Answer Sketch for Homework G

431 Staff and Professor Love

*Due **2019-11-01** at 2 PM. Last Edited 2019-10-02 11:59:36*

Contents

R Setup

Here's the complete R setup we used.

```
knitr::opts_chunk$set(comment=NA)
options(width = 60)

library(here); library(janitor); library(broom)
library(PropCIs); library(exact2x2); library(Epi)
library(patchwork); library(magrittr); library(tidyverse)

source(here("R", "Love-boost.R"))

q1_raw <- read_csv(here("data", "hwG_q1.csv")) %>%
  clean_names()

q2_raw <- read_csv(here("data", "hwG_q2.csv")) %>%
  clean_names()
```

Question 1

Here's the raw data, as imported. We have two rows for each subject, although sometimes the "at follow up" assessment is missing (see S-003 for instance.)

```
q1_raw
```

```
# A tibble: 550 x 3
   subject_id assessment     phq9_score
   <chr>      <chr>               <dbl>
 1 S-001      start of study         16
 2 S-001      at follow up           19
 3 S-002      start of study         17
 4 S-002      at follow up           13
 5 S-003      start of study         20
 6 S-003      at follow up           NA
 7 S-004      start of study         10
 8 S-004      at follow up           11
 9 S-005      start of study         14
10 S-005      at follow up           10
# ... with 540 more rows
```

Part A

The outcome is PHQ-9 score. I suppose you could say that the key outcome is each subject's *change* in PHQ-9 score from baseline to follow-up, if you like.

Part B

The exposure groups being compared are

1. patients before the administration of the sertraline medication regimen, and
2. then those same patients again after completing the regimen.

Part C

The data were collected using paired samples, where each PHQ-9 measure is part of a pair of assessments for the same subject. We will want to rearrange the data, which were provided to us in "long" format to a "wider" structure for calculation and plotting of paired differences in PHQ-9 score, etc. We'll also see that we have missing "follow-up" values for some subjects, which we'll deal with by dropping those subjects (since we want a complete cases analysis.)

Specifically, here's what I did to manage the data.

```
q1_wider <- q1_raw %>%
  pivot_wider(names_from = assessment,
              values_from = phq9_score)

q1_wider
```

```
# A tibble: 275 x 3
   subject_id `start of study` `at follow up`
   <chr>                 <dbl>          <dbl>
 1 S-001                    16             19
 2 S-002                    17             13
 3 S-003                    20             NA
 4 S-004                    10             11
 5 S-005                    14             10
 6 S-006                    16             11
 7 S-007                    23             18
 8 S-008                    15              6
 9 S-009                    10              7
10 S-010                     7              5
# ... with 265 more rows
```

```
summary(q1_wider)
```

```
  subject_id        start of study   at follow up
 Length:275         Min.   : 0.00   Min.   : 0.00
 Class :character   1st Qu.: 9.00   1st Qu.: 3.00
 Mode  :character   Median :12.00   Median : 8.50
                    Mean   :12.16   Mean   : 8.75
                    3rd Qu.:15.00   3rd Qu.:13.00
                    Max.   :26.00   Max.   :24.00
                                    NA's   :15
```

I'm going to rename the two columns with spaces in their names, and then drop the cases with missing data at follow up. Then, I'll calculate the paired (baseline - follow up) differences, and place them in a variable called `PHQ9_diff`. Improvements (reductions) in PHQ-9 scores will be represented by positive `PHQ9_diff` values.

```
q1_wider <- q1_wider %>%
  rename(baseline = "start of study", follow = "at follow up") %>%
  drop_na %>%
  mutate(PHQ9_diff = baseline - follow)

q1_wider
```

```
# A tibble: 260 x 4
   subject_id baseline follow PHQ9_diff
   <chr>         <dbl>  <dbl>     <dbl>
```

```
 1 S-001            16     19        -3
 2 S-002            17     13         4
 3 S-004            10     11        -1
 4 S-005            14     10         4
 5 S-006            16     11         5
 6 S-007            23     18         5
 7 S-008            15      6         9
 8 S-009            10      7         3
 9 S-010             7      5         2
10 S-011            14     10         4
# ... with 250 more rows
```

OK. We have 260 paired differences now, with no missing values, and an average reduction of 3.3 points on the PHQ-9 scale from baseline to follow-up:

```
mosaic::favstats(~ PHQ9_diff, data = q1_wider)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

```
 min Q1 median Q3 max     mean       sd   n missing
  -7  1      3  6  18 3.315385 4.056616 260       0
```

Part D

Professor Love provided no substantial information to address the issue of whether the samples were taken at random from the population of patients at the six clinics involved. There is no reason provided that justifies the assumption of random sampling. So whether this sample can be generalized effectively to the population of all patients at the six clinics involved is unclear. He also provided no details on what the characteristics of the patients were that included them as "participants." We need more information to answer this question.

Part E

The significance level should be $\alpha = 0.10$ so we'll create a 90% confidence interval.

Part F

The PI wants a two-sided confidence interval.

Part G

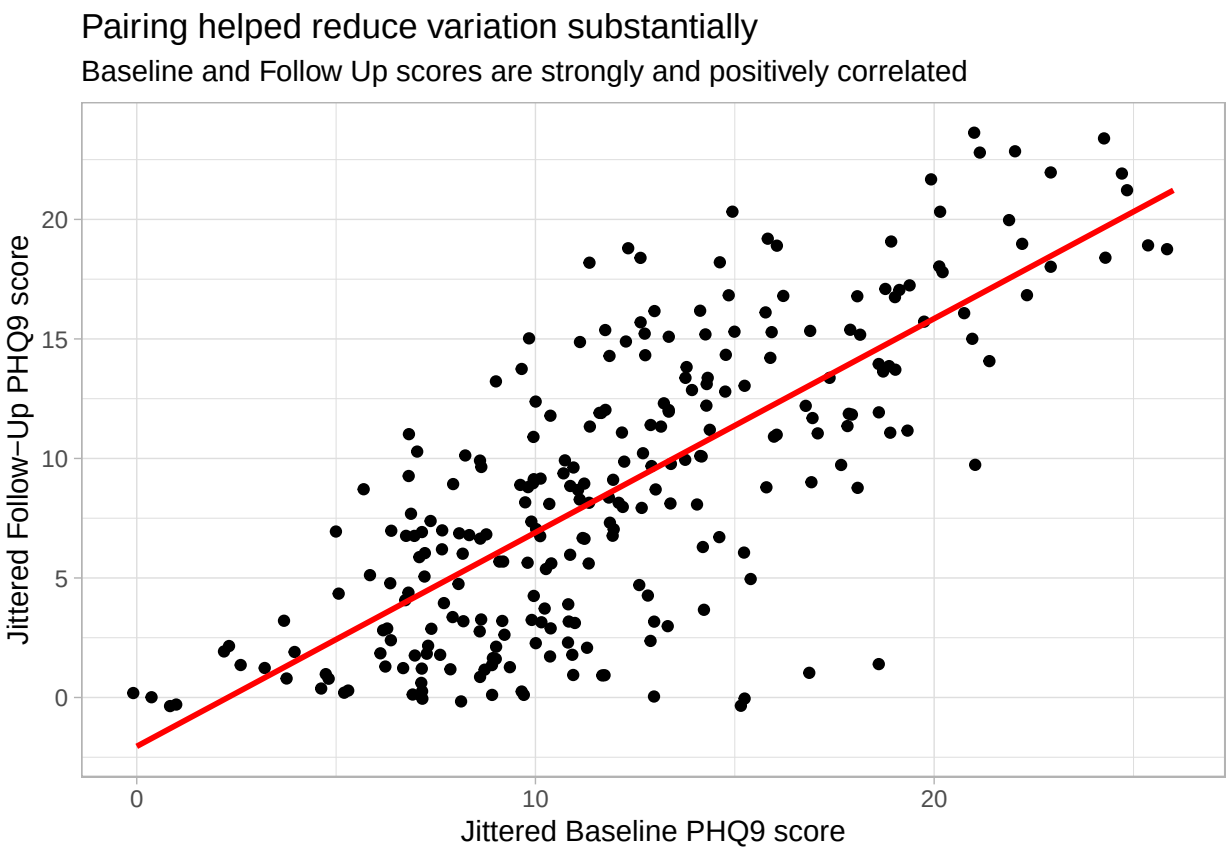
We have paired samples. Pairing helped because the correlation of the PHQ-9 scores before the study and after the study (for the subjects with data at both time points) is quite large and positive, at about 0.75.

```
q1_wider %$% cor(baseline, follow)
```

```
[1] 0.7499678
```

```
ggplot(q1_wider, aes(x = baseline, y = follow)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  theme_light() +
```

```
  labs(title = "Pairing helped reduce variation substantially",
       subtitle = "Baseline and Follow Up scores are strongly and positively correlated",
       x = "Jittered Baseline PHQ9 score",
       y = "Jittered Follow-Up PHQ9 score")
```



Part H

We have paired samples. The plots of paired differences presented below suggest that a Normal model is probably a reasonable choice in this situation, so that if we want to build a confidence interval for the difference in means, we can likely use a t-test approach fairly safely.

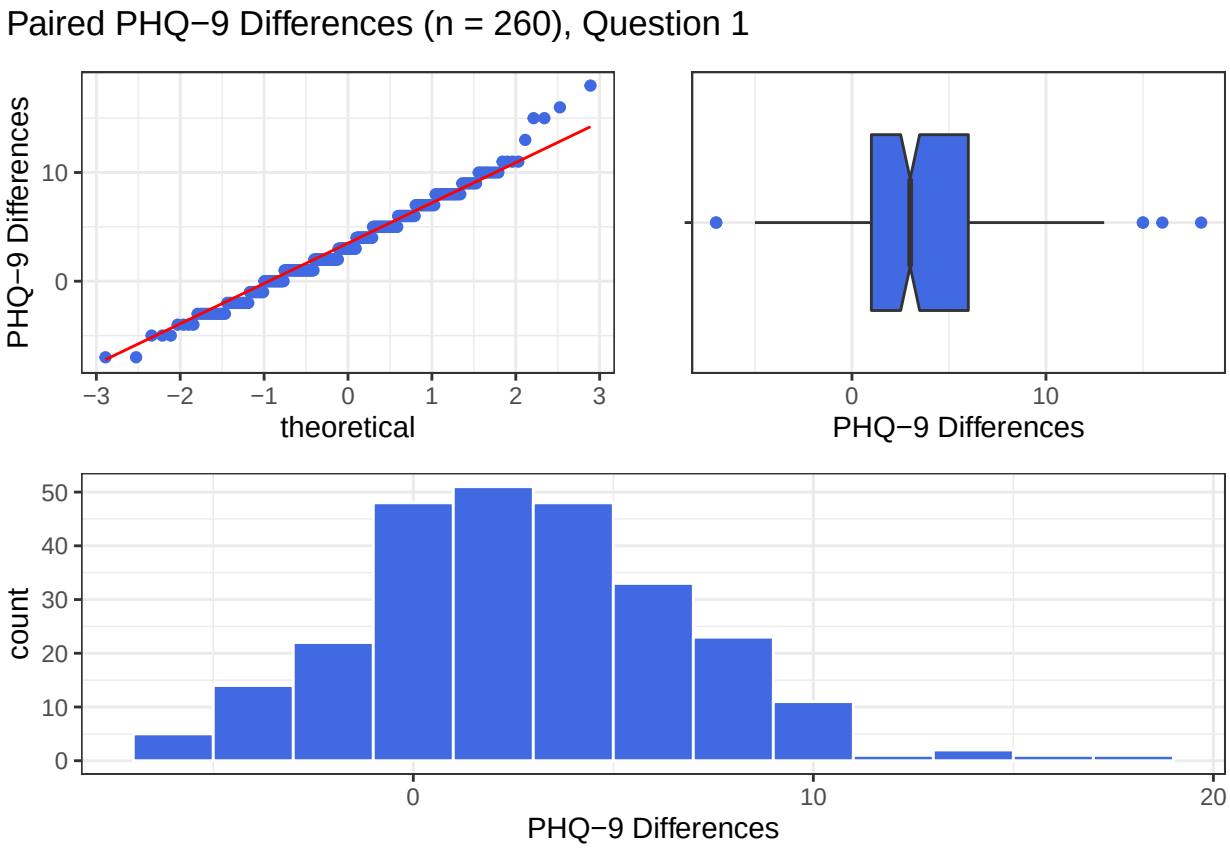
```
p1 <- ggplot(q1_wider, aes(sample = PHQ9_diff)) +
  geom_qq(col = "royalblue") + geom_qq_line(col = "red") +
  theme_bw() +
  labs(y = "PHQ-9 Differences")

p2 <- ggplot(q1_wider, aes(x = "", y = PHQ9_diff)) +
  geom_boxplot(width = 0.7, fill = "royalblue",
               outlier.color = "royalblue",
               notch = TRUE) +
  coord_flip() +
  labs(x = "", y = "PHQ-9 Differences") +
  theme_bw()

p3 <- ggplot(q1_wider, aes(x = PHQ9_diff)) +
```

```
  geom_histogram(fill = "royalblue", col = "white",
                 binwidth = 2) +
  theme_bw() +
  labs(x = "PHQ-9 Differences")

p1 + p2 - p3 +
  plot_layout(ncol = 1) +
  plot_annotation(title = "Paired PHQ-9 Differences (n = 260), Question 1")
```



Part I

These are not independent samples, so we skip this part of the question.

Part J

The 90% confidence interval based on the t test is an improvement (reduction) of (2.9, 3.7) points on the PHQ-9 scale (which ranges from 0 to 27.) Our point estimate is 3.3 points of improvement on that scale.

```
model1 <- lm(PHQ9_diff ~ 1, data = q1_wider)

tidy(model1, conf.int = TRUE, conf.level = 0.90) %>%
  select(estimate, std.error, conf.low, conf.high)
```

```
# A tibble: 1 x 4
```

```
  estimate std.error conf.low conf.high
     <dbl>     <dbl>    <dbl>     <dbl>
1     3.32     0.252     2.90      3.73
```

What if we used a bootstrap instead?

I wouldn't have, but you might have argued for a bootstrap approach, in which case your results would have looked something like this, if you used `431` as your seed.

```
set.seed(431)
q1_wider %$% Hmisc::smean.cl.boot(PHQ9_diff)
```

```
    Mean    Lower    Upper 
3.315385 2.834615 3.796250 
```

The bootstrap confidence interval I obtained is pretty close to the interval I got from the t test, although it's a bit wider.

Question 2.

Here's the raw data, as imported. We have one row for each county, although we will need to collapse the `county_type` information into a comparison of rural vs. urban counties. The rural counties include those with `county_type` of micropolitan or noncore, and the urban counties will include the other four `county_type` groups.

```
q2_raw
```

```
# A tibble: 487 x 3
   county_number county_type  drive_in_minutes
   <chr>         <chr>                   <dbl>
 1 C_001         micropolitan               42
 2 C_002         noncore                    31
 3 C_003         noncore                    54
 4 C_004         small metro                25
 5 C_005         medium metro               17
 6 C_006         small metro                 5
 7 C_007         noncore                    63
 8 C_008         small metro                26
 9 C_009         micropolitan               36
10 C_010         noncore                    55
# ... with 477 more rows
```

Part A

The outcome is drive time in minutes to the nearest certified opioid treatment program from the center of the county.

Part B

The exposure groups being compared are

1. urban counties (those with `county_type` of small metro, medium metro, large fringe metro or large central metro)

2. rural counties (those with `county_type` of micropolitan or noncore)

Part C

The data were collected using independent samples of 217 urban and 270 rural counties. The data are in the correct (long) format for analysis of independent samples, but, as mentioned, we need to collapse the levels of `county_type` to specify our two exposure groups.

Specifically, here's what I did to manage the data.

```
q2_better <- q2_raw %>%
  mutate(exposure = fct_collapse(county_type,
                                 rural = c("micropolitan",
                                           "noncore"),
                                 urban = c("large central metro",
                                           "large fringe metro",
                                           "medium metro",
                                           "small metro")))

q2_better %>% tabyl(county_type, exposure) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_title()
```

```
                    exposure
         county_type    urban rural Total
 large central metro       25     0    25
  large fringe metro       42     0    42
        medium metro       50     0    50
        micropolitan        0   110   110
             noncore        0   160   160
         small metro      100     0   100
               Total      217   270   487
```

OK. We have 217 urban and 270 rural counties. Do we have complete data on `driving_time` in each group?

```
mosaic::favstats(drive_in_minutes ~ exposure, data = q2_better)
```

```
  exposure min Q1 median Q3 max     mean       sd   n
1    urban   1 15     24 32  86 26.74654 19.13755 217
2    rural   1 37     44 55 100 45.82593 16.98817 270
  missing
1       0
2       0
```

Looks good. The sample mean time in the rural group of counties is much longer, it appears, than in the urban group of counties.

Part D

With two exceptions (according to the footnote), we have data here from all of the counties in 5 states. This isn't a random sample of counties. It's (essentially) all of them from five states hit hard by the opioid epidemic. It is probably reasonable to generalize from this sample to the population of those states, but they are very different from other states, potentially.

Part E

The original paper used 95% confidence, so the significance level should be $\alpha = 0.05$.

Part F

The PI wants a one-sided confidence interval, as they ask only for whether times are detectably longer in rural than in urban counties.

Part G

These are not paired samples, so we skip this part of the question.

Part H

These are not paired samples, so we skip this part of the question.

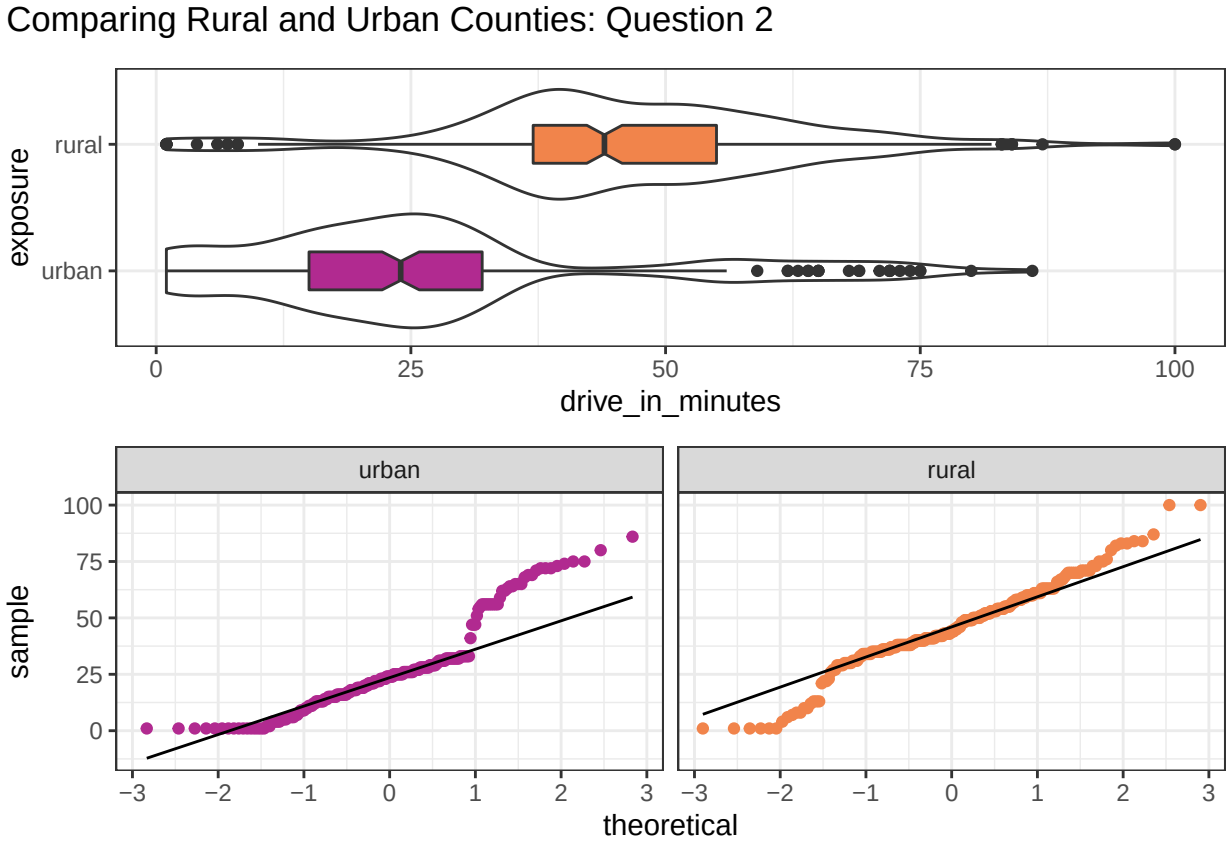
Part I

Here are some plots comparing the driving times for rural vs. urban counties.

```
p1 <-
  ggplot(q2_better, aes(x = exposure, y = drive_in_minutes)) +
  geom_violin() +
  geom_boxplot(aes(fill = exposure), notch = TRUE, width = 0.3) +
  coord_flip() +
  scale_fill_viridis_d(option = "C", begin = 0.4, end = 0.7) +
  guides(fill = FALSE) +
  theme_bw()

p2 <-
  ggplot(q2_better, aes(sample = drive_in_minutes)) +
  geom_qq(aes(col = exposure)) + geom_qq_line(col = "black") +
  facet_wrap(~ exposure) +
  scale_color_viridis_d(option = "C", begin = 0.4, end = 0.7) +
  guides(col = FALSE) +
  theme_bw()

p1 + p2 + plot_layout(ncol = 1) +
  plot_annotation(title = "Comparing Rural and Urban Counties: Question 2")
```



I don't think you can make a good case for treating these as well described by a Normal model. The rural data appears heavy-tailed, and the urban data appears to have a heavy right tail in particular, and may be a bit right skewed (although the mean and median are close.) I'd suggest that a bootstrap may be a better choice for comparing means here.

Part J

A one-tailed 95% confidence interval using the bootstrap may be found by calculating the 90% two-sided confidence interval, and using its upper bound. Using `431` as my seed, I get:

```
set.seed(431)
q2_better %$% bootdif(drive_in_minutes, exposure, conf.level = 0.90)
```

```
Mean Difference            0.05            0.95
       19.07938        16.27898        21.73546
```

So the one-sided 95% confidence interval would have a lower bound of 16.28 minutes for the population mean driving time difference between rural and urban counties.

What if we used a t-based procedure instead?

I'll just emphasize that I don't think this is particularly justifiable, since the data in neither sample is well-approximated by a Normal model.

```
model2 <- lm(drive_in_minutes ~ exposure, data = q2_better)
```

```
tidy(model2, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low)
```

```
# A tibble: 2 x 3
  term          estimate conf.low
  <chr>            <dbl>    <dbl>
1 (Intercept)       26.7     24.7
2 exposurerural     19.1     16.4
```

The one-sided 95% confidence interval based on the t distribution uses the same lower bound as the two-sided 90% confidence interval, and that bound is 16.38 minutes.

To fit this directly, we could use the pooled t test:

```
q2_better %$% t.test(drive_in_minutes ~ exposure,
                     conf.level = 0.95,
                     var.equal = TRUE,
                     alt = "less")
```

```

    Two Sample t-test

data:  drive_in_minutes by exposure
t = -11.641, df = 485, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -16.37834
sample estimates:
mean in group urban mean in group rural
           26.74654            45.82593
```

Note that the sample sizes (217 vs. 270) and sample variances (19.14^2 = 366.34, and 16.99^2 = 288.66 are somewhat different from each other.

```
mosaic::favstats(drive_in_minutes ~ exposure, data = q2_better)
```

```
  exposure min Q1 median Q3 max     mean       sd   n
1    urban   1 15     24 32  86 26.74654 19.13755 217
2    rural   1 37     44 55 100 45.82593 16.98817 270
  missing
1       0
2       0
```

So we might instead prefer to use the Welch t method to obtain our confidence interval:

```
q2_better %$% t.test(drive_in_minutes ~ exposure,
                     conf.level = 0.95,
                     alt = "less")
```

```

    Welch Two Sample t-test

data:  drive_in_minutes by exposure
t = -11.491, df = 435.85, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -16.34259
sample estimates:
```

```
mean in group urban mean in group rural
           26.74654            45.82593
```

Question 3

This is a paired samples analysis, comparing the proportions with PHQ-9 values of 10 or higher before and after the administration of the medication regimen.

```
q1_wider %>% count(baseline >= 10, follow >= 10)
```

```
# A tibble: 4 x 3
  `baseline >= 10` `follow >= 10`     n
  <lgl>            <lgl>          <int>
1 FALSE            FALSE             77
2 FALSE            TRUE               6
3 TRUE             FALSE             70
4 TRUE             TRUE             107
```

Let's create two new variables to give us the 2x2 table of paired comparisons that we need.

```
q1_wider <- q1_wider %>%
  mutate(early = ifelse(baseline >= 10, "10+", "Below 10"),
         late = ifelse(follow >= 10, "10+", "Below 10"))

q1_wider %>%
  tabyl(early, late) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_title()
```

```
          late
    early  10+ Below 10 Total
      10+  107       70   177
 Below 10    6       77    83
    Total  113      147   260
```

So we have 76 discordant pairs here, of which 70 show improvements (patients who went from 10+ to < 10).

Summarizing with the McNemar Odds Ratio

Suppose we want to use the McNemar odds ratio to summarize this relationship. We'll build the desired 99% confidence interval using the exact McNemar test:

```
q1_wider %$%
  exact2x2(early, late, paired = TRUE,
           conf.int = TRUE, conf.level = 0.99)
```

```
    Exact McNemar test (with central confidence
    intervals)

data:  early and late
b = 70, c = 6, p-value = 6.312e-15
alternative hypothesis: true odds ratio is not equal to 1
99 percent confidence interval:
  4.17436 47.30917
```

```
sample estimates:
odds ratio
  11.66667
```

The odds of having a PHQ-9 score of 10 or higher are estimated to be 11.67 times higher (99% confidence interval: 4.17, 47.31) for subjects at the start than after the medication regimen.

Summarizing with the difference in proportions

Alternatively, we could have provided a confidence interval for the difference in proportions rather than the odds ratio.

```
diffpropci.Wald.mp(b = 70, c = 6, n = 260, conf.level = 0.99)
```

```
data:

99 percent confidence interval:
 -0.3230507 -0.1692570
sample estimates:
[1] -0.2461538
```

So the proportion with PHQ-9 of 10 or higher is estimated to be 0.246 (or 24.6 percentage points) lower after the medication regimen (99% CI: 0.169, 0.323) than it is before the medication regimen.

We could also have used the Agresti-Min procedure instead of the Wald CI...

```
diffpropci.mp(b = 70, c = 6, n = 260, conf.level = 0.99)
```

```
data:

99 percent confidence interval:
 -0.3212908 -0.1672588
sample estimates:
[1] -0.2442748
```

This approach estimates that the proportion with PHQ-9 of 10 or higher is 0.244 (or 24.4 percentage points) lower after the medication regimen (99% CI: 0.167, 0.321) than it is before the medication regimen.

Question 4.

```
q2_better %>% count(exposure, drive_in_minutes <= 20)
```

```
# A tibble: 4 x 3
  exposure `drive_in_minutes <= 20`     n
  <fct>    <lgl>                    <int>
1 urban    FALSE                      130
2 urban    TRUE                        87
3 rural    FALSE                      253
4 rural    TRUE                        17
```

Our 2x2 table (arranged in standard epidemiological format) is:

```
table_q4 <- q2_better %>%
  mutate(dr_length = ifelse(drive_in_minutes <= 20, "Short", "Too_Long")) %>%
  tabyl(exposure, dr_length)

table_q4
```

```
 exposure Short Too_Long
    urban    87      130
    rural    17      253
```

and now we can use our `twobytwo` approach from `Love-boost.R` to obtain a set of 99% confidence intervals.

```
twobytwo(87, 130, 17, 253,
         "Urban", "Rural", "Short", "Too_Long",
         conf.level = 0.99)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Short
Comparing : Urban vs. Rural

      Short Too_Long    P(Short) 99% conf. interval
Urban    87      130      0.4009     0.319   0.4888
Rural    17      253      0.6630     0.034   0.1136

                                  99% conf. interval
            Relative Risk: 6.3676    3.3529  12.0930
         Sample Odds Ratio: 9.9597    4.7641  20.8215
Conditional MLE Odds Ratio: 9.9101    4.7507  22.6905
    Probability difference: 0.3380    0.2426   0.4299

             Exact P-value: 0.0000
        Asymptotic P-value: 0.0000
------------------------------------------------------
```

In our data, 40.1% of urban and 6.3% of rural counties were centered within a 20 minute drive of an OHP facility. The estimated difference in proportions is 0.338 (33.8 percentage points) and the 99% confidence interval around that difference is (0.243, 0.430).

We could instead interpret this in terms of the relative risk, or the sample (cross-product) odds ratio and that would be equally appropriate. In any case, the difference seems vast.

Question 5

As always, we don't write sketches for essay questions. We will share some of the stronger ones.

On Grading Homework G

Grades on Homework G are on a 0-100 scale.

General/Administrative

- Subtract 20 points if they fail to turn in both Markdown and HTML on time (on time = within 1 hour of the deadline) but still get it in by 6 PM.
- Award zero points on the entire assignment to anyone whose first submission of the assignment is after 6 PM, unless excused from the assignment by Professor Love.

Question 1. (30 points)

Parts A, B, D, E and F are worth 2 points each for a correct response.

Parts C, G and H, taken together, are worth 10 points.

Part J is also worth 10 points. Their interpretation needs to be accurate, and their calculation needs to match the decisions they made in the previous parts of the question.

If a student incorrectly identifies these are independent samples, they should not score more than 20 points on the Question, but if they do everything correctly given their bad decision on independent vs. paired, they should only lose the 10 points associated with parts C, G and H.

If they failed to rearrange the data in such a way as to let them do a paired samples analysis, they should score no more than 20/30 on the Question. If they failed to read the section on "complete case only" and thus thought of the data as independent samples, then they also lose the 10 points for parts C/G/H.

If they provide an answer to Part I, even though they thought the data were paired, I would ignore that. If they thought the data were independent, they should have answered I rather than G and H, and they'll lose the 10 points anyway.

Question 2. (30 points)

Parts A, B, D, E and F are worth 2 points each for a correct response.

Parts C and I taken together, are worth 10 points.

Part J is also worth 10 points. Their interpretation needs to be accurate, and their calculation needs to match the decisions they made in the previous parts of the question.

If a student incorrectly identifies these are paired samples, they should not score more than 20 points on the Question, but if they do everything correctly given their bad decision on independent vs. paired, they should only lose the 10 points associated with parts C and I.

If they provide an answer to Parts G and H, even though they thought the data were independent, that should cost them 5 points.

Question 3. (10 points)

I think of this question as falling into two parts:

- Up to 5 points for how well they were able to correctly isolate the necessary paired comparisons table.
- Up to 5 points for identifying an appropriate confidence interval given whatever they did create and providing an appropriate interpretation of their result.

Question 4. (10 points)

Again, I think of this question as falling into two parts:

- Up to 5 points for how well they were able to correctly isolate the necessary independent samples 2x2 table.
- Up to 5 points for identifying an appropriate confidence interval given whatever they did create and providing an appropriate interpretation of their result.

Question 5 (20 points)

You need to identify (as a group) the 6-8 best essays (of the complete set of 60) that were read by the TAs (so that's choosing from the best two that each of you read, probably). In the Comments to Professor Love, please briefly identify the top 6-8 and specify the topic of these 6-8 best essays so I can read through them before returning them to the students, and select 2-4 to share.

No more than 14 out of 20 points should be given unless:

- the essay is clear,
- it answers the questions posed,
- it meets the word limit, and
- it has generally good grammar and spelling.

Students should receive **18-20 points** if they meet all of the standards above, and are one of the 6-8 best essays in the group.

Students should receive **15-17 points** if they meet all of the standards above, but were not in that top group.

Students should receive **12-14 points** if they meet three of the standards above but miss on one of them.

Students should receive **fewer than 12 points** if they fail to meet at least two of the standards above.

All students should receive some feedback (at least a "Nice job! I found this interesting and well-written") from the TA who did the initial grading of the work. I expect about 50 of our 60 students will receive grades on the essay between 12 and 17.