My answers to assignment of week 2

Hongzhe Zhang_JJ

Question 1: Blood types probability

The following table shows the prevalence of ABO block groups for the U.S. Caucasian and American Black populations. The table specifies the probabilities of a specific race and blood type. For example, the probability that a randomly selected individual is black with blood type O is 0.098.

	A	B	AB	O
Caucasian	0.352	0.064	0.024	0.360
American Black	0.054	0.040	0.008	0.098

Question 1.1

What is the probability that a randomly selected individual does not have blood type AB?

```
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
(1 - 0.024 - 0.008) %>%
  print()
```

```
## [1] 0.968
```

The probability that a randomly selected individual does not have blood type AB is 0.968.

Question 1.2

What is the probability that two people selected at random both have blood type B?

```
((0.064 + 0.04) ^ 2) %>%
  print()
```

```
## [1] 0.010816
```

Assume we select with replacement, the probability that two people selected at random both have blood type B is approximately 0.011.

Question 1.3

Are the events "blood type B" and "American black" race statistically independent?

```
((0.064 + 0.04) * (0.054 + 0.040 + 0.008 + 0.098) == 0.04 )%>%
  print()
```

```
## [1] FALSE
```

Two events are independent if the occurrence of one does not affect the probability of the other. As the probability of event "blood type B American black" is not equal to the multiplication of probabilities of events "blood type B" and "American black", theu are not statistically independent.

Question 1.4

Are the events "blood type O" and "blood type A" mutually exclusive?

Yes, they are. As the probability for events "blood type o" and "blood type A" is zero.

Question 1.5

Given an example of two events that are not mutually exclusive.

The events "blood type B" and "American black" are not mutually exclusive.

Question 1.6

What is the conditional probability of "blood type A" given that the race is "American black"?

```
(0.054 /(0.054 + 0.040 + 0.008 + 0.098)) %>%
  print()
```

```
## [1] 0.27
```

Given that the race is "American black", the conditional probability of "blood type A" is

Question 2: Death probability

The following data records the reported proportions of 20- to 25-year-old White males causes of death within a five year period. In other words, it gives the probability of death within five years.

Cause	Probability
Suicide	0.00126
Homicide	0.00063
Auto accident	0.00581
Leukemia	0.00023
all other causes	0.00788

Question 2.1

What is the probability that a White male aged 20 to 25 dying from any cause of death within the next five years?

```
(0.00126 + 0.00063 + 0.00581 + 0.00023 + 0.00788) %>%
  print()
```

```
## [1] 0.01581
```

The probability that a White male aged 20 to 25 dying from any cause of death within the next five years according to the given records is 0.01581.

Question 2.2

Out of 100,000 white males in the 20 to 25 age group, how many deaths would you expect in the next 5 years from leukemia?

```
(0.00023 * 100000) %>%
  print()
```

```
## [1] 23
```

Out of 100,000 white males in the 20 to 25 age group, according to the data given, I expect 23 deathes in the next 5 years from leukemia.

Question 2.3

Given that a White male aged 20 to 25 years has died, what is the most likely cause of death? Assume that nothing else is known. Provide rationale for your answer.

Given that a White male aged 20 to 25 years has died, the most likely cause of death from my perspective is Auto Accident. First of all, the probability of Auto accident is the biggest out of all the existing explicit causes. Secondly, although the casue "all other causes" is actullay larger, it is unlikely that it will contain any causes with probability larger than the probability of Leukemia.

Question 3: Lab test one

Suppose that a lab test values are normally distributed with $\mu = 80$ and $\sigma = 3$.

Question 3.1

What is the probability that a randomly selected test has a value that is within one standard deviation of the mean? Round to three decimal places.

```
(pnorm(83, 80, 3) - pnorm(77, 80, 3)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.683
```

The probability that a randomly selected test has a value that is within one standard deviation of the mean is 0.683.

Question 3.2

What is the probability that a randomly selected test has a value that is within two standard deviations of the mean? Round to three decimal places.

```
(pnorm(86, 80, 3) - pnorm(74, 80, 3)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.954
```

The probability that a randomly selected test has a value that is within two standard deviation of the mean is 0.954.

Question 3.3

What is the probability that a randomly selected test has a value that is within three standard deviations of the mean? Round to three decimal places.

```r
(pnorm(89, 80, 3) - pnorm(71, 80, 3)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.997
```

The probability that a randomly selected test has a value that is within three standard deviation of the mean is 0.997.

Question 4: Lab test two

Suppose that another lab test values record changes from a reference value and these are normally distributed with $\mu = -16$ and $\sigma = 1.5$.

Question 4.1

What is the probability that a randomly selected test has a value that is within one standard deviation of the mean? Round to three decimal places.

```r
(pnorm(-14.5, -16, 1.5) - pnorm(-17.5, -16, 1.5)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.683
```

The probability that a randomly selected test has a value that is within one standard deviation of the mean is 0.683.

Question 4.2

What is the probability that a randomly selected test has a value that is within two standard deviations of the mean? Round to three decimal places.

```r
(pnorm(-13, -16, 1.5) - pnorm(-19, -16, 1.5)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.954
```

The probability that a randomly selected test has a value that is within two standard deviation of the mean is 0.954.

Question 4.3

What is the probability that a randomly selected test has a value that is within three standard deviations of the mean? Round to three decimal places.

```r
(pnorm(-11.5, -16, 1.5) - pnorm(-20.5, -16, 1.5)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.997
```

The probability that a randomly selected test has a value that is within three standard deviation of the mean is 0.997.

Question 4.4

How do your answers compare to those of Question 3? Does this surprise you?

The answers are correspondingly the same. It does not suprise me, as they both are normally distributed. If we are caculating the probability that the random variables falling in the range within standard deviations of the mean, we are actually in a sense adjusting them into the standard normal distribution. As they were both adjusted into standard normal distribution, the probabilities should be the same as the units of the standard deviation assigned are the same.

Question 5: Randomization verification

A clinical trial was designed to determine whether radiation therapy or surgery was more effective in reducing disease recurrence in men with prostate cancer. Patients were randomized to each treatment in a 1:1 manner, which implies the chance a patient is randomized to either arm is 0.50. Suppose that in the first 20 patients randomized at a particular site participating in the trial, 17 patients were assigned to the surgery arm. The site complains that the randomization does not appear to be working. Does the site's suspicion have merit?

When you answer this, you should consider all scenarios that would support the suspicions that the randomization is broken. Specifically, what other outcomes out of 20 would be more extreme to the one observed. You might want to calculate the probability of getting the result observed or one more extreme.

```
##numbers used t support claims
pbinom(17, 40 ,0.5) %>%
  print()
```

```
## [1] 0.2147953
```

```
dbinom(20, 40, 0.5) %>%
  print()
```

```
## [1] 0.1253707
```

To randomize the patients is to given the patients different treatments randomly, which means that the numbers of patients recieving surgery arm or radiation therapy are both binomially distrbuted with n = 40 and p = 0.5. The sites complains the quality of the randomization from my perspectiive is merely because the numbers of people for each group are not even(not equals 20). Though 20 is the expectation of the distribution of number of people getting a specific treatment, the actual probability of this is merely 0.125, which is just above the half of the probability that the number of people is less than or equal to 17. Practically, the number of people could vary between 0 and 40.

Hypertension

The remainder of the questions are relevant to this scenario.

The proportion of U.S. adults over the age of 20 with hypertension is 0.335, or 33.5%.

Question 6: Hypertension

Suppose you take a random sample of 64 U.S. adults over the age of 20.

Question 6.1

What is the chance that 32 people have hypertension? (Round to 4 decimal places.)

```r
(dbinom(32, 64, 0.335)) %>%
  round(4) %>%
  print()
```

```
## [1] 0.0025
```

The chance that 32 people have hypertension 0.0025.

Question 6.2

What is the chance that more than 20 individuals in the sample have hypertension? (Round to 3 decimal places.)

```r
(1 - pbinom(20, 64, 0.335)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.593
```

The chance that more than 20 individuals in the sample have hypertension is equal to 1 - the chance that less than 21 individuals in the sample have hypertension, which is 0.593.

Question 6.3

What is the chance that between 21 and 24 individuals, inclusive, out of 64 have hypertension? (Round to 3 decimal places.)

```r
-(pbinom(21, 64, 0.335) - pbinom(24, 64, 0.335)) %>%
  round(3) %>%
  print()
```

```
## [1] 0.28
```

The chance that between 21 and 24 individuals, inclusive, out of 64 have hypertension is equal to 0.28.

Question 7: Hypertension

Now suppose you take your random sample (of size 64) from one city in the United States. You get a sample proportion of 0.211.

Question 7.1

How likely are you to get a sample proportion this small or smaller if this city is representative of the US population?

```r
pbinom(0.211 * 64, 64, 0.335) %>%
  round(3) %>%
  print()
```

```
## [1] 0.015
```

As this city is representative of the US population, we use p = 0.335. We get that the chance that I get a sample proportion this small or smaller is 0.006.

Question 7.2

What do you conclude?

As it is highly unlikely to get a sample like this, and if it is not plausible to conduct a another similar sample; I may conclude that the this city is not representative of the US population and actually has a smaller proportion of people having hypertension.

Question 8: Hypertension

Give a range for which you are 95% certain would contain the sample proportion of adults with hypertension if you were to take a random sample of 64 adults in the U.S. over the age of 20. (Round to three decimal places.)

```
(qbinom(0.95, 64, 0.335)) %>%
  round(3) %>%
  print()
```

```
## [1] 28
```

```
(print(29 / 64)) %>%
  round(3)
```

```
## [1] 0.453125
```

```
## [1] 0.453
```

As the probability for the number of people having hypertension is below or equal to 29 is 0.95; [0,0.453] is a range for which I are 95% certain would contain the sample proportion of adults with hypertension

Question 9: Hypertension

We want to compare the sampling distribution of our sample proportion to that of a normal distribution. The normal distribution should have the same mean and standard deviation as the sampling distribution. To do this, we want to use `qqplot()`. We will simulate the sampling distribution of both the sampling proportion as well as the normal distribution. After using `qqplot()`, use `qqline()` to add a line to the plot.

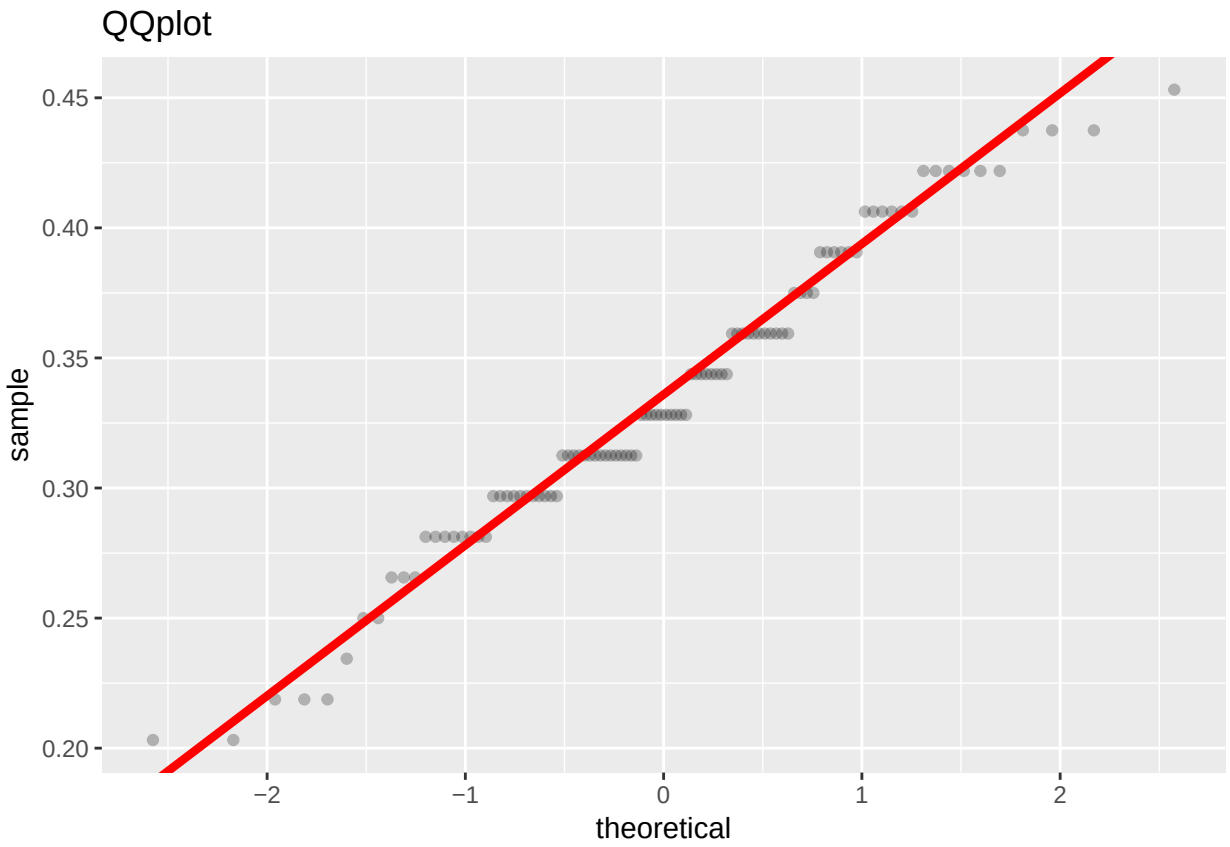
Question 9.1

Make a qqplot comparing the sampling distribution of the sample proportion to that of a normal distribution, including a line that indicates a perfect relationship.

```
simu_data <- (rbinom(100, 64, 0.335)/64)

slope <- diff(quantile(simu_data, c(0.25, 0.75))) / diff(qnorm(c(0.25, 0.75)))
intercept <- quantile(simu_data, 0.25) - slope * qnorm(0.25)
simu_data_df <- data.frame(simu_data)

simu_data_df %>%
ggplot(aes(sample = simu_data)) +
  stat_qq(alpha = 0.25) +
  geom_abline(aes(slope = slope, intercept = intercept), col = "red", lwd = 1.5) +
  ggtitle("QQplot")
```



Question 9.2

How do these two distributions compare?

Two distributions bisically match. Except the mild signs of deviation in both ends indicating the sample is mildly heavy-tailed.

Question 9.3

To verify your answer to Question 9.2, use a normal distribution with the same mean and standard deviation of the binomial distribution to determine the likelihood that we observe between 21 and 24 individuals, inclusive, out of 64 have hypertension (Question 6.3). How do these values compare?

```
(pnorm(24, 64 * 0.335, sqrt(64 * 0.335 * 0.645) )- pnorm(21, 64 * 0.335, sqrt(64 * 0.335 * 0.645)) )%>%
  print()
```

```
## [1] 0.3014961
```

```
(pbinom(24, 64, 0.335 ) - pbinom(20, 64 , 0.335 )) %>%
  print()
```

```
## [1] 0.385344
```

The values differ a lot, though it cannot be considered as an evidence against our previos conclusion. The reason why sampling distribution was camparable to normal distribution was due to CLT which is only applied when the sampling population is large(for example 100). Only one observation which is the case of question 9.3 is far from being enough.