhino_long, aes(x = location, y = size)) +
    geom_violin(aes(fill = location),) +
    geom_boxplot(width = 0.2) +
    stat_summary(aes(col = location), fun = mean,
                 geom = "point", size = 4) +
    guides(fill = "none", col = "none") +
    coord_flip()
```



The two distributions show no outliers, nor do they show any substantial skew, and although they have wildly different sample variances, this is a balanced design. We should perform a pooled t test, as a result. Remember that we want a 90% confidence interval. Note that the sample mean (shown in the plot with dots) is larger in Hole B than Hole A.

```
t_q15 <- rhino_long %$%
    t.test(size ~ location, var.equal = TRUE,
           conf.level = 0.90)

tidy(t_q15)
```

```
# A tibble: 1 x 10
  estimate estimate1 estimate2 statistic p.value parameter conf.low conf.high
     <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
1    -19.7      501.      520.     -2.10  0.0391        78    -35.2     -4.06
# ... with 2 more variables: method <chr>, alternative <chr>
```

15.1 Results for Question 15 (6 points)

Question 15	Result
% responses that were correct	2
% of available Points Awarded	57

Here were the things I was looking for in evaluating your responses.

1. You were using independent, rather than paired samples.
2. You settled on a pooled t test, either using a regression model or t.test().
3. You estimated the population mean difference as -19.65
4. You estimated the 90% confidence interval as (-35.24, -4.06).
5. You concluded that the rhinos near Hole B were detectably larger than those living near Hole A. You needed to state something like that explicitly ("the effect is negative" without specifying that you're looking at A-B is not enough.)
6. You didn't write anything that was incorrect, anything that was clumsy or unclear or anything that misinterpreted your results.

- If you did all six of those things, you got six points, but only one person did. Another person scored 5.5, while everyone else scored 5 or fewer points.
  - If you switched the sign of the mean difference or the 90% CI, you got half credit on those elements.
  - If you slightly mis-rounded an estimate, you lost 0.5 point.
  - If you used a bootstrap you should have had a different 90% CI than I did, specifically (-34.10, -5.38). If you got that, you lost no points on element 4, just element 2. If you wrote it as (-5.38, -34.10) you lost 0.5 point for that. If you used a seed other than the one we specified, then I wasn't happy about that, but I checked to see if your response matched your seed.
  - If you used the Welch t test instead of the pooled t test, you should have had a 90% CI of (-35.31, -3.99). If you got that, you lost no points on element 4, just element 2.
  - Note that if you said you used a pooled t test but actually obtained the interval using something else, you got no credit for the confidence interval.
  - If you said you paired the data, but you actually didn't, you lost 2 points for that.
  - If you said you paired the data and you actually did, you'd lose most (if not all) of the credit in the question.
  - If you used a 95% confidence interval (or any other level) instead of 90% but the rest of your work was correct, you lost 1 point.
  - If you misspelled something or wrote something grammatically incorrect, you lost at least 0.5 point for that in the "nothing wrong" element. In fact, most people lost 0.5 or 1 point on the "Not writing anything incorrect" element.

The student with a 6 point answer's response was...

> Boxplots of rhino size visually appeared to follow a normal distribution without outliers and the question stem did not indicate intentional pairing was performed when rhinos were sampled, so a two sample T-test was used to build my estimate. The point estimate difference of Hole A - Hole B rhino size was -19.65 with a 90% confidence interval of -35.24 to -4.06 for the difference in size. The model indicates that rhinos sampled from the heavily poached watering Hole A are smaller than rhinos sampled from the less poached watering Hole B. The 90% CI does not include zero so the model would indicate a statistically detectable difference in size of rhinos from Hole A compared to Hole B.

The student with a 5.5 point answer's response (they misrounded the confidence interval estimate) was...

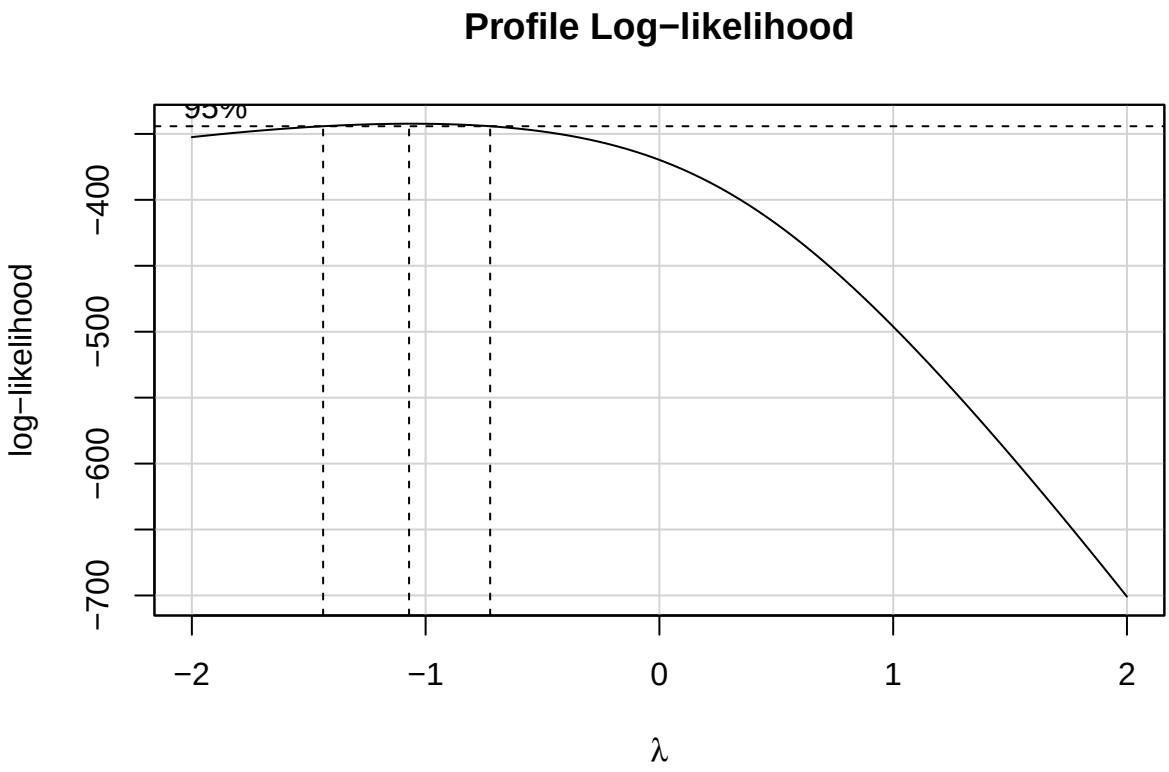
> After viewing the normal Q-Q plots, histograms, and boxplots with violin plots, hole_a and hole_b were assumed to be normally distributed (the deviations in the hole_b plots were minor). The samples are independent because there is no deliberate matching between the two groups, and the samples are equally sized, so a pooled t-test was used. The size outcome is -19.65 with a 90% confidence interval of (-35.25, -4.05). There is a clear difference in the average rhino size: the rhinos near watering hole A have a smaller average size than the rhinos near watering hole B. The confidence interval does not include 0 (entirely negative), so there is a statistically detectable difference (according to this t-test procedure with 90% confidence).

16 Answer 16 is a.

```
score_dat <- read_csv("data/score_dat.csv") %>%
    clean_names()

m16 <- lm(score ~ x1 + x2 + x3 + x4, data = score_dat)

boxCox(m16)
```



The suggested power is clearly near -1, which indicates the inverse of our outcome. That's choice a.

16.1 Results for Question 16 (3 points)

Question 16	Result
% responses that were correct	over 95
% of available Points Awarded	over 95

- No partial credit was available.

17 Answer 17 is 2.07, with 95% CI (1.42, 3.01)

```
hospsim <- read_csv("data/hospsim.csv",
                    show_col_types = FALSE) %>%
    clean_names()

hospsim %$% table(sex, statin)
```

```
        statin
sex       NO YES
  FEMALE 107 305
  MALE    49 289
```

So we need to rearrange the table to get the right odds ratio comparison. I did this with the `twobytwo` function from `Love-boost.R`.

```
twobytwo(289, 49, 305, 107,
         "Male", "Female", "Yes", "No")
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Yes
Comparing : Male vs. Female

       Yes  No    P(Yes) 95% conf. interval
Male   289  49    0.8550    0.8133   0.8887
Female 305 107    0.7403    0.6958   0.7803

                                  95% conf. interval
            Relative Risk: 1.1550    1.0747   1.2413
         Sample Odds Ratio: 2.0691    1.4229   3.0088
Conditional MLE Odds Ratio: 2.0672    1.4034   3.0763
    Probability difference: 0.1147    0.0571   0.1704

             Exact P-value: 0.0001
        Asymptotic P-value: 0.0001
------------------------------------------------------
```

17.1 Results for Question 17 (3 points)

Question 17 Overall	Result
% responses that were correct	87
% of available Points Awarded	92

Point Estimate	Result
% responses that were correct	90
% of available Points Awarded	92

CI Endpoints	Result
% responses that were correct	88
% of available Points Awarded	92

- The point estimate was worth 1 point.
- The confidence interval endpoints were worth 2 points (one point per endpoint.)
- Students who typed in an extra decimal place got full credit.
- Rounding errors in the point estimate (2.06 or 2.08) or endpoints (1.41 or 1.43 on the low end or 3.00 or 3.02 on the high end) got half-credit for responses with that problem.

18 Answer 18 is c.

```
anova(lm(hsgrads ~ insurance, data = hospsim))
```

```
Analysis of Variance Table

Response: hsgrads
           Df Sum Sq Mean Sq F value    Pr(>F)
insurance   3   3580 1193.46  10.794 5.992e-07 ***
Residuals 746  82481  110.56
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov(hsgrads ~ insurance, data = hospsim),
         ordered = TRUE)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = hsgrads ~ insurance, data = hospsim)

$insurance
                          diff       lwr      upr     p adj
UNINSURED-MEDICAID   2.7250163 -2.948734 8.398767 0.6036857
MEDICARE-MEDICAID    5.6336133  2.563514 8.703713 0.0000163
COMMERCIAL-MEDICAID  6.4368450  3.406237 9.467453 0.0000004
MEDICARE-UNINSURED   2.9085971 -2.362698 8.179893 0.4867175
COMMERCIAL-UNINSURED 3.7118288 -1.536564 8.960222 0.2641848
COMMERCIAL-MEDICARE  0.8032317 -1.383020 2.989483 0.7799896
```

The ANOVA F test shows a p value well below 0.05. According to the Tukey HSD comparison, Medicaid is detectably lower than either Commercial or Medicare. That's option c.

18.1 Results for Question 18 (3 points)

Question 18	Result
% responses that were correct	over 95
% of available Points Awarded	over 95

- No partial credit was available.

19 Answer 19 is b.

```
model_q19 <-
    lm(ldl ~ clinic_type + age + sex +
           insurance + hsgrads + a1c + bmi +
           sbp + statin, data = hospsim)

tidy(model_q19, conf.int = TRUE) %>%
    select(term, estimate, conf.low, conf.high, p.value) %>%
    kable(digits = 2)
```

term	estimate	conf.low	conf.high	p.value
(Intercept)	76.18	34.18	118.17	0.00
clinic_typeOLD	9.30	3.20	15.40	0.00
age	-0.21	-0.55	0.13	0.23
sexMALE	-7.77	-13.18	-2.36	0.00
insuranceMEDICAID	-2.25	-11.18	6.69	0.62
insuranceMEDICARE	-4.01	-10.84	2.83	0.25
insuranceUNINSURED	12.57	-1.55	26.69	0.08
hsgrads	0.07	-0.18	0.33	0.57
a1c	2.48	1.12	3.84	0.00
bmi	-0.37	-0.74	-0.01	0.05
sbp	0.27	0.09	0.45	0.00
statinYES	-10.39	-17.20	-3.59	0.00

```
glance(model_q19) %>%
    select(r.squared)
```

```
# A tibble: 1 x 1
  r.squared
      <dbl>
1    0.0921
```

Older Clinics have detectably higher LDL cholesterol after accounting for all other variables, thanks to the exclusively positive 95% confidence interval for the slope of the indicator variable clinic_typeOLD. The overall R-square is 9.2%, well less than 20%. So the correct answer is b.

19.1 Results for Question 19 (3 points)

Question 19	Result
% responses that were correct	88
% of available Points Awarded	92

- I gave one point to choice c and also 1 point to choice d for getting part of the answer right.
- The most common incorrect response was d.

20 Answer 20 is e.

The smaller (stepwise) model still has 7 predictors. The only variables that drop out from the full model are hsgrads and age.

```
step(model_q19)
```

```
Start:  AIC=5402.89
ldl ~ clinic_type + age + sex + insurance + hsgrads + a1c + bmi +
    sbp + statin

              Df Sum of Sq    RSS    AIC
- hsgrads      1     420.3 977110 5401.2
- age          1    1888.9 978579 5402.3
- insurance    3    7305.6 983995 5402.5
<none>                     976690 5402.9
- bmi          1    5279.5 981969 5404.9
- sex          1   10517.7 987207 5408.9
- sbp          1   11091.1 987781 5409.4
- clinic_type  1   11874.6 988564 5410.0
- statin       1   11892.7 988582 5410.0
- a1c          1   16875.3 993565 5413.7

Step:  AIC=5401.21
ldl ~ clinic_type + age + sex + insurance + a1c + bmi + sbp +
    statin

              Df Sum of Sq    RSS    AIC
- age          1    1820.3 978930 5400.6
- insurance    3    7411.8 984522 5400.9
<none>                     977110 5401.2
- bmi          1    5217.7 982328 5403.2
- sex          1   10176.9 987287 5407.0
- sbp          1   10808.7 987919 5407.5
- clinic_type  1   11501.0 988611 5408.0
- statin       1   12105.1 989215 5408.4
- a1c          1   16683.1 993793 5411.9

Step:  AIC=5400.61
ldl ~ clinic_type + sex + insurance + a1c + bmi + sbp + statin

              Df Sum of Sq    RSS    AIC
<none>                     978930 5400.6
- bmi          1    4078.3 983009 5401.7
- insurance    3   12120.4 991051 5403.8
- sex          1    9592.1 988522 5405.9
- sbp          1    9789.2 988720 5406.1
- clinic_type  1   13593.0 992523 5408.9
- statin       1   15069.2 994000 5410.1
- a1c          1   19161.7 998092 5413.1

Call:
lm(formula = ldl ~ clinic_type + sex + insurance + a1c + bmi +
    sbp + statin, data = hospsim)
```

```
Coefficients:
       (Intercept)      clinic_typeOLD             sexMALE   insuranceMEDICAID
           70.9329              9.7166             -7.3483             -1.2067
 insuranceMEDICARE  insuranceUNINSURED                 a1c                 bmi
           -6.1510             12.8761              2.5997             -0.3167
               sbp           statinYES
            0.2470            -11.3866
```

Comparing the two models, the seven-predictor (smaller) model has a higher adjusted R^2, and lower AIC and BIC, so it shows better performance using all three measures, which is option `e`.

```
model_smaller <-
    lm(ldl ~ clinic_type + sex +
           insurance + a1c + bmi +
           sbp + statin, data = hospsim)

bind_rows(
    glance(model_q19),
    glance(model_smaller)) %>%
    mutate(model = c("full", "smaller")) %>%
    select(model, adj.r.squared, AIC, BIC) %>%
    kable()
```

model	adj.r.squared	AIC	BIC
full	0.0785625	7533.296	7593.357
smaller	0.0789447	7531.015	7581.836

20.1 Results for Question 20 (3 points)

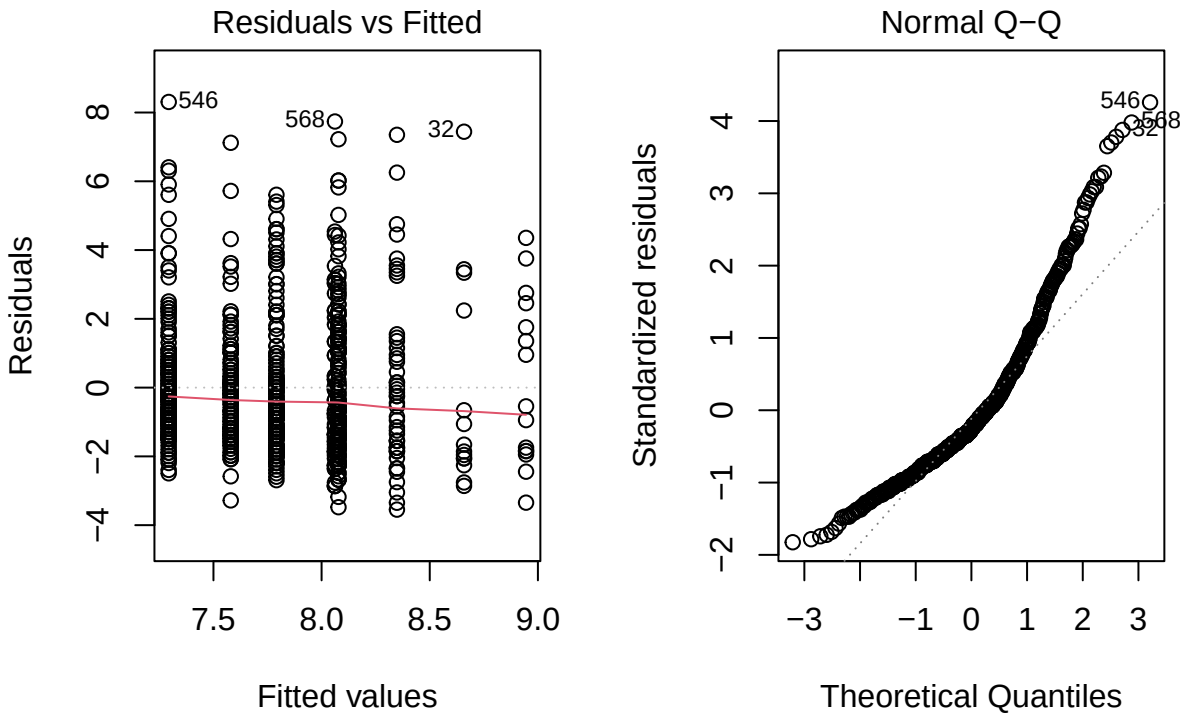
Question 20	Result
% responses that were correct	80
% of available Points Awarded	83

- I gave one point to students who chose `c`, since that was more correct than the other incorrect responses, essentially.
- The most common incorrect response was `c`.

21 Answer 21 is `b`.

```
model_q21 <- lm(a1c ~ sex + insurance, data = hospsim)

par(mfrow=c(1,2)); plot(model_q21, which = 1:2);
```




```
par(mfrow=c(1,1))

hospsim %>% slice(546)
```

```
# A tibble: 1 x 11
  subject_id clinic_type   age sex    insurance hsgrads   a1c   ldl   bmi   sbp
  <chr>      <chr>       <dbl> <chr>  <chr>       <dbl> <dbl> <dbl> <dbl> <dbl>
1 X1546      OLD            28 FEMALE MEDICARE     89.9  15.6   179  25.1   118
# ... with 1 more variable: statin <chr>
```

This is subject X1546, who is a Female Medicare patient visiting an OLD clinic. That's option `b`.

21.1 Results for Question 21 (3 points)

Question 21	Result
% responses that were correct	93
% of available Points Awarded	93

- No partial credit was available.
- The most common incorrect response was `e`.

22 Answer 22 is `a`, `b`, `c` and `e`.

All of these are useful except the adjusted R^2 in the training sample, which is a measure of performance within the sample, but doesn't tell us anything directly about predictions out of sample.

22.1 Results for Question 22 (3 points)

Question 22	Result
% responses that were correct	12
% of available Points Awarded	61

- Four of the responses were correct. You got 0.75 point for each correct response you selected, so long as you didn't choose option d.
- You received no credit on the question if you selected `d`, either alone or with other responses.
- The most common response was `a`, `b` and `c`, forgetting `e`.

23 Answer 23 is `b`.

The essential problem here is indicated by the curve in the residuals vs. fitted values plot on the top left. That's a problem with the linearity assumption, and our next step should be to consider transformations of the outcome (and if that doesn't work, perhaps transformations of the predictors.)

While there is an influential point here, that may no longer be true after the transformation, and also you wouldn't want to simply drop a point because it was influential.

23.1 Results for Question 23 (3 points)

Question 23	Result
% responses that were correct	77
% of available Points Awarded	77

- No partial credit was available.
- The most common incorrect response was `a`.

24 Answer 24 is `d` and `e`.

Approaches `d` and `e` accomplish the specified task, as shown below. The others do not. Note in particular that option `a` would create a test sample with 80% of the observations and a training sample with the remaining 20%, which is the opposite of what we're trying to do. Also note that option `b` would not select a random sample of observations, but rather just the first 100 observations, to be in the test sample. Option `c` is wrong because there's no `partition()` function.

```
mydat <- tibble(x1 = rnorm(n = 500, mean = 100, sd = 1),
                x2 = rnorm(n = 500, mean = 200, sd = 2),
                x3 = rnorm(n = 500, mean = 100, sd = 3),
                x4 = rnorm(n = 500, mean = 200, sd = 4),
```

```
                x5 = rnorm(n = 500, mean = 100, sd = 5),
                y = rnorm(n = 500, mean = 25, sd = 4))

nrow(mydat)
```

```
[1] 500
```

```
## approach d

mydat_test <- slice_sample(mydat, n = 100)
mydat_train <- anti_join(mydat, mydat_test)
```

```
Joining, by = c("x1", "x2", "x3", "x4", "x5", "y")
```

```
nrow(mydat_test); nrow(mydat_train)
```

```
[1] 100
```

```
[1] 400
```

```
## approach e

mydat_train <- slice_sample(mydat, prop = 0.80)
mydat_test <- anti_join(mydat, mydat_train)
```

```
Joining, by = c("x1", "x2", "x3", "x4", "x5", "y")
```

```
nrow(mydat_test); nrow(mydat_train)
```

```
[1] 100
```

```
[1] 400
```

a. `mydat_test <- slice_sample(mydat, prop = 0.80)` and `mydat_train = anti_join(mydat, mydat_test)`
b. `mydat_test <- slice_head(mydat, 100)` and `mydat_train = anti_join(mydat, mydat_test)`
c. `mydat_test <- partition(mydat, 400:100)` and `mydat_train = anti_join(mydat, mydat_test)`
d. `mydat_test <- slice_sample(mydat, n = 100)` and `mydat_train = anti_join(mydat, mydat_test)`
e. `mydat_train <- slice_sample(mydat, prop = 0.80)` and `mydat_test = anti_join(mydat, mydat_train)`
f. None of these approaches would work.

24.1 Results for Question 24 (3 points)

Question 24	Result
% responses that were correct	72
% of available Points Awarded	82

- Students choosing only `d` or only `e` got 1.5 points.
- People choosing `d`, `e` and something that was incorrect got 1 point.
- The most common incorrect response was `b`, `d` and `e`.

25 Answer 25 estimate is 0.520, with 90% CI (0.471, 0.568)

```
survA <- read_csv("data/survA.csv",
                      show_col_types = FALSE) %>%
    clean_names()

survA %>% tabyl(statfuture, year)
```

```
 statfuture 2014 2015 2016 2017 2018 2019 2020 2021
          3    0    1    0    0    0    0    0    0
          4    1    1    1    0    0    0    2    0
          5    8   10    7    8    6    4    9    8
          6   14   18   19   10   14   14   28   18
          7   18   19   37   30   31   42   28   28
         NA    1    0    0    0    0    1    0    4
```

So 2015-2018 and 2020 are the years without missing responses on `statfuture`.

```
q25 <- survA %>%
    filter(year %in% c(2015:2018, 2020))

q25 %>% count(year)
```

```
# A tibble: 5 x 2
   year     n
  <dbl> <int>
1  2015    49
2  2016    64
3  2017    48
4  2018    51
5  2020    67
```

```
q25 %>% count(statfuture == 7)
```

```
# A tibble: 2 x 2
  `statfuture == 7`     n
  <lgl>             <int>
1 FALSE               134
2 TRUE                145
```

So, 145 out of $(145 + 134 = 279)$ or 51.97% of the `statfuture` responses in 2015-18 and 2020 were 7.

```
mosaic::binom.test(x = 145, n = 279, conf.level = 0.90,
                   ci.method = "Agresti-Coull")
```

```
    Exact binomial test (Agresti-Coull CI)

data:  145 out of 279
number of successes = 145, number of trials = 279, p-value = 0.5495
alternative hypothesis: true probability of success is not equal to 0.5
90 percent confidence interval:
 0.4705609 0.5684870
sample estimates:
probability of success
             0.5197133
```

25.1 Results for Question 25 (3 points)

Question 25 Overall	Result
% responses that were correct	53
% of available Points Awarded	70

Point Estimate	Result
% responses that were correct	72
% of available Points Awarded	74

CI Endpoints	Result
% responses that were correct	55
% of available Points Awarded	68

- The point estimate was worth 1 point.
- The confidence interval endpoints were worth 2 points (one point per endpoint.)
- Students who typed in an extra decimal place got full credit.
- Rounding errors in the point estimate (0.519) or endpoints (0.470 on the low end or 0.569 on the high end) got half-credit for responses with that problem.

26 Answer 26 is a.

How about the 2021 data?

```
survA %>% filter(year == 2021) %>%
    count(statfuture == 7)
```

```
# A tibble: 3 x 2
  `statfuture == 7`     n
  <lgl>             <int>
1 FALSE                26
2 TRUE                 28
3 NA                    4
```

Ignoring the missing values, we have a sample proportion of $28/(28 + 26) = 28/54 = 0.519$, which is within the 90% CI presented in response to Question 25. So that's choice `a`.

26.1 Results for Question 26 (3 points)

Question 26	Result
% responses that were correct	over 95
% of available Points Awarded	over 95

- No partial credit was available.

27 Answer 27 is 265 subjects.

See Answer 28 for the R code I used. Here are the important points.

- This is a paired samples study.
- We're comparing means, so we need `power.t.test`.
- The observed effect size was 10, so we need at least half that for our `delta`. (Note that whether you use -5 or +5 for delta has no effect.)
- The standard deviation in the sample set of differences was 20 (or technically 19.99). Increasing that by 20% means 19.99(1.25) = 24.9875. (If you use 25 instead you get the same conclusion.)
- We want a 95% confidence level, so we need sig.level = 0.05
- We want 90% power, so power = 0.90.

We obtain a paired t test with 265 observations (always round up to ensure we have the sample size necessary) and we realize that while each of those 265 subjects will be measured twice, that doesn't make them new subjects, so 265 is the minimum number of subjects.

27.1 Results for Question 27 (3 points)

Question 27	Result
% responses that were correct	68
% of available Points Awarded	70

- Students who wrote "265 pairs" got 2 points out of 3.
- Students who wrote "530 subjects" got 1 point.

28 Answer 28 is the following R code.

```
power.t.test(delta = -5, sd = 19.99*1.25,
             sig.level = 0.05, power = 0.90,
             type = "paired")
```

```
     Paired t test power calculation

              n = 264.3512
          delta = 5
             sd = 24.9875
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

28.1 Results for Question 28 (3 points)

Question 28	Result
% responses that were correct	72
% of available Points Awarded	77

- Those who had the right test (paired samples) and delta, significance level and power specifications but incorrectly specified the standard deviation got 2 points.
- Those who did everything right except specified the wrong significance level got 2 points.
- Those who used "one.sample" or "paired" appropriately but misspecified the delta got 1 point.
- Those who used independent samples instead of the correct paired samples approach got no credit, even if you did it correctly for an independent samples setting. This includes those of you who actively didn't specify the type of test in your power.t.test statement.
- Those who included n and tried to estimate the power got no credit here, either.

29 Answer 29 is that none of the statements are appropriate.

- Statement a is not appropriate. The confidence interval is a statement about a population mean, not about individual subjects (in this case, not about individual cows.)
- Statement b is not appropriate for the same reason as statement a. The confidence interval is a statement about a population mean, not about individual cows.
- Statement c is not appropriate. We know with 100% certainty that the average weight gain in this study was 56 pounds. 95% sure is not the same thing as 100% sure, and this is not an appropriate interpretation of a confidence interval, although I understand that some of you likely figured "well, if we're 100% sure, we're at least 95% sure".
- Statement d is not appropriate either. The average weight gain of cows fed this supplement does not vary. It's what we are trying to estimate.
- Statement e is not appropriate. There is not a 95% chance for another sample to have an average weight gain between 45 and 67 pounds. There is a 95% chance that another sample will have its average weight gain within two standard errors of the true (population) mean.

An appropriate statement would have been that we were 95% confident that the true (population) mean weight gain for all cows fed this supplement would be between 45 and 67 pounds. Caution: this probably wasn't a random sample of cows. It was more likely to be a convenience (or convenient) sample, making us a bit reluctant to extend this inference to other situations.

29.1 Results for Question 29 (5 points)

Question 29	Result
% responses that were correct	10
% of available Points Awarded	69

- There were 5 parts (a-e) to this Question. You got one point for each correct (i.e. 2. Not appropriate) response.

Part	a	b	c	d
% correct	95	83	55	70

30 Answer 30 is an essay.

We don't write sketches for essay questions.

The essay topics for 2021 (as summarized by me) were:

- Accessibility of Findings (3)

- Algorithmic Accuracy (3)
- Algorithmic Transparency (1)
- Bayesian Methods (4)
- Classification Trees (1)
- Confounding / Lurking Factors (3)
- Deduction vs. Induction (1)
- Ethics and Transparency (4)
- Exploration vs. Confirmation (2)
- Inductive Inference (1)
- Intention to Treat (1)
- Measurement Bias (2)
- Media and Communication (5)
- Meta-Analysis (1)
- Metaphorical Populations (1)
- Mister P (Multi-level regression and post-stratification) (2)
- Over-fitting (1)
- P-hacking (1)
- Post-Data Hypothesizing (or HARKing) (1)
- Prosecutor's Fallacy (1)
- Publication Bias (2)
- Questionable Research Practices (5)
- Receiver Operating Characteristic Curves (1)
- Reproducibility Crisis (6)
- Simpson's Paradox (2)
- Skepticism and its Importance (1)
- Social Physics (1)
- SPRT (1)
- Statistical "Humility" (1)

I gave this question in Fall 2020 as well, and there the essay topics were (as summarized by me):

- Algorithmic Transparency (2)
- Ascertainment Bias
- Association vs Causation (4)
- Bayes factors
- Bayes' Theorem
- Bayesian "smoothing"
- Bias/Variance Tradeoff
- Bradford-Hill Criteria (2)
- Causality and Data Reversibility
- Central Limit Theorem
- Classification Trees (2)
- Communication & Transparency (5)
- Communication Pipeline (3)
- Complexity vs Performance
- Confounding
- Convenience Sampling
- Data Ethics
- Designing Observational Studies
- Ethics / Statistical Fraud
- Evaluating Algorithms (3)
- Evaluating Published Claims (2)
- Exploration vs. Confirmation
- External Validity
- Feature Engineering (2)

- Making Daily Decisions
- Metaphorical Populations
- Mixing Fisher and Neyman-Pearson
- Neural Networks
- Over-fitting
- P-hacking
- Poisson Distribution (2)
- Positive vs. Negative Framing
- PPDAC cycle
- Pre-registration of analyses
- Publication Bias (3)
- Questionable Research Practices / Researcher Degrees of Freedom (7)
- Randomized Clinical Trials
- Regression to the Mean (2)
- Reproducibility Crisis (4)
- ROC curves
- Simpson's Paradox

That's a handy index of things I would hope you would have a better understanding of after reading Spiegelhalter.

30.1 Results for Question 30 (10 points)

This was originally meant to be an 8 point question, but I decided to grade it on a scale from 0 to 10 points, although the highest score I gave was actually 9. Most scores on Question 30 were between 6 and 8.5 points. I'm still treating the Quiz as if it were out of 100 points, not 102.

I reviewed the essays with regard to the following issues:

1. Did the essay meet the general specifications, regarding length and topic?
2. Were there multiple problems with grammar/syntax/spelling?
3. Was the topic of the essay clear to me on a single reading?
4. Was the topic something that is discussed in The Art of Statistics?
5. Had we discussed the topic of the essay meaningfully in class?
6. What was the overall quality of the essay?

Assuming items 1-4 went fairly well, my reactions to items 5 and 6 led me to your grade. If items 1-4 above didn't go so well, then you lost somewhere between 0.5 and 3 points for those concerns.

- A score of 10 in my mind would be an essay that was essentially flawless, which is probably too tough a standard in this context.
- A score of 9 was the best I actually gave, and indicated a really good essay about a topic we hadn't really discussed, or an excellent essay about something we had discussed.
- A score of 8 would usually be a good essay about a topic we hadn't really discussed, or a very good essay about something we had discussed.
- A score of 7 would usually be an OK essay about a topic we hadn't really discussed, or a good essay about something we had discussed.
- A score below 7 would usually indicate some substantial problems with items 1-4 above, or that the essay itself just wasn't very effective or clear.

31 Results Across All Items

Take your raw score, obtained by adding up all of your points on Questions 1-30.

Then add 5 points. That's your score on Quiz 3, and the median score is thus 90.

Here's the distribution of that Quiz 3 score.

Scoring Range	Students
95 or above	14
90 to 94.75	17
85 to 89.75	13
80 to 84.75	4
75 to 79.75	5
70 to 74.5	some
Below 70	a few