Answer Sketch for Homework 3

431 Staff and Professor Love

'Due 2018-09-21, version 2018-09-15

Contents

0.1 R Setup

Here's the complete R setup we used.

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

library(magrittr); library(tidyverse)
```

Then we read in the data set, which we'd stored in the project directory.

```
LBWunicef <- read.csv("LBWunicef.csv") %>% tbl_df
```

We could use `glimpse` to take a look at the data...

```
glimpse(LBWunicef)
```

```
Observations: 180
Variables: 3
$ nation    <fct> Albania, Algeria, Angola, Antigua and...
$ lbw.pct   <int> 4, 6, 12, 5, 7, 8, 7, 7, 10, 11, 8, 2...
$ least.dev <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
```

Or we could just list the tibble, as a check on what we've done...

```
LBWunicef
```

```
# A tibble: 180 x 3
   nation              lbw.pct least.dev
   <fct>                 <int>     <int>
 1 Albania                   4         0
 2 Algeria                   6         0
 3 Angola                   12         1
 4 Antigua and Barbuda       5         0
 5 Argentina                 7         0
 6 Armenia                   8         0
 7 Australia                 7         0
 8 Austria                   7         0
 9 Azerbaijan               10         0
10 Bahamas                  11         0
# ... with 170 more rows
```

The data includes 180 nations, and information on `lbw.pct` and `least.dev` status.

1 Question 1

How many nations have non-missing low birth weight percentage estimates?

There are 180 nations with non-missing low birth weight percentage estimates.

1.1 Using `dim` or `nrow` and `summary`

We can use the `dim` function, or the `nrow` function to determine the number of rows in the `LBWunicef` data, and we can use the `summary` function to see if there are any missing values in the `LBWunicef` data:

```
dim(LBWunicef)
```

```
[1] 180   3
```

```
nrow(LBWunicef)
```

```
[1] 180
```

```
summary(LBWunicef)
```

```
                 nation       lbw.pct        least.dev
 Albania             :  1   Min.   : 0.00   Min.   :0.00
 Algeria             :  1   1st Qu.: 6.00   1st Qu.:0.00
 Angola              :  1   Median : 9.00   Median :0.00
 Antigua and Barbuda:  1   Mean   :10.08   Mean   :0.25
 Argentina           :  1   3rd Qu.:12.00   3rd Qu.:0.25
 Armenia             :  1   Max.   :35.00   Max.   :1.00
 (Other)             :174
```

If there were any NA values in `lbw.pct`, the summary would indicate that. Since it doesn't, we must have 180 nations with a value of `lbw.pct`.

1.2 Using `favstats`

We need to figure out the total sample size, and the number of missing `lbw.pct` values. Perhaps we could use the `favstats` function from the `mosaic` package.

```
mosaic::favstats(~ lbw.pct, data = LBWunicef)
```

```
 min Q1 median Q3 max     mean      sd   n missing
   0  6      9 12  35 10.07778 5.46646 180       0
```

And in fact, there are no missing values in the low birth weight percentage data.

1.3 Using `skim` from `skimr`

```
skimr::skim_with(integer = list(hist = NULL))
## did that just to leave out the sparkline histograms

skimr::skim(LBWunicef)
```

```
Skim summary statistics
 n obs: 180
 n variables: 3
```

```
-- Variable type:factor ------------------------------------------------------------------------
 variable missing complete   n n_unique
   nation       0      180 180      180
                    top_counts ordered
 Alb: 1, Alg: 1, Ang: 1, Ant: 1   FALSE

-- Variable type:integer -----------------------------------------------------------------------
  variable missing complete   n  mean   sd p0 p25 p50   p75
   lbw.pct       0      180 180 10.08 5.47  0   6   9 12
 least.dev       0      180 180  0.25 0.43  0   0   0  0.25
 p100
   35
    1
```

1.4 Using anyNA and length

Alternatively, using the %$% version of the pipe available in the magrittr package, we could use:

```
LBWunicef %$% anyNA(lbw.pct)
```

```
[1] FALSE
```

```
## returns TRUE if there are any missing (NA) values
## in the lbw.pct variable

## to get the number of values in lbw.pct, could use:
LBWunicef %$% length(lbw.pct)
```

```
[1] 180
```

For more on piping like this, visit the Pipes section in *R for Data Science* at http://r4ds.had.co.nz/pipes.html. The %$% function is described there as "exploding" out the variables in a data frame so that you can refer to them explicitly.

2 Question 2

> Which nations have the three largest low birth weight percentages? Are each of these considered by the UN to be "least developed" nations or not?

The three largest low birth weight percentages in the data are Mauritania (35%), Pakistan (32%), and India (28%). Of these three nations, only the troubled Northern African nation of Mauritania falls in the "least developed nations" category.

2.1 Using dplyr and the tidyverse

We can use dplyr, specifically the arrange function, to show a tibble that has been sorted in descending order of lbw.pct. R Studio's cheat sheet for Data Transformation at https://www.rstudio.com/resources/cheatsheets/ is very helpful here.

```
LBWunicef %>% arrange(desc(lbw.pct))
```

```
# A tibble: 180 x 3
   nation      lbw.pct least.dev
   <fct>         <int>     <int>
```

```
 1 Mauritania       35         1
 2 Pakistan         32         0
 3 India            28         0
 4 Nauru            27         0
 5 Niger            27         1
 6 Comoros          25         1
 7 Haiti            23         1
 8 Bangladesh       22         1
 9 Philippines      21         0
10 Chad             20         1
# ... with 170 more rows
```

And, if we wanted to view just the first three rows, we could arrange and then slice...

```
LBWunicef %>%
  arrange(desc(lbw.pct)) %>%
  slice(1:3)
```

```
# A tibble: 3 x 3
  nation     lbw.pct least.dev
  <fct>        <int>     <int>
1 Mauritania      35         1
2 Pakistan        32         0
3 India           28         0
```

2.2 A fast, one-line alternative with `rank`

```
## The fastest one-line alternative I know
LBWunicef[which(rank(LBWunicef$lbw.pct) > length(LBWunicef$lbw.pct) - 3),]
```

```
# A tibble: 3 x 3
  nation     lbw.pct least.dev
  <fct>        <int>     <int>
1 India           28         0
2 Mauritania      35         1
3 Pakistan        32         0
```

2.3 `sort`, `which` and brute force

Clearly, we could solve this problem through simple brute force, inspecting the data until we find the largest values, and then associating them with Nations. The `sort` and `which` commands can help us here.

```
LBWunicef %$% sort(lbw.pct)
```

```
  [1]  0  3  3  3  3  4  4  4  4  4  4  4  4  4  4  4  5  5
 [19]  5  5  5  5  5  5  5  5  5  5  5  6  6  6  6  6  6  6
 [37]  6  6  6  6  6  6  6  6  6  6  6  6  6  6  7  7  7  7
 [55]  7  7  7  7  7  7  7  7  7  8  8  8  8  8  8  8  8  8
 [73]  8  8  8  8  8  8  8  8  8  8  8  9  9  9  9  9  9  9
 [91]  9  9  9  9 10 10 10 10 10 10 10 10 10 10 10 10 10 10
[109] 10 10 10 10 11 11 11 11 11 11 11 11 11 11 11 11 11 11
[127] 11 11 11 11 11 12 12 12 12 12 12 12 13 13 13 13 13 13
[145] 13 13 14 14 14 14 14 14 14 14 14 15 15 15 15 16 16 17
[163] 17 17 18 18 18 18 19 20 20 21 22 23 25 27 27 28 32 35
```

OK. So the three largest values have `lbw.pct` greater than 27. How do we identify which nations those are?

```
LBWunicef %$% which(lbw.pct > 27)
```

```
[1]  73 103 122
```

And now that we know which row numbers are the top 3, we can show all of the available data related to those three row numbers (including their names) using `slice` to identify specific rows in the data.

```
LBWunicef %>% slice(c(73, 103, 122))
```

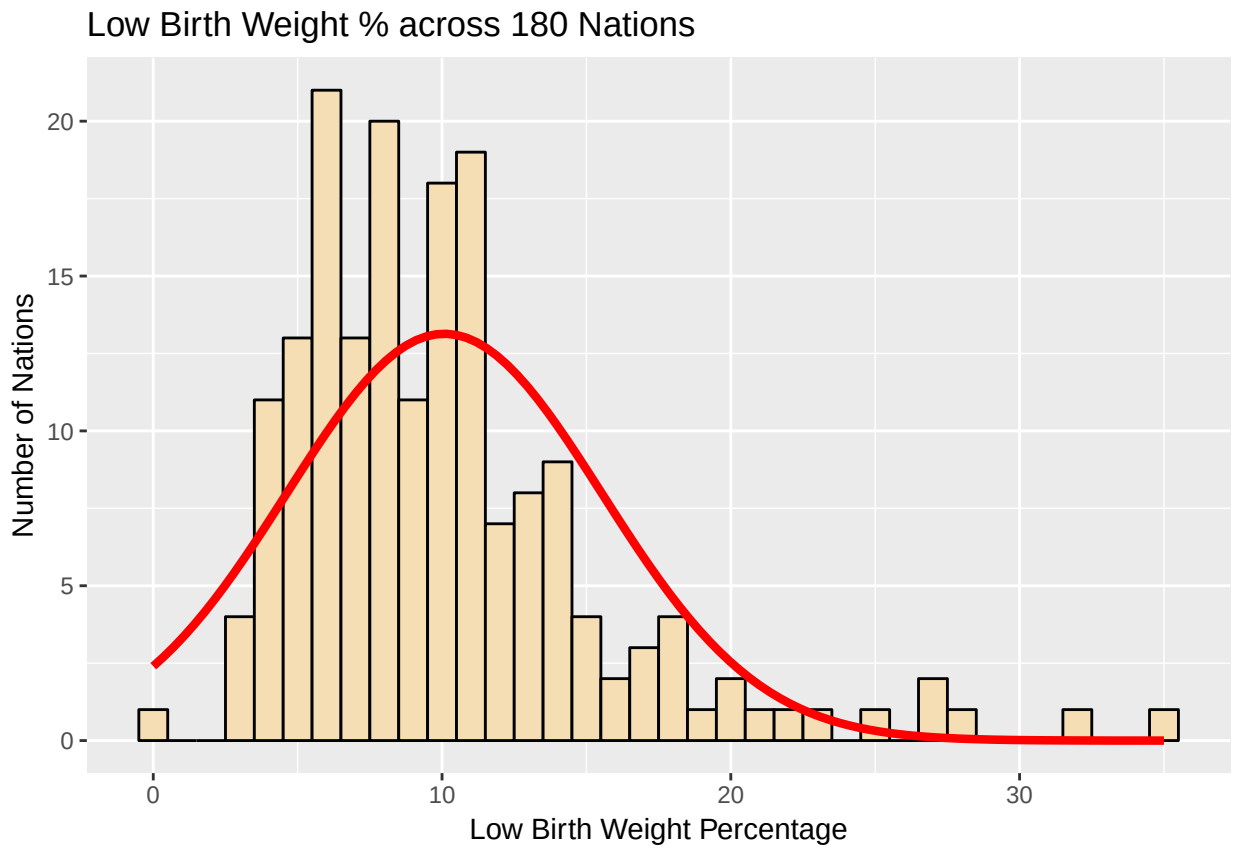
```
# A tibble: 3 x 3
  nation     lbw.pct least.dev
  <fct>        <int>     <int>
1 India           28         0
2 Mauritania      35         1
3 Pakistan        32         0
```

3 Question 3

Create a histogram of the low birth weight percentages, then superimpose a normal density function with the same mean and standard deviation in red. Based on your plot, is the standard deviation or the inter-quartile range a more appropriate measure of variation in the low birth weight rates? Why?

Here's one approach.

```
ggplot(LBWunicef, aes(x = lbw.pct)) +
  geom_histogram(fill = "wheat", col = "black",
                 binwidth = 1) +
  stat_function(fun = function(x, mean, sd, n)
    n * dnorm(x = x, mean = mean, sd = sd),
    args = with(LBWunicef,
                c(mean = mean(lbw.pct),
                  sd = sd(lbw.pct),
                  n = length(lbw.pct))),
    col = "red", lwd = 1.5) +
  labs(title = "Low Birth Weight % across 180 Nations",
       x = "Low Birth Weight Percentage",
       y = "Number of Nations")
```



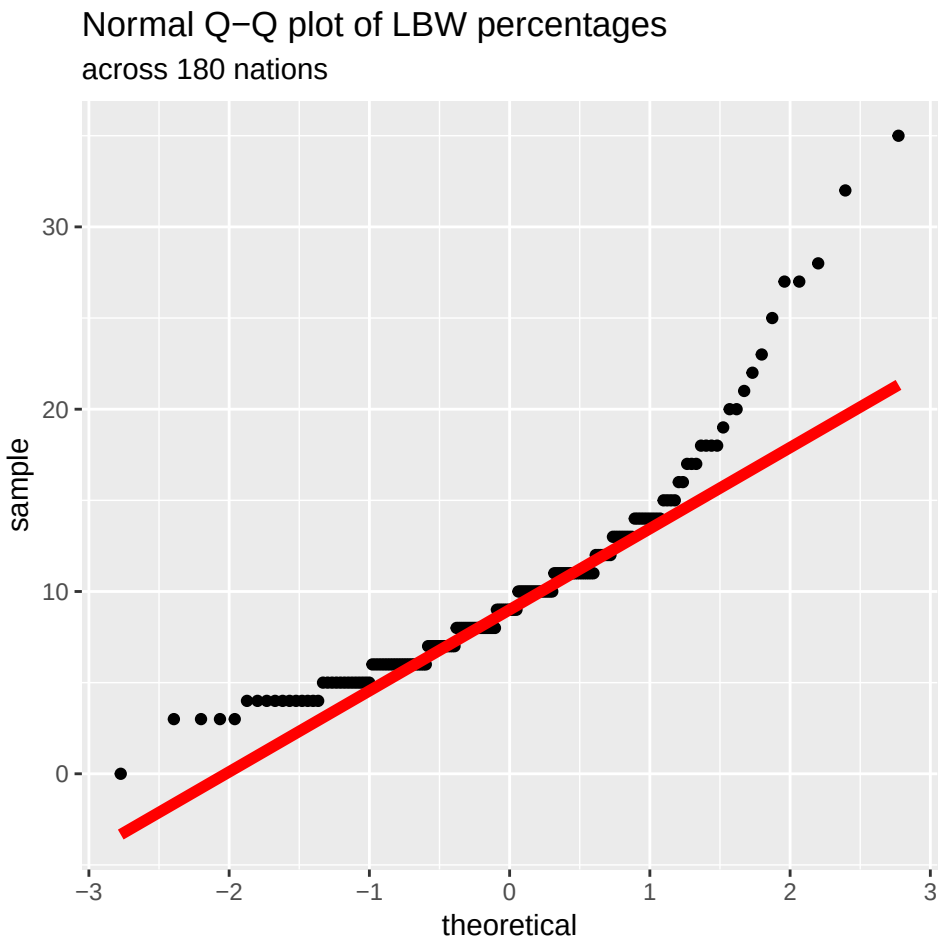
Clearly, the plot shows substantial right skew, so assuming a Normal model is not well justified. Thus, the standard deviation is less appropriate as a measure of spread than the interquartile range.

4 Question 4

Create a normal Q-Q plot for the low birth weight percentage estimates. Would you say that the data are approximately Normally distributed, or not approximately Normally distributed? Justify your answer by interpreting what you see in your plot, and whatever summary statistics you deem to be useful in making your decision.

Again, the data are clearly right skewed, as indicated by the curve in the normal Q-Q plot.

```
ggplot(LBWunicef, aes(sample = lbw.pct)) +
  geom_qq() + geom_qq_line(col = "red", lwd = 2) +
  labs(title = "Normal Q-Q plot of LBW percentages",
       subtitle = "across 180 nations")
```



4.1 Using Numerical Summaries to Assess Normality

As usual, we should focus first on the plots to assess Normality, which might realistically have included a boxplot or violin plot along with the histogram and Normal Q-Q plot we've seen. Summary statistics should play a supporting role.

4.1.1 Thinking about A Skewness Measure

```
mosaic::favstats(~ lbw.pct, data = LBWunicef)
```

```
 min Q1 median Q3 max     mean      sd   n missing
   0  6      9 12  35 10.07778 5.46646 180       0
```

As for summary statistics, the mean (10.08) is well to the right of the median (9), and, since the standard deviation is 5.47. So the skew_1 value is also indicative of right skew, with $\text{skew}_1 = 0.197$, which is essentially the value we usually use as a minimum indicator of substantial right skew.

```
LBWunicef %>%
  summarize(mean(lbw.pct), median(lbw.pct), sd(lbw.pct),
    skew1 = ( mean(lbw.pct) - median(lbw.pct) ) /
                                          sd(lbw.pct) )
```

```
# A tibble: 1 x 4
  `mean(lbw.pct)` `median(lbw.pct)` `sd(lbw.pct)` skew1
            <dbl>             <dbl>         <dbl> <dbl>
1            10.1                 9          5.47 0.197
```

4.1.2 Thinking about the Empirical Rule

We've already decided now that the data aren't symmetric enough for a Normal model to be a particularly good choice. If we wanted, we could also determine whether the Empirical Rule holds well for these data, and use that to help guide our understanding of whether the Normal model would work well (although at this point, that seems pretty settled.)

For instance, if a Normal model held, then about 68% of the nations would fall within two standard deviations of the mean. Is that true here?

```
LBWunicef %>%
  count(mean_pm_1sd = lbw.pct > mean(lbw.pct) - sd(lbw.pct) &
        lbw.pct < mean(lbw.pct) + sd(lbw.pct) )
```

```
# A tibble: 2 x 2
  mean_pm_1sd     n
  <lgl>       <int>
1 FALSE          37
2 TRUE          143
```

In fact, 143/180 is 79.4% of the nations that fall within 1 SD of the mean. That's higher than we would expect in data that followed a Normal distribution, so this pushes us slightly further in the direction we were already going when we just had the pictures - of concluding that the Normal model isn't a good choice for these data.

If a Normal model held, for instance, then about 95% of the data would fall within two standard deviations of the mean. Is that true here?

```
LBWunicef %>%
  count(mean_pm_2sd = lbw.pct > mean(lbw.pct) - 2*sd(lbw.pct) &
        lbw.pct < mean(lbw.pct) + 2*sd(lbw.pct) )
```

```
# A tibble: 2 x 2
  mean_pm_2sd     n
  <lgl>       <int>
1 FALSE           8
```

```
2 TRUE          172
```

And 172/180 is 95.6% of the nations that fall within 2 SD of the mean value of `lbw.pct`. That's pretty close to expectations, but, again, the 1 SD empirical rule doesn't hold so well.

4.1.3 Thinking about Hypothesis Testing (Shapiro-Wilk Test)

A really, really bad idea is to use a hypothesis test to assess Normality. Such a test is essentially valueless without first looking at a plot of the data. But such tests are available. None are great, specifically because they only test for specific types of non-Normality, and most people can visualize several types of non-Normality simultaneously, making that (visualization) a much more powerful tool (even if it seems less "objective").

One of the simplest of such tests to run is the Shapiro-Wilk test of Normality. That test estimates a p value, something that's very easy to misinterpret. In the case of a Shapiro-Wilk test, if you see a p value that is less than a given value (the most common choice is 0.05), then that suggests that there is some evidence of non-Normality in the way the Shapiro-Wilk test tries to find it, or at least there's more evidence than if the p value were larger. The p value is a conditional probability, so it will always fall between 0 and 1.

```
LBWunicef %$% shapiro.test(lbw.pct)
```

```
    Shapiro-Wilk normality test

data:  lbw.pct
W = 0.87017, p-value = 2.462e-11
```

Here, the p value is very small, which pushes us slightly further in the direction of concluding that the Normal model isn't a good choice for these data.

Other hypothesis tests are available for assessing non-Normality. Again, none are great. In fact, I can't remember the last time I reported a Shapiro-Wilk test (or any other hypothesis test for non-Normality) in my practical work.

5 Question 5

> Display an effective graph comparing the two development groups (least developed nations vs. all other nations) in terms of their percentages of low birth weight births. What conclusions can you draw about the distribution of low birth weight rates across the two development groups? Be sure to label your graph so it stands alone, and also supplement your graph with separate text discussing your conclusions.

Generally, the low birth weight percentages are higher in the nations which are least developed, but there is considerable overlap.

5.1 Preliminaries: Creating a Factor

Before I build my plot, I'll create a new factor variable in the `LBWunicef` data, which I'll call `least_developed` and which will contain the levels No and Yes, for the original numeric 0 and 1.

```
LBWunicef <- LBWunicef %>%
  mutate(least_developed = fct_recode(factor(least.dev), "Yes" = "1", "No" = "0"))
```

Just as a sanity check, I'll be sure I've recoded appropriately with a frequency table:

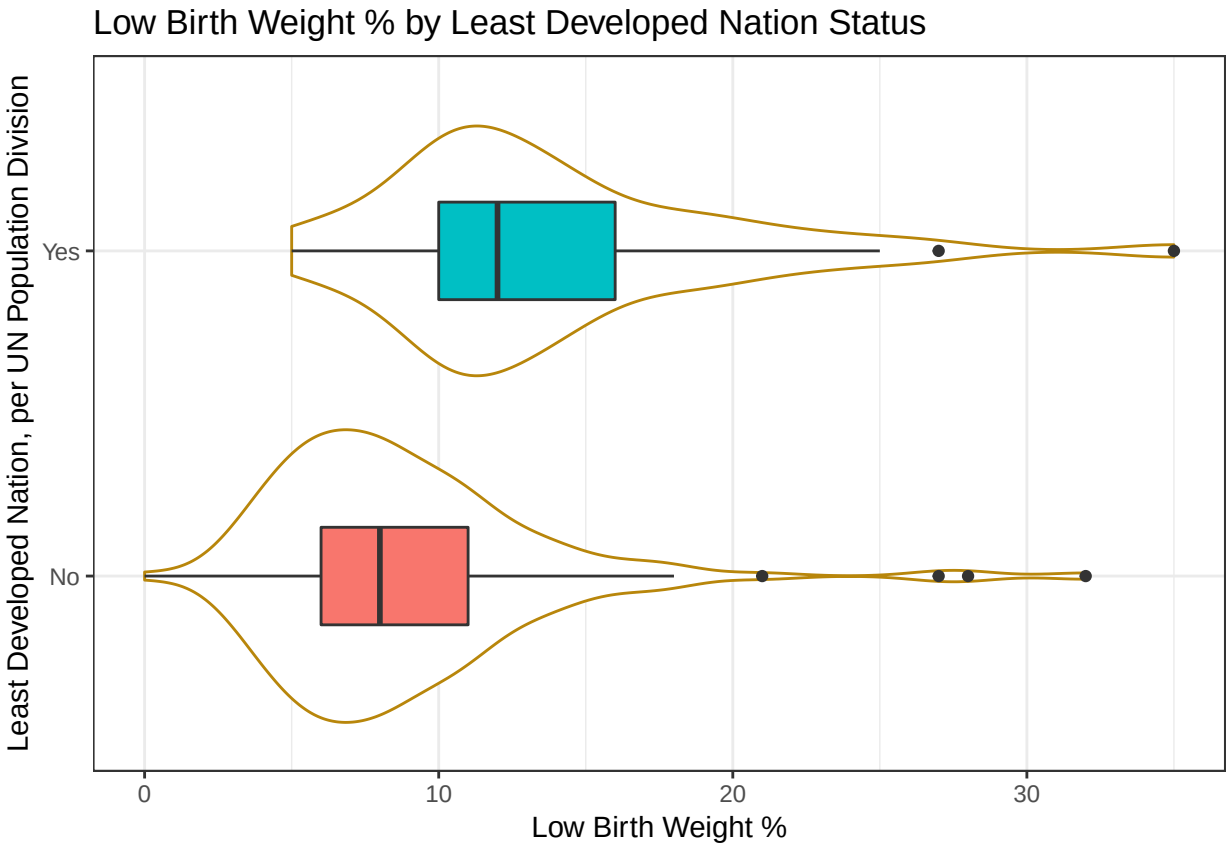
```
LBWunicef %>% count(least_developed, least.dev)
```

```
# A tibble: 2 x 3
  least_developed least.dev     n
  <fct>               <int> <int>
1 No                      0   135
2 Yes                     1    45
```

5.2 A Comparison Boxplot (and Violin Plot)

Now, I'll build a comparison boxplot. I'll get a little fancy and create violin plots while I am at it.

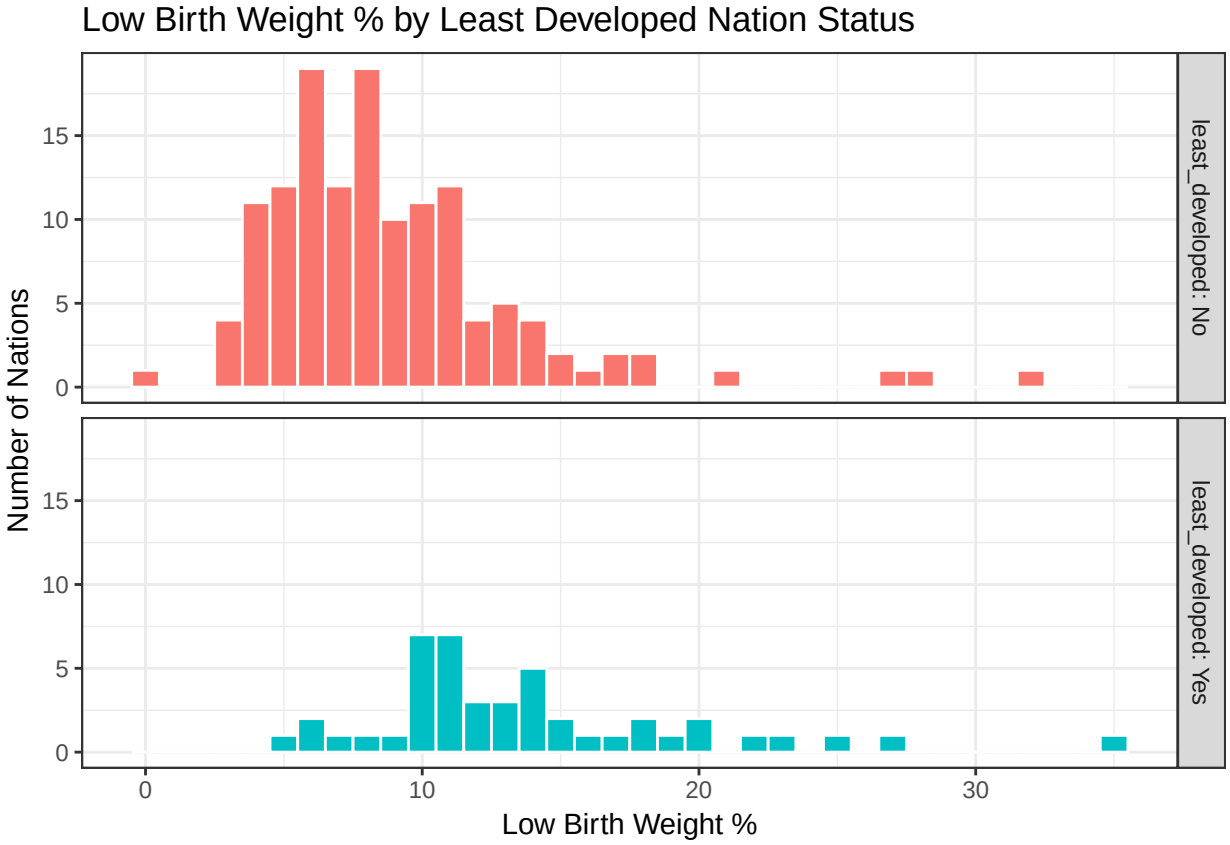
```
ggplot(LBWunicef, aes(x = least_developed, y = lbw.pct)) +
  geom_violin(col = "darkgoldenrod") +
  geom_boxplot(aes(fill = least_developed), width = 0.3) +
  guides(fill = FALSE) +
  coord_flip() +
  labs(title = "Low Birth Weight % by Least Developed Nation Status",
       y = "Low Birth Weight %",
       x = "Least Developed Nation, per UN Population Division") +
  theme_bw()
```



5.3 Another Reasonable Choice: Faceted Histograms

You could certainly have built a set of faceted histograms instead, but ideally, you'd have them arranged so that the distributions were easy to compare (the two histograms on top of each other, as these boxplots are, rather than just plotted next to each other.) That's part of the reason I flipped those boxplots. Here's our attempt.

```
ggplot(LBWunicef, aes(x = lbw.pct, fill = least_developed)) +
  geom_histogram(binwidth = 1, col = "white" ) +
  facet_grid(least_developed ~ ., labeller = "label_both") +
  guides(fill = FALSE) +
  labs(title = "Low Birth Weight % by Least Developed Nation Status",
       y = "Number of Nations",
       x = "Low Birth Weight %") +
  theme_bw()
```



This does convey a bit more effectively that the "least developed" nations comprise one-quarter (45/180) of the total set of nations, but I think on the whole I prefer the boxplot here.

6 Question 6

> Read the Introduction and Chapter 1 of Nate Silver's *The Signal and the Noise*. One possible takeaway, particularly from the Introduction, suggested, for example in a review by Jonah Sinick, might be that increased access to information can do more harm than good.
>
> Tell us about an example in your own field/work/experience where a "surplus" of information made (or makes) it easier for people dealing with a complex system to cherry-pick information that supports their prior positions. What were the implications of your example in terms of lessons that can be learned? If you can connect your example to some of the lessons described in the Chapter 1 discussion of the failure to predict the 2008 catastrophe on the US economy, that would be welcome.
>
> Please feel free to supply as many supporting details as are useful to you in relating the story. An appropriate response to Question 6 will use complete English sentences with proper grammar and syntax, will cite a link or two to a Web URL or other published work, and be between 200 and 400 words long.

We don't write answer sketches for essay questions. We'll gather a few of the more interesting and enlightening responses, and share de-identified excerpts with the group after grading.

7 Question 7

> Generate a "random" sample of 75 observations from a Normal distribution with mean 100 and standard deviation 10 using R. The `rnorm` function is likely to be helpful. Now, display a normal Q-Q plot of these data, using the `ggplot2` package from the `tidyverse`. How well does the Q-Q plot approximate a straight line?
>
> Repeat this task for a second sample of 150 Normally distributed observations, again with a mean of 100 and a standard deviation of 10. Then repeat it again for samples of 25 and 225 Normally distributed observations with a different mean and variance. Which of the four Q-Q plots you have developed better approximates a straight line and what should we expect the relationship of sample size with this phenomenon to be?

From a coding perspective, I'm just looking for you to properly draw a random sample from a Normal distribution and then produce the necessary plots.

Note that you either want to use four different random seeds here, and build each sample separately, or build one long set of 475 samples (`75 + 150 + 25 + 225 = 475`) to cover all four needs, and then split the group of 475 values accordingly.

Building four separate samples might look like this:

```
set.seed(43101); x <- rnorm(n = 75, mean = 100, sd = 10)
  samp1 <- data_frame(value = x, grp = rep("S-75", 75))
set.seed(43102); x <- rnorm(n = 150, mean = 100, sd = 10)
  samp2 <- data_frame(value = x, grp = rep("S-150", 150))
set.seed(43103); x <- rnorm(n = 25, mean = 100, sd = 10)
  samp3 <- data_frame(value = x, grp = rep("S-25", 25))
set.seed(43104); x <- rnorm(n = 225, mean = 100, sd = 10)
  samp4 <- data_frame(value = x, grp = rep("S-225", 225))

q7_first_try <- bind_rows(samp1, samp2, samp3, samp4)
rm(samp1, samp2, samp3, samp4, x) # drop these vectors
```

But what I actually did was build a single set of 475 values, and then split them, using this code:

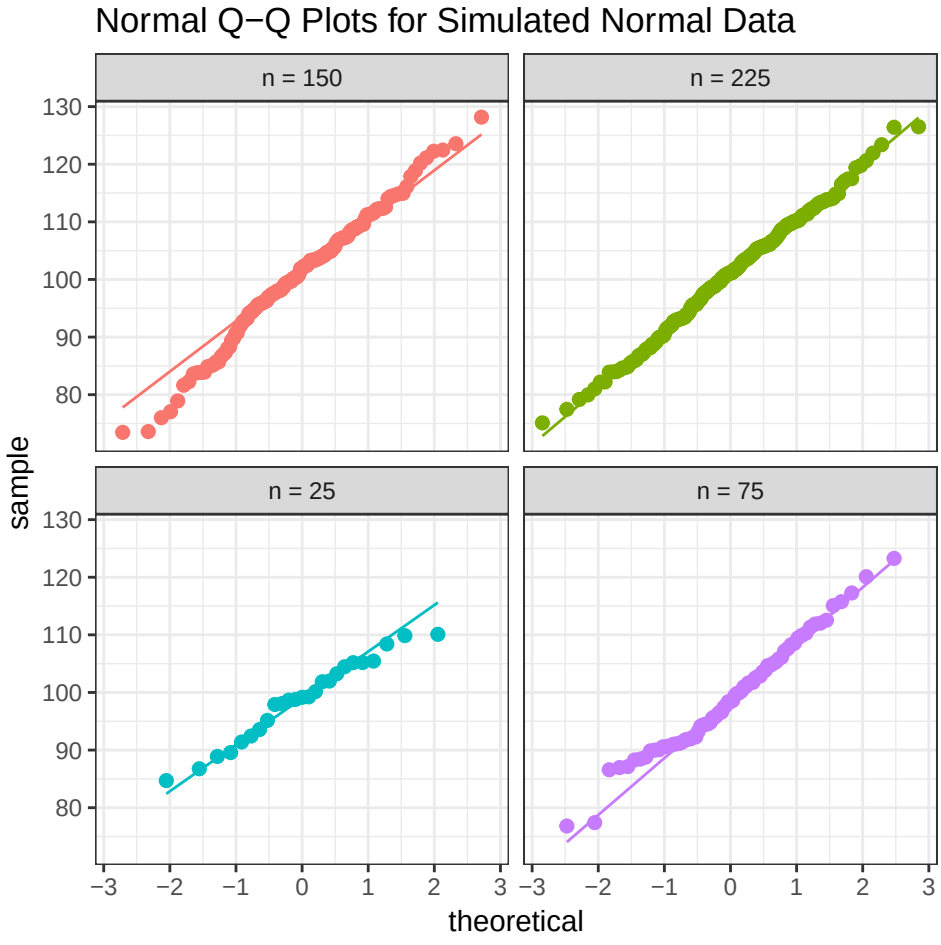
```
set.seed(20180921); big.sample <- rnorm(n = 475, mean = 100, sd = 10)
big.grp <- c(rep("n = 75", 75), rep("n = 150", 150),
             rep("n = 25", 25), rep("n = 225", 225))

q7_data <- data_frame(value = big.sample, grp = big.grp)
rm(big.sample, big.grp) # we won't need those vectors again
```

So, now we are ready to build the four Normal Q-Q plots.

All four of these plots show fairly modest deviations from what we would expect a Normal distribution to look like, usually in terms of showing a few outlying values.

```
ggplot(q7_data, aes(sample = value, col = grp)) +
  geom_qq(size = 2) + geom_qq_line() +
  guides(color = FALSE) +
  facet_wrap(~ grp) +
  labs(title = "Normal Q-Q Plots for Simulated Normal Data") +
  theme_bw()
```



With larger sample sizes, there's **no real reason** to assume that the plots will improve substantially in terms of eliminating outliers, in fact. Once we have at least 25 points (as in all of these cases) it appears that the results are fairly reasonable (in terms of suggesting that a Normal approximation is generally valid) in all of these plots.