STATS 60 Summer 2020: HW3

Remarks

Please show your work, and provide brief explanation for your answers.

If a problem has multiple parts, each will be graded separately. We will not subtract grades if your solution to one part is correct but the number you use from previous parts is wrong.

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)
library(lubridate)
```

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 (we don't really multiply numbers to years). Genre is qualitative and nominal scale. Total is quantitative and ratio scale.

(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, perhaps they like reading more so the answer here is biased upwards. Another issue is this may not be comprehensive, people may buy books, or borrow from other libraries. You can complement this data with surveys about how many books people read, and what percentage do they borrow.

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

$$\Pr(\text{all healthy}) = (1 - 0.014)^{20} = 0.754.$$

(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.

$$\begin{aligned}
\Pr(P) &= \Pr(P \cap H) + \Pr(P \cap \text{not}H) \\
&= \Pr(P \mid H)\Pr(H) + \Pr(P \mid \text{not}H)\Pr(\text{not}H) \\
&= 0.002 \times 0.754 + 0.999 \times (1 - 0.754) \\
&= 0.247.
\end{aligned}$$

(c)

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

Answer

$$\begin{aligned}
\Pr(\text{not}H \mid P) &= 1 - \Pr(H \mid P) \\
&= 1 - \frac{\Pr(H \cap P)}{\Pr(P)} \\
&= 1 - \frac{0.002 \times 0.754}{0.247} \\
&= 0.994.
\end{aligned}$$

The posterior odds is

$$\frac{\Pr(\text{not}H \mid P)}{\Pr(H \mid P)} = \frac{0.994}{1 - 0.994} = 165.7.$$

(d)

Suppose we test 5000 separate groups, which corresponds to testing 100,000 people. What's the probability that 1200 groups test positive? What about 1500?

Extra How many groups do you expect to test positive?

Answer Use the binomial distribution.

```
dbinom(1200, 5000, 0.247)
```

```
## [1] 0.006807583
```

```
dbinom(1500, 5000, 0.247)
```

```
## [1] 2.472365e-18
```

Problem 3 Sampling distribution of the mean

The Current Population Survey 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 each income bracket. For example, if you download the data table, you can see there are 4,283 households out of 128,579 whose yearly income is less than $5000.

From this source, we generate a hypothetical population of size 128, 579 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 128,579 and contains hypothetical incomes of every household in the survey. We will work with this hypothetical income population `all_income`.

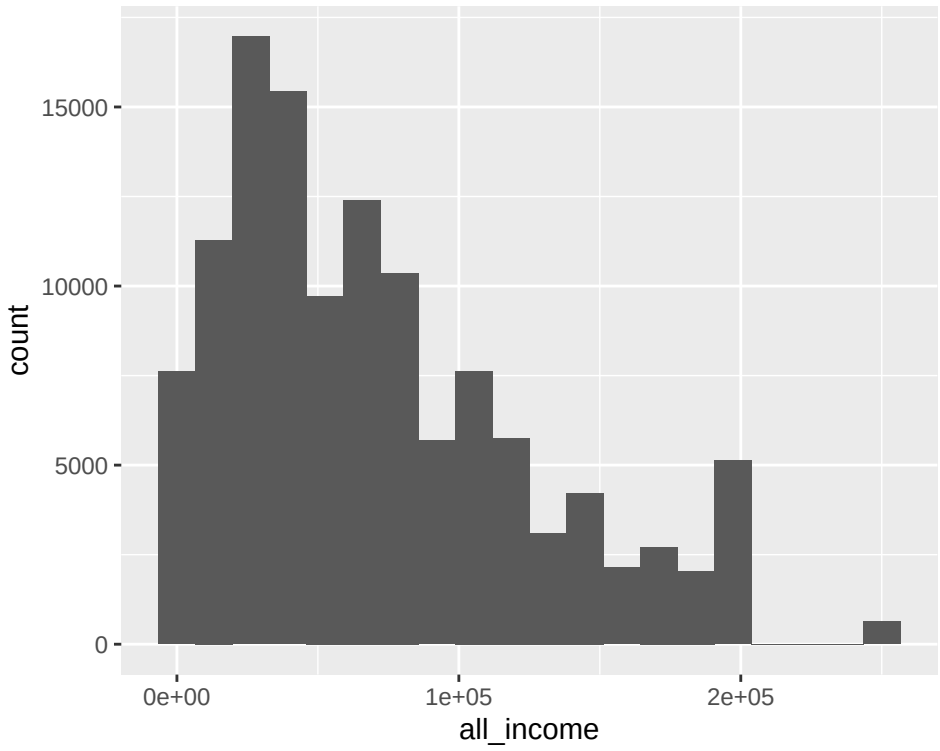
```
income <- c(seq(0, 200000, by = 5000), 250000) # income brackets
mid_income <- c((income[-42] + income[-1]) / 2, 250000) # mid-points of the bracket
N <- c(4283, 3337, 5510, 5772, 5672, 5469, 5822, 5404, 5195, 4839,
       5300, 4417, 4604, 3999, 3795, 3950, 3349, 3064, 3102, 2581,
       2866, 2449, 2318, 1971, 2004, 1780, 1678, 1426, 1414, 1316,
       1492, 978, 1161,  970 , 905 , 835 , 772 , 686  ,584 , 565 ,
       4572, 637) # number of households in each bracket
all_income <- rep(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. 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)
```



(b)

What is the population mean, median and interquartile range?

Answer

```
mean(all_income)
```

```
## [1] 71035.88
```

```
median(all_income)
```

```
## [1] 60000
```

```
IQR(all_income)
```

```
## [1] 70000
```

(c)

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

Answer We need to subtract the mean and divide by the SD. The SD is 5.4543011×10^4.

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

```
## [1] -0.5690165
```

(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)
```

(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.

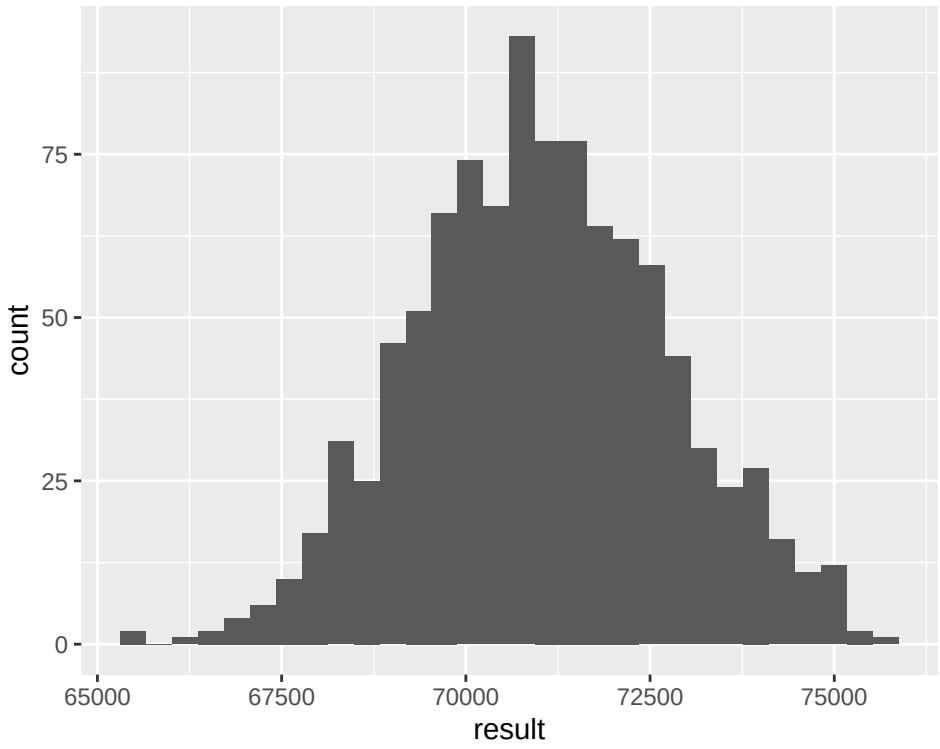
```
result <- numeric(length = 10) # a vector of 100 zero
for(i in 1:10){ # iterate i from 1 to 100
   # write the action to perform in one loop inside of bracket { }
  result[i] <- i
}
result
```

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

Answer The histogram of the sample mean is approximately normal, its mean is 5.5 and its standard deviation is 3.0276504

```
B <- 1000
result <- numeric(length = B)
for(i in 1:B){
  result[i] <- mean(sample(all_income, 1000))
}
ggplot()+geom_histogram(aes(x = result))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



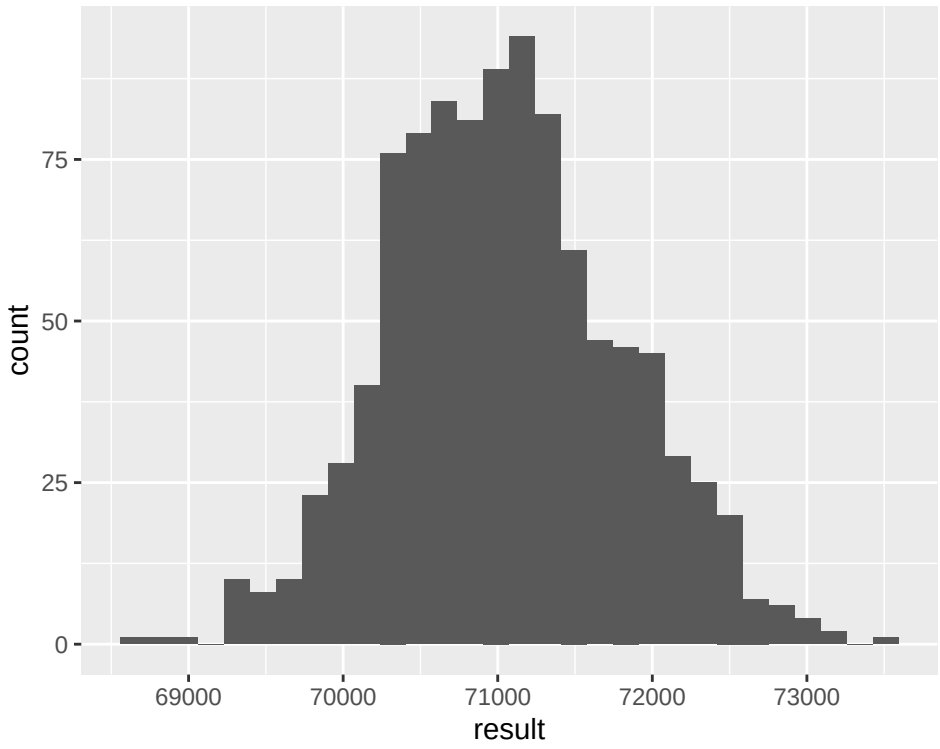
(f)

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

Answer

```
B <- 1000
result <- numeric(length = B)
for(i in 1:B){
  result[i] <- mean(sample(all_income, 5000))
}
ggplot()+geom_histogram(aes(x = result))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mean(result)
```

```
## [1] 71047.99
```

```
sd(result)
```

```
## [1] 748.3762
```

Problem 4 Drought data

This exercise explores more the US drought data we looked at in the Rlab. The data we use is downloaded from US drought monitor website, the state drought level in terms of percent area from 2015-01-01 to 2019-12-31.

```
drought <- read.csv("/Users/zq/Desktop/Teaching/60_summer2020/homework/hw3/drought.csv")
```

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

```
drought <- drought %>%
  mutate(
    date = ymd(MapDate),
    state = as.character(StateAbbreviation)
    ) %>%
  select(date, state, None, D0:D4)
```

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

```
glimpse(drought)
```

```
## Observations: 13,624
## Variables: 8
## $ date  <date> 2019-12-31, 2019-12-24, 2019-12-17, 2019-12-10, 2019-12-03, ...
## $ state <chr> "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", "...
## $ None  <dbl> 93.18, 93.18, 93.18, 92.48, 92.48, 91.17, 91.17, 90.29, 90.29...
## $ D0    <dbl> 6.82, 6.82, 6.82, 7.52, 7.52, 8.83, 8.83, 9.71, 9.71, 9.71, 1...
## $ D1    <dbl> 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 3.96, 3.96, 3.96, 4...
## $ D2    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 2.00, 2...
## $ D3    <dbl> 0.00, 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, 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 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

```
drought_ca <- drought %>% filter(state == "CA")
```

(2)

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

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

Answer

```
drought_ca2019 <- drought_ca %>% filter(year(date) == 2019)
```

(3)

Compute the average DSCI of California in the year 2019.

Answer

```
drought_ca2019 %>%
  mutate(dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  summarize(avg = mean(dsci))
```

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

(4)

Compute the average DSCI of every state in the year 2019.

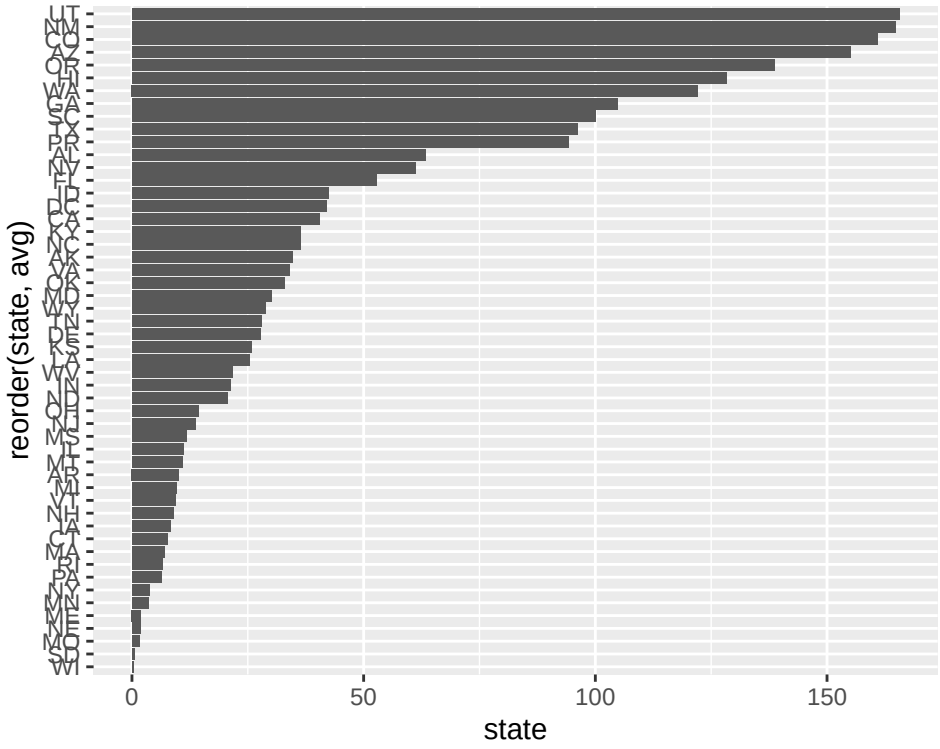
```
dsci_2019 <- drought %>%
  filter(year(date) == 2019) %>%
  mutate(dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  group_by(state) %>%
  summarize(avg = mean(dsci))
```

(5)

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

Answer

```
ggplot(dsci_2019) +
    geom_col(aes(x = reorder(state, avg), y = avg)) +
    ylab("state") +
    coord_flip()
```



(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.

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

Answer Drought in January is the most severe.

```
drought_ca2019_month <- drought_ca2019 %>%
  transmute(month =month(date),
         dsci = D0 + 2*D1 + 3* D2 + 4*D3 + 5*D4) %>%
  group_by(month) %>%
  summarize(avg = mean(dsci))

drought_ca2019_month
```

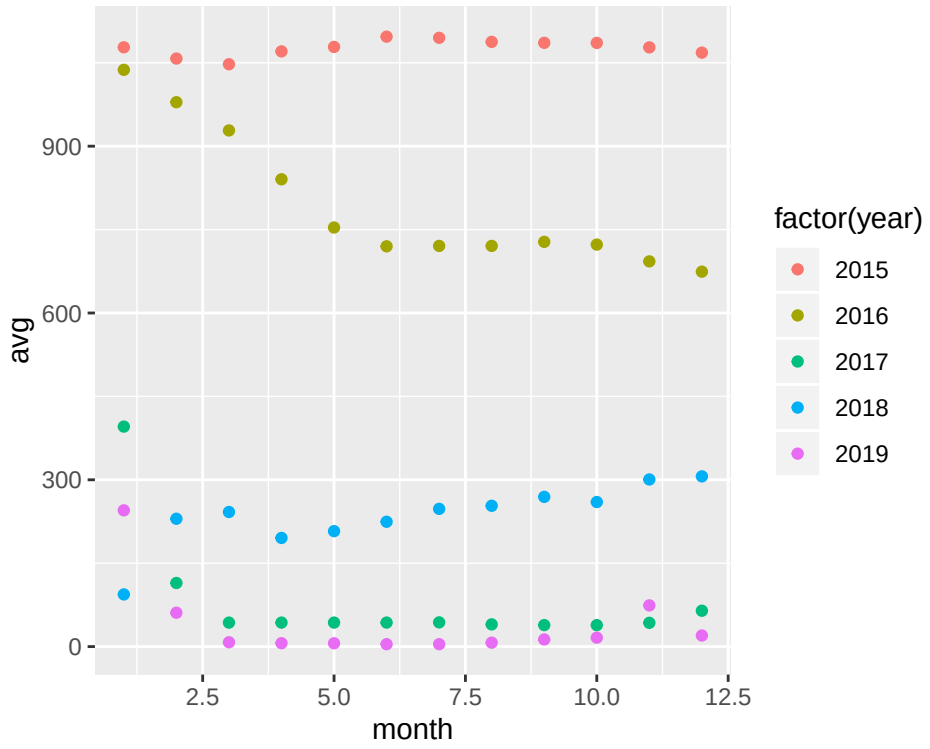
```
## # A tibble: 12 x 2
##    month    avg
##    <dbl>  <dbl>
##  1     1 245.
##  2     2  60.9
##  3     3   7.78
##  4     4   6.09
##  5     5   5.97
##  6     6   4.32
##  7     7   4.32
##  8     8   7.09
##  9     9  12.9
## 10    10  16.0
## 11    11  74.1
## 12    12  19.9
```

(2)

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

```
library(tidyr)
monthly_drought_ca <- 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))
```

```
ggplot(monthly_drought_ca) +
  geom_point(aes(x = month, y = avg, color = factor(year)))
```



In 2015 the worst drought occured in June and in 2016 the worst drought occured in January. Since there is a decreasing trend in the overall drought level, to fully understand the seasonal effect we should remove the trend first.

(c) Your question

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