STATS 60 Summer 2020: HW3 Solution

Remarks

Some of the solutions use submissions from students. Thanks and acknowledgments to these students.

You will need to use the packages `ggplot2`, `dplyr` and `lubridate` for this problem set. Install the package `lubridate` with `install.packages("lubridate")` and load the package with `library(lubridate)`

```
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.2
```

Problem 1 Reading habits

Your friend wants find out reading habits of residents of Palo Alto. In particular, how many books people read per year and which types of books they read most.

(a) He has access to the checkout records at Palo Alto library, which contain the following variables
Year Which year is it?
Genre Book category, such as "Action", "Anthology", "History" etc.
Total Number of checkouts

Describe the data type and measurement scale of each of these variable. Is it qualitative or quantitative? Is it nominal, ordinal, interval or ratio?

Answer Year is quantitative and interval scale, because difference between years has a consistent meaning but we don't multiply numbers to years. Genre is qualitative and nominal scale. Total is quantitative and ratio scale, because it has absolute zero and can be multiplied.

(b) He divides the total checkouts by the total visitors to the library in each year and obtain the ratio. He also finds the most popular categories in terms of this ratio. Do you think his approach answers the questions? Explain your answers in a paragraph. If yes, explain why. If the answer is no, explain what more information you need to get a better answer.

Answer One issue is that borrowers at library may not be representative of all the residents at Palo Alto, for instance they might read more so the the average book they read may be higher than general public. Another issue is this data is not comprehensive, since people purchase books, or borrow from other libraries. He can complement this data with surveys about how many books people read, and what percentage of those books do they borrow.

Many of you noticed other issues with this approach. For example, a single number does not provide a comprehensive picture of the reading habbits. Also, the number of visitors may not be the good denominator to use because people may be at the library to read books but not checkout, they may also check in and out multiple times. Further, the number of books checkout from each category is limited by the stock of the library. Finally, in part 2, some of you suggested using the checkout proportion of each genre among total checkouts to measure popularity of book categories.

Problem 2 Pooled testing

Recently there's some discussions about group screening of COVID-19. By pooling samples from many people into groups, and evaluating pools rather than individuals, we can potentially reduce the number of tests and increase the number of people tested. This problem explores the cost-effectiveness of this idea.

Suppose a test has sensitivity 99.9% and specificity 99.8%. For simplicity, we assume that this is true for pooled tests as well, that is to say, we can treat the pooled sample as a single sample, which is infected if at least one person in the group is infected. In this problem, we consider groups of size 10. Finally assume the infection probability is 14 in 1000 people.

(a) What's the probability that everyone is healthy in a group?

Answer

$$\Pr(\text{all healthy}) = (1 - 0.014)^{10} = 0.868.$$

(b) What's the probability a group tests positive?

Answer Let H denote the event that everyone in the group are healthy and let P denote the event that the group tests positive.

$$\begin{aligned}\Pr(P) &= \Pr(P \cap H) + \Pr(P \cap \text{not}H)\\ &= \Pr(P \mid H)\Pr(H) + \Pr(P \mid \text{not}H)\Pr(\text{not}H)\\ &= 0.002 \times 0.868 + 0.999 \times (1 - 0.868)\\ &= 0.134.\end{aligned}$$

(c) Now suppose a group tests positive, what's the posterior odds that at least one person from the group is infected?

Answer

$$\begin{aligned}\Pr(\text{not}H \mid P) &= 1 - \Pr(H \mid P)\\ &= 1 - \frac{\Pr(H \cap P)}{\Pr(P)}\\ &= 1 - \frac{0.002 \times 0.868}{0.134}\\ &= 0.987.\end{aligned}$$

The posterior odds is

$$\frac{\Pr(\text{not}H \mid P)}{\Pr(H \mid P)} = \frac{0.987}{1 - 0.987} = 76.2.$$

(d) Suppose we test 5000 separate groups, which corresponds to testing 50,000 people. What's the probability that 650 groups test positive? What about 800?

Answer The number of groups tested positive is from a binomial distribution with size 5000 and success probability 0.134.

```r
dbinom(650, 5000, 0.134)
```

```
## [1] 0.01184836
```

```r
dbinom(800, 5000, 0.134)
```

```
## [1] 1.503753e-08
```

Extra How many groups do you expect to test positive?

Answer The expected number of groups tested positive is 5000 * 0.134 = 670.

Problem 3 Sampling distribution of the mean

The Current Population Survey at this link provides summary data of individual education, work and income. Specifically, we use the household income data from year 2018 at this link, which contains the number of households (in thousands) in each income bracket. For example, if you download the data table, you can see there are 4,283 (in thousands) households out of 128,581 whose yearly income is less than $5000.

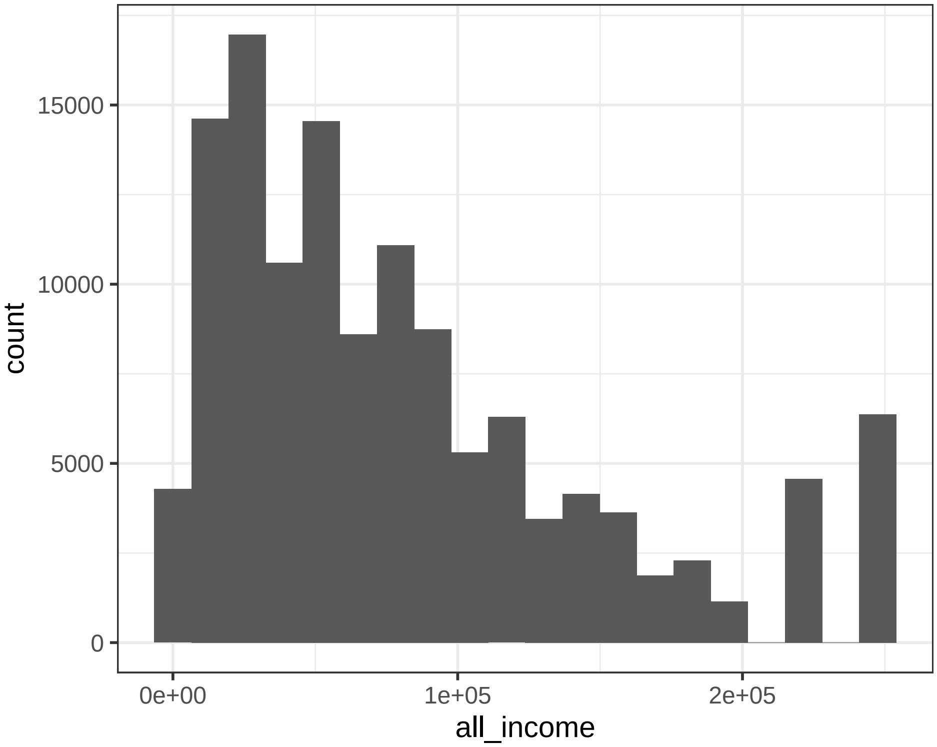
From this source, we generate a hypothetical population of size 128,581, i.e. we take one household in every thousand, assuming the income of all the households in one bracket is the midpoint of that bracket. Finally, for households in the last bracket "$250,000 and above", we treat their income as 250,000. The following R code creates this population and store it in the vector `all_income`, whose length is 128,581 and contains hypothetical incomes. We will work with this hypothetical income population `all_income`.

```
income <- c(seq(0, 200000, by = 5000), 250000) # income brackets
mid_income <- c((income[-42] + income[-1]) / 2, 250000) # mid-points of the bracket
N <- c(4283, 3337, 5510, 5772, 5672, 5469, 5822, 5404, 5195, 4839,
       5300, 4417, 4604, 3999, 3795, 3950, 3349, 3064, 3102, 2581,
       2866, 2449, 2318, 1971, 2004, 1780, 1678, 1426, 1414, 1316,
       1492, 978, 1161,  970 , 905 , 835 , 772 , 686  ,584 , 565 ,
       4572, 6375) # number of households in each bracket
all_income <- rep(mid_income, N)
```

(a) Make a histogram of this hypothetical income population. Why is there a uptick at 250,000? Is it from a normal distribution? Why?

Answer This distribution is skewed to the right and not normal, because the normal distribution is symmetric around the mean. There's an uptick at 250,000 because we truncated all income values above 250,000 to 250,000.

```
ggplot() +
  geom_histogram(aes(x = all_income), bins = 20) +
  theme_bw()
```



(b) What is the population mean, median and interquartile range?

Answer

```
mean(all_income)
```

```
## [1] 82198.34
```

```
median(all_income)
```

```
## [1] 62500
```

```
IQR(all_income)
```

```
## [1] 80000
```

(c) What is the z-score of a household income of $40,000?

Answer We subtract the mean and divide by the standard deviation. The SD is 6.6156843×10^4.

```
(40000 - mean(all_income))/sd(all_income)
```

```
## [1] -0.637853
```

(d) Draw a random sample of size 1000 of all the incomes, what is the mean of this random sample? In R, we can draw a random sample using the `sample` function, for example the following R code draws a random sample without replacement from the population and stores it to the vector `s`.

```
s <- sample(all_income, size = 1000)
```

Answer

```
mean(s)
```

```
## [1] 83727.5
```

(e) Draw 1000 random samples of size $n = 1000$, compute the mean of each sample. Draw a histogram of all the sample means. Is it approximately normal? What is the mean and standard deviation of these sample means?

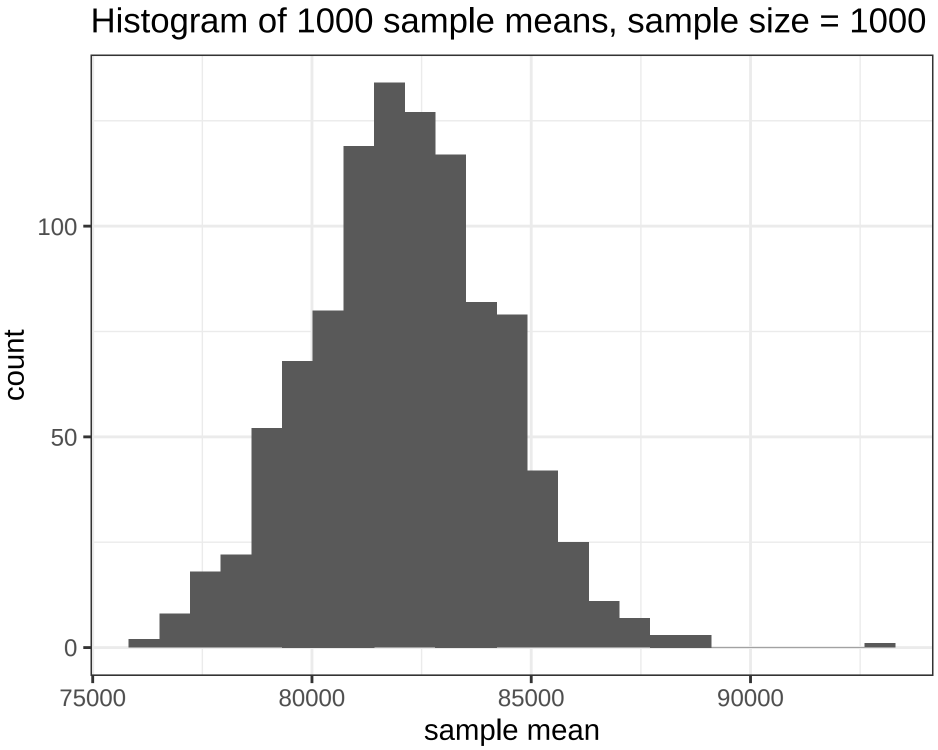
You can use a for loop iterate a process many times. Here's an example.

```
result <- numeric(length = 10) # a vector of 100 zero
for(i in 1:10){ # iterate i from 1 to 100
  # write the action to perform in one loop inside of bracket { }
  # here we assign the i-th element of result to i
  result[i] <- i
}
result
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

Answer

```
set.seed(1) # set seed to get the same anwer
B <- 1000
result <- numeric(length = B)
for(i in 1:B){
  result[i] <- mean(sample(all_income, 1000)) # plug in the sample mean
}
ggplot()+
  geom_histogram(aes(x = result), bins = 25) +
  xlab("sample mean") +
  ggtitle("Histogram of 1000 sample means, sample size = 1000")+
  theme_bw()
```



```
sd(result)
```

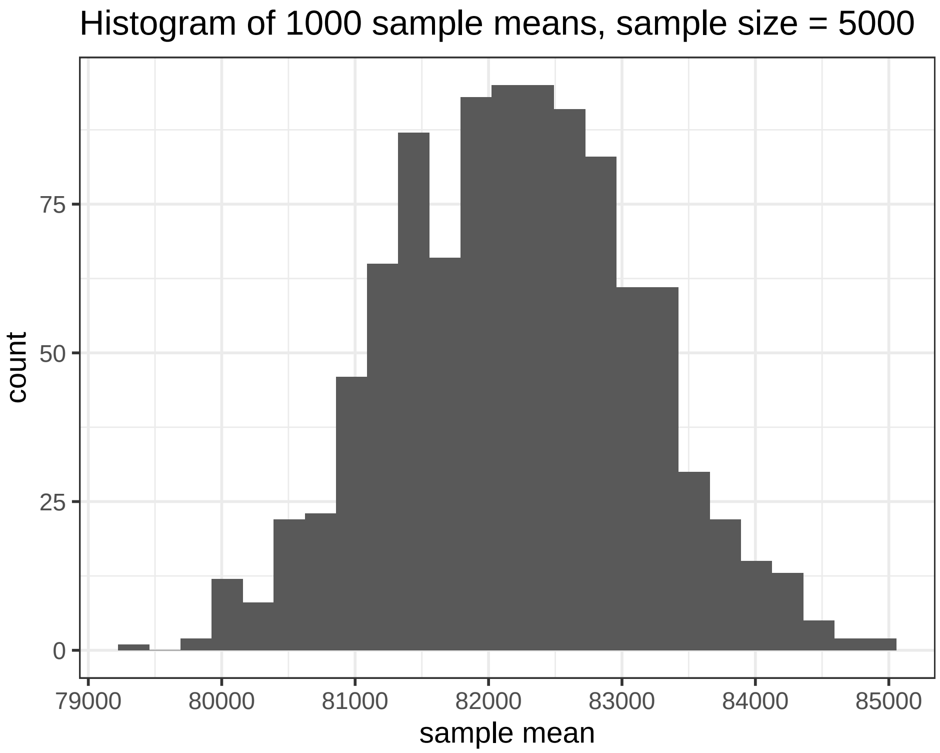
```
## [1] 2166.719
```

The histogram of the sample mean is approximately normal, its mean is 8.2124803×10^4 and its standard deviation is 2166.7190262

(f) Repeat part (d) and (e) with sample size $n = 5000$, what do you observe?

Answer We know that the standard deviation of the sample mean is $\sigma/\sqrt{n}$ where n is the sample size. The standard deviation of the sample means in part(f) are about 923, and it is 2141 from part(e). $2141/923 \approx \sqrt{5}$, as expected.

```
B <- 1000
result <- numeric(length = B)
for(i in 1:B){
  result[i] <- mean(sample(all_income, 5000))
}
ggplot()+
  geom_histogram(aes(x = result), bins = 25) +
  xlab("sample mean") +
  ggtitle("Histogram of 1000 sample means, sample size = 5000") +
  theme_bw()
```



```
mean(result)
```

```
## [1] 82202.13
```

```
sd(result)
```

```
## [1] 926.6946
```

Problem 4 Drought data

This exercise explores more the US drought data we looked at in the Rlab. The data we use was downloaded from US drought monitor website, the state drought level in terms of percent area from 2015-01-01 to 2019-12-31. You can download the dataset in the course website assignment webpage and read the data into a data frame `drought` using the following code. Make sure you change the path to where your data is located.

```
drought <- read.csv("drought.csv") # use this path if your data is at "Desktop" folder
```

Before we start, we modify column names and select columns of interest. For simplicity, we assume each row to correspond to 7 day drought level starting at `ValidStart`. We convert `ValidStart` to a `Date` object and store it at the column `date`.

```
drought <- drought %>%
  mutate(
    date = ymd(ValidStart), # ymd is a function from lubridate package
    state = as.character(StateAbbreviation)
    ) %>%
  select(date, state, None, D0:D4)
```

You can take a quick look at the data with `glimpse`.

```
glimpse(drought)
```

```
## Rows: 13,624
## Columns: 8
## $ date  <date> 2019-12-31, 2019-12-24, 2019-12-17, 2019-12-10, 2019-12...
## $ state <chr> "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "A...
## $ None  <dbl> 93.18, 93.18, 93.18, 92.48, 92.48, 91.17, 91.17, 90.29, ...
## $ D0    <dbl> 6.82, 6.82, 6.82, 7.52, 7.52, 8.83, 8.83, 9.71, 9.71, 9....
## $ D1    <dbl> 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 3.96, 3.96, 3....
## $ D2    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 2....
## $ D3    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0....
## $ D4    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

(a) Drought severity in each state

In this exercise we will compute the average drought severity in each state in the year 2019. The drought severity index at this link is given by

$$1(D_0) + 2(D_1) + 3(D_2) + 4(D_3) + 5(D_4) = DSCI$$

(1) Filter the data to California.

Answer

```r
# store california data to drought_ca
drought_ca <- drought %>% filter(state == "CA")
```

(2) Filter the California data to the year 2019. To extract year, you can use

```r
year(drought$date)
```

Answer

```r
# store 2019 california data to drought_ca2019
drought_ca2019 <- drought_ca %>% filter(year(date) == 2019)
```

(3) Compute the average DSCI of California in the year 2019.

Answer

```r
drought_ca2019 %>%
  mutate(dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>% # create a column of dsci
  summarize(avg = mean(dsci)) # compute average
```

```
##        avg
## 1 40.55302
```

(4) Compute the average DSCI of every state in the year 2019.

Answer

```r
# store the answer to dsci_2019
dsci_2019 <- drought %>%
  filter(year(date) == 2019) %>%
  mutate(dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  group_by(state) %>% # group by states
  summarize(avg = mean(dsci))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
# inspect the first few rows
head(dsci_2019)
```

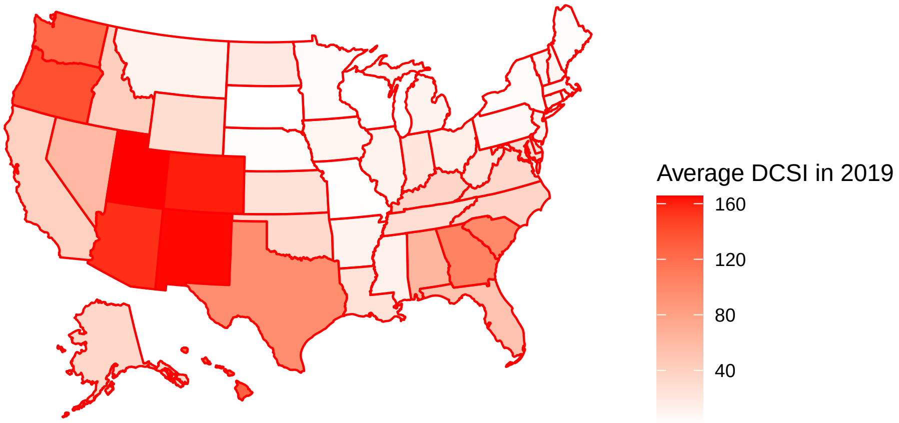
```
## # A tibble: 6 x 2
##   state   avg
##   <chr> <dbl>
## 1 AK     34.5
## 2 AL     63.4
## 3 AR     10.2
## 4 AZ    155.
## 5 CA     40.6
## 6 CO    161.
```

(5) Display your findings in part (4) in a visualization.

Answer Answer 1 (Thanks to Ashley)

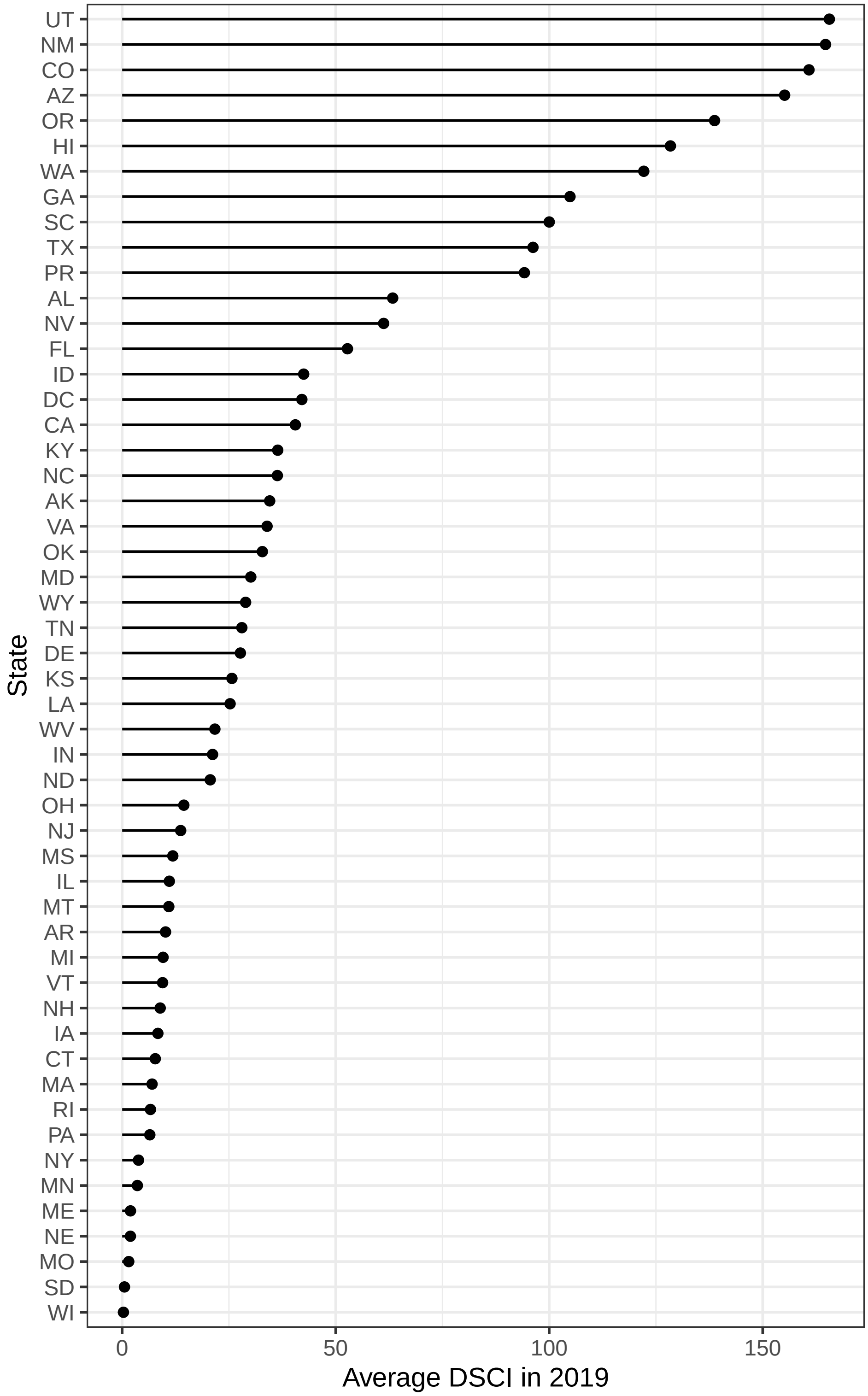
```
# install.packages("usmap") # install package usmap if you do not have the package
library(usmap)
plot_usmap(data = dsci_2019, values = "avg", color = "red")+
  scale_fill_continuous(
    low = "white", high = "red", name = "Average DCSI in 2019",
    label =scales::comma
    ) +
  theme(legend.position = "right")
```

```
## Warning: Use of `map_df$x` is discouraged. Use `x` instead.

## Warning: Use of `map_df$y` is discouraged. Use `y` instead.

## Warning: Use of `map_df$group` is discouraged. Use `group` instead.
```



Answer 2 A lollipop plot showing drought level of each state, ordered from highest to lowest.

```
ggplot(dsci_2019) +
  geom_point(aes(x = reorder(state, avg), y = avg)) +
  geom_segment( aes(x=reorder(state, avg), xend=reorder(state, avg), y=0, yend=avg)) +
  xlab("State") +
  ylab("Average DSCI in 2019") +
  coord_flip() +
  theme_bw()
```



(b) Is drought seasonal?

(1) In the state of california, compute the average drought severity index (see question (a) part (1)) for each month in the year 2019. Which month sees the most severe drought? Use `month()` to extract month from a `Date` object.

```r
month(drought$date)
```

Answer The drought was most severe in January.

```r
# average DSCI in each month
drought_ca2019 %>%
  transmute(month = month(date),
            dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  group_by(month) %>%
  summarize(avg_dsci = mean(dsci))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 12 x 2
##    month avg_dsci
##    <dbl>    <dbl>
##  1     1   245.
##  2     2    60.9
##  3     3     7.78
##  4     4     6.09
##  5     5     5.97
##  6     6     4.32
##  7     7     4.32
##  8     8     7.09
##  9     9    12.9
## 10    10    16.0
## 11    11    74.1
## 12    12    19.9
```

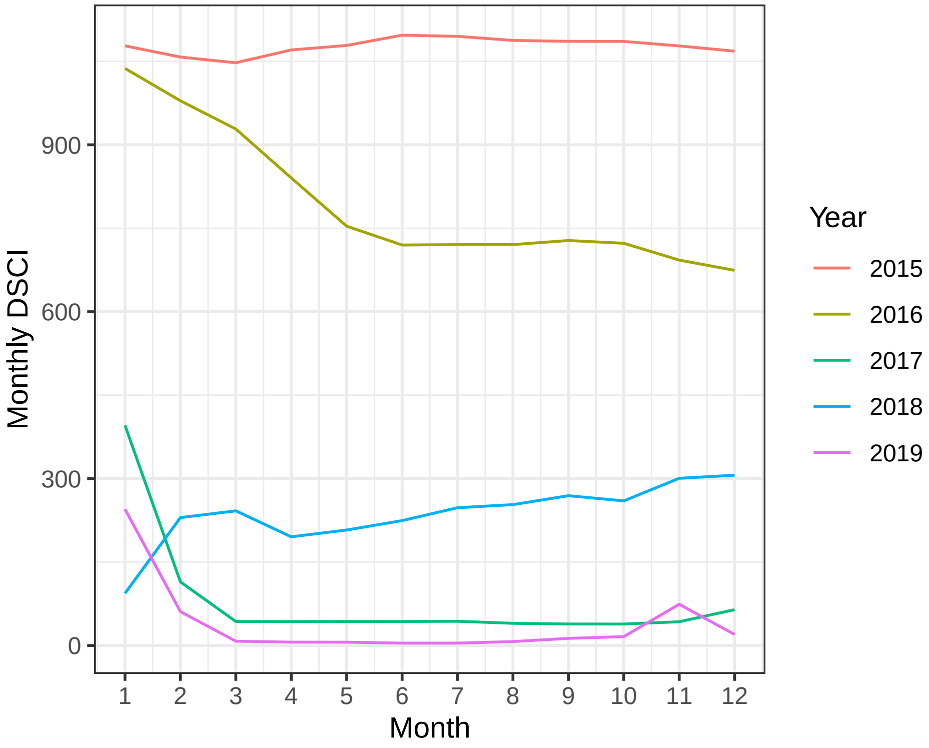
(2) Is your finding in part (1) consistent in all five years from 2015 - 2019?

Answer We plot the average dsci in each month. Drought was more severe in January in 2016, 2017 and 2019. In 2015 drought was most severe in June and 2018 in December. Thus January does not always see the most severe drought.

```r
ca_drought_monthly <- drought %>% filter(state == "CA") %>%
  mutate(year = year(date),
         month =month(date),
         dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  filter(year != 2014) %>%
  group_by(year, month) %>%
  summarize(avg = mean(dsci))
```

```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```

```r
ggplot(ca_drought_monthly) +
  geom_line(aes(x = month, y = avg, color = factor(year))) +
  scale_x_continuous(breaks = 1:12) +
  ylab("Monthly DSCI") +
  xlab("Month") +
  labs(color = "Year")+
  theme_bw()
```



(c) Your question

Describe a question you have about drought levels and use summary statistics or a visualization from this data to answer it.

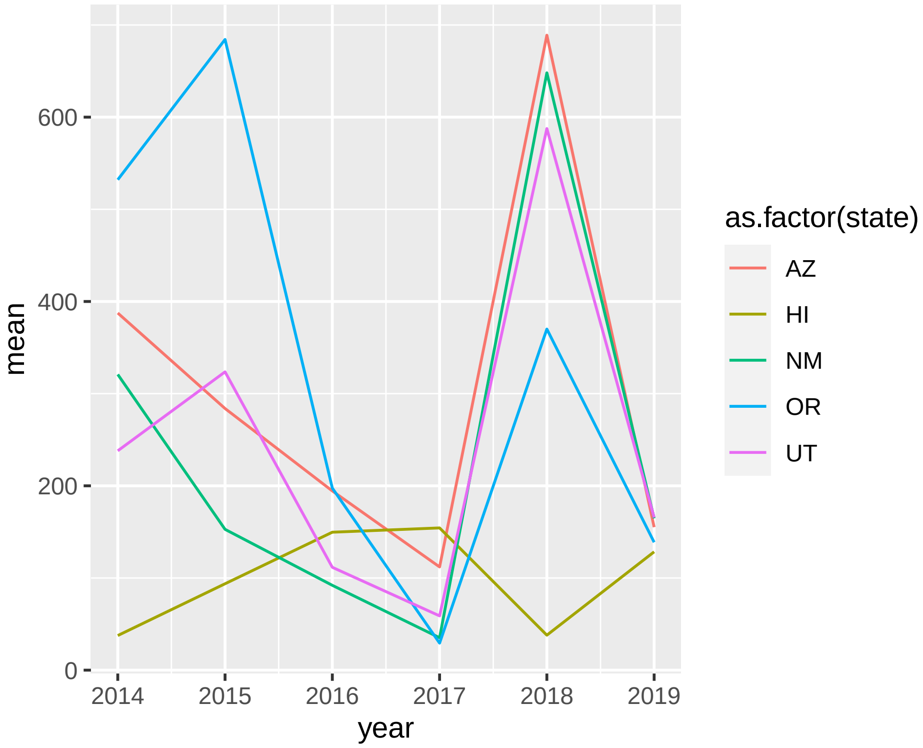
Answer thanks to Arden (with minor edits from TA)

After making the graph of which states had the highest DSCI in 2019, I was curious abou the 5 most severe states. Specifically, I was curious about whether the states followed the same drought trends over the years and whether there were certain years where droughts were more severe among all of them, or less severe among all the them. I made a line graph of the mean DSCI per year with different colors representing the states with the highest DSCI in 2019. From the graph, it is clear that in 2016 and 2017, droughts were less severe among all of the states except for Hawaii. In 2018, all of the states had more severe droughts, except for Hawaii. So, this graph shows that the year did affect some states in the same ways and there were worse and better years acorss the board. But, it also shows that it may depend on the region because Hawaii did not follow the same trends as the contiguous US states, Arizona, New Mexico, Oregon, and Utah.

```
topfive <- drought %>%
  mutate(
    year = year(date),
    dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  select(year, state, dsci) %>%
  filter(state %in% c("UT", "NM","AZ","OR", "HI")) %>%
  group_by(year, state) %>%
  summarize(mean = mean(dsci))
```

```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```

```
ggplot(topfive, aes(x = year, y = mean)) +
  geom_line(aes(group = state, color = as.factor(state)))
```


600
400
200
0
mean
2014
2015
2016
2017
2018
2019
year
as.factor(state)
AZ
HI
NM
OR
UT