Midterm - Solutions

EPIB607 - Inferential Statistics - McGill University

2020-10-27

- The exam consists of 4 problems and 8 pages. The exam is worth 100 points, and the point value for each question is displayed in square brackets.
- This is a 2 hour exam. You will be given one extra hour to upload your solutions to each question as a pdf document to Crowdmark. Note that you must upload a separate pdf for each question and code.
- You must respond to the questions in an RMarkdown document in the provided template and knitted to pdf. The template will automatically put the code at the end of the document. The code should be uploaded separately to Question 5 on Crowdmark.
- Be sure to read the questions carefully. Some parts of a problem statement may ask for more than one calculation.
- Some parts of a question may require the answer to an earlier part. If you cannot solve the earlier part, you can still receive partial credit for the later parts; make up a reasonable answer for the earlier part to use in subsequent parts of the problem.
- Show your work and explain your reasoning; the final answer is not as important as the process by which you arrived at that answer. We can more easily give partial credit if you have written out your steps clearly.
- State your assumptions in the context of the problem. Avoid generic statements when possible.
- Label your axes, provide units and answer in complete sentences. Clear and concise answers will be rewarded.
- All your work must be your own. Use of the internet is not permitted. Collaboration is strictly forbidden, including any discussion about resolving technical issues related to knitting files (other than with the course instructor). You are allowed to access the internet only for the course materials on myCourses. Using the internet or obtaining help from anyone else is considered Cheating as per Article 17 of the Code of Student Conduct and Disciplinary Procedures.
- Answers must be in your own words. Plagiarism is not acceptable and we will very likely detect if an answer has been copied from a website.
- Finally, I would like to remind you about Academic Integrity. McGill places a great deal of importance on honest work, the art of scholarship, and the fair treatment of all members of the university community, and demands a rigid insistence on giving credit where credit is due. Offences such as cheating and breaches of research ethics undermine not only the value of our collective work, but also the academic integrity of the University and the value of a McGill degree.

1 (15 points) Flu Season

According to data from the CDC, about 37.1% of adults (individuals 18 years of age or older) in the United States and 57.9% of children (individuals between 6 months and 17 years of age) in the United States received a flu vaccine during the 2017-2018 flu season.

a. [4] Consider a random sample of 50 adults.

   i. Calculate the probability that exactly 20 adults received a flu vaccine.

   The probability that exactly 20 out of 50 adults received a flu vaccine is 0.105.

```
dbinom(20, 50, 0.371)
```

```
## [1] 0.1047823
```

   ii. Calculate the probability that exactly 30 adults did not receive a flu vaccine.

This is equivalent to the probability that exactly 20 out of 50 randomly sampled adults received a flu vaccine, which was calculated above as 0.105:

```
dbinom(30, 50, 1 - 0.371)
```

```
## [1] 0.1047823
```

b. [4] Consider a random sample of 20 children.

   i. What is the probability that at most 10 children received a flu vaccine?

   This probability asks for "at most 10", which equals the probabilities that either 0, 1, 2, ... or 10 children received a flu vaccine out of 20 children. The probability that at most 10 children out of 20 randomly sampled children received a flu vaccine is 0.310.

```
## [1] 0.309678
```

   ii. What is the probability that at least 11 children received a flu vaccine?

   This probability asks for "at least 11", which equals the probabilities that either 11, 12, 13, ... or 20 children received a flu vaccine out of 20 children. The probability that at least 11 children out of 20 randomly sampled children received a flu vaccine is 0.690. Note that the event of "at least 11" is the complement of "at most 10".

```
pbinom(10, 20, 0.579, lower.tail = FALSE)
```

```
## [1] 0.690322
```

```
# alternatively
1 - pbinom(10, 20, 0.579)
```

```
## [1] 0.690322
```

c. [4] State two assumptions you needed to make in order to answer parts a) and b). Briefly comment on the extent to which those assumptions were reasonable.

   In order to use the binomial distribution, it was necessary to assume that the people picked in a random sample do not have contact with each other, such that the decision to get a flu vaccine is independent; this is reasonable if the sample is selected from a large, geographically diverse population, such as the the adult population in Northeast California, for instance.

   Using the binomial distribution also requires assuming that each person in a sample has an equal probability of receiving a flu vaccine. This may be less reasonable since, for example, access to flu vaccines may be different across groups of people in the United States. Someone whose workplace or university campus offers free flu vaccines may be more likely to receive a vaccine than someone who needs to make an appointment at a doctor's office to be vaccinated.

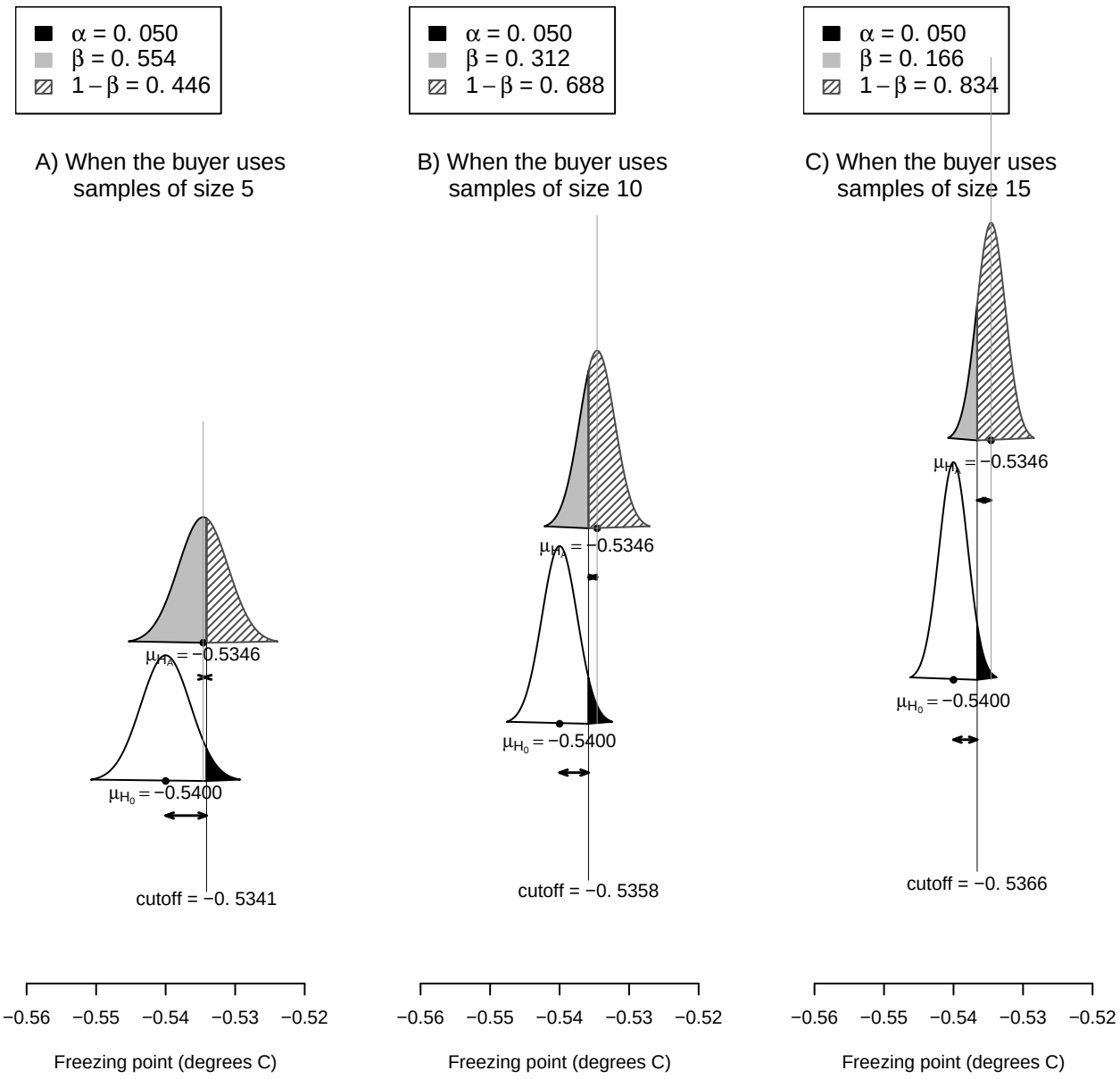
d. [3] Consider a random sample of $n = 70$ individuals, which consists of 50 adults and 20 children. Let Z represent the total number of individuals who received the flu vaccine in the sample of 70 individuals. Does Z follow a binomial distribution? Explain your answer.

No, Z does not follow a binomial distribution. A binomial distribution requires a constant probability p of success in all n trials. Here, the sample of 70 individuals explicitly consists of 20 children and 50 adults; it is known that the probability of success differs between children and adults. This is an example of a *stratified* random sample, in which subgroups are sampled randomly, then combined.

Note that Z would follow a binomial distribution if a simple random sample of 70 children and adults were taken, in which the number of children and adults in the sample were not fixed; under this scenario, the probability of success p would be somewhere in between 0.417 and 0.593, and calculated based on the ratio of children to adults in the population.

2 (20 points) The Cheese Maker

A cheese maker buys milk from several suppliers. It suspects that some suppliers are adding water to their milk to increase their profits. Excess water can be detected by measuring the freezing point of the "liquid". The freezing temperature of natural milk varies according to a Gaussian distribution, with mean $\mu = -0.540°$ Celsius (C) and standard deviation $\sigma = 0.008°$C. Added water raises the freezing temperature toward 0°C, the freezing point of water. Assume the supplier is mixing 99% milk with 1% water and that the σ would remain unchanged. Such a mixture would freeze at -0.5346°C. In the Figure below, null and alternative distributions are plotted for different samples sizes used by the buyer to detect cheating.



a. [3] What do α, β and $1 - \beta$ (shown in the Figure above) represent in the context of the problem?

α is the Type I error. In the context of the problem, it is the probability of accusing the supplier of adding excess water to their milk, when in fact they are not.

β is the Type II error, i.e., the probability that the supplier is in fact cheating (adding water to the milk), but you are not be able to detect this cheating.

$1 - \beta$ is the power, i.e., the probablity of correctly detecting that the supplier was in fact adding water to the milk (cheating).

b. [4] Why does the cutoff value decrease as we move from left (Panel A) to right (Panel C) in the Figure above?

For a fixed α, the cutoff is a function of the standard error which decreases as we increase the sample size: cutoff = -0.54 + 1.96 * 0.008 / $\sqrt{n}$

c. [4] Explain why β decreases as we move from left (Panel A) to right (Panel C) in the Figure above.

For a fixed α, β decreases as sample size increases. This is again due to the standard error of the mean which decreases with increasing sample size, making the null and alternative distributions narrower (i.e. more separation between the null and the alternative). Therefore as the distributions become narrower, it becomes easier to detect cheating (power increases), therefore type II error decreases.

d. [3] Describe the factors which affect $1 - \beta$.

The factors which affect power (for a fixed α) can be seen by the balancing equation:

$$\Delta = z_{1-\alpha/2} \times SEM + z_{1-\beta} \times SEM,$$

where $SEM = \sigma/\sqrt{n}$. That is, the effect size Δ, the noise σ, and the sample size n all affect power.

e. [3]

i) What would happen to the cutoff value if the supplier was mixing 98% milk with 2% water? Nothing. The cutoff value is fixed under the null hypothesis and is not affected by what the supplier does.

ii) What would happen to the cutoff value if $\sigma = 0.010$?

The cutoff would increase:

$$-0.54 + 1.96 * 0.010/\sqrt{5} = -0.531,$$

making it harder to reject the null. As the noise increases, so does the cutoff.

f. [3]

i) What would happen to $1 - \beta$ if the supplier was mixing 98% milk with 2% water? Justify your answer.

Power would increase at a fixed α. The signal (Δ) increases which makes it easier to detect cheating.

ii) What would happen to $1 - \beta$ if $\sigma = 0.010$. Justify your answer.

Power would decrease at a fixed α. The noise (σ) increases which makes it harder to detect cheating, as the null and alternative distributions overlap more.

3 (35 points) Youth Risk Behavioral Surveillance System (YRBSS)

This question uses data from the Youth Risk Behavioral Surveillance System (YRBSS), a yearly survey conducted by the US Centers for Disease Control to measure health-related activity in high-school aged youth. The dataset `yrbss` contains responses from the 12,579 participants in 2013 for a the following variables:

- `age`: age in years
- `gender`: gender of participant, recorded as either `female` or `male`
- `height`: height, in meters (1 m = 3.28 ft)
- `weight`: weight, in kilograms (1 kg = 2.2 lbs)
- `hispanic`: Indicates a participant's status as hispanic (TRUE) or not (FALSE)
- `race`: Indicates a participant's race
- `physically.active.7d`: Days per week that the participant is physically active
- `hours.tv.per.school.day`: The average number of hours of TV watched by the participant on a schoolday.
- `strength.training.7d`: Out of the 7 days preceding the survey, how many days the participant did exercises to strengthen or tone their muscles (such as push-ups, sit-ups, or weight lifting).

For the purposes of this question only, we will be treating the 12,579 individuals in `yrbss` as the **entire target population** of high-school aged youth. The dataset can be read into `R` as follows

```
yrbss <- readr::read_csv(file = "yrbss.csv")
dplyr::glimpse(yrbss)
```

```
## Rows: 12,579
## Columns: 9
## $ age                    <dbl> 15, 15, 15, 15, 15, 14, 15, 15, 15, 15, 15,...
## $ gender                 <chr> "female", "female", "female", "female", "fe...
## $ height                 <dbl> 1.73, 1.60, 1.50, 1.57, 1.65, 1.88, 1.75, 1...
## $ weight                 <dbl> 84.37, 55.79, 46.72, 67.13, 131.54, 71.22, ...
## $ hispanic               <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
## $ race                   <chr> "Native Hawaiian or Other Pacific Islander"...
## $ physically.active.7d   <dbl> 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 7, 7, 7, 4...
## $ hours.tv.per.school.day <chr> "5+", "2", "3", "5+", "5+", "5+", "5+", "do...
## $ strength.training.7d   <dbl> 0, 0, 1, 0, 2, 0, 3, 0, 3, 0, 0, 7, 7, 7, 3...
```

a) [2] List the variable types for `weight`, `hispanic`, `race`, and `hours.tv.per.school.day`

Weight is a continuous variable, hispanic is a binary indicator or categorical variable or logical (any of these three is correct), race is a categorical variable, hours.tv.per.school.day is categorical ordinal since there is a natural ordering of the categories 1 all the way to 5+

b) The following graph plots the weight against height colored by race.

i. [1.5] List all the scales being used in the figure.

three scales: x-axis position for height, y-axis position for weight, and color for race

ii. [2] Describe any patterns you see in the figure.

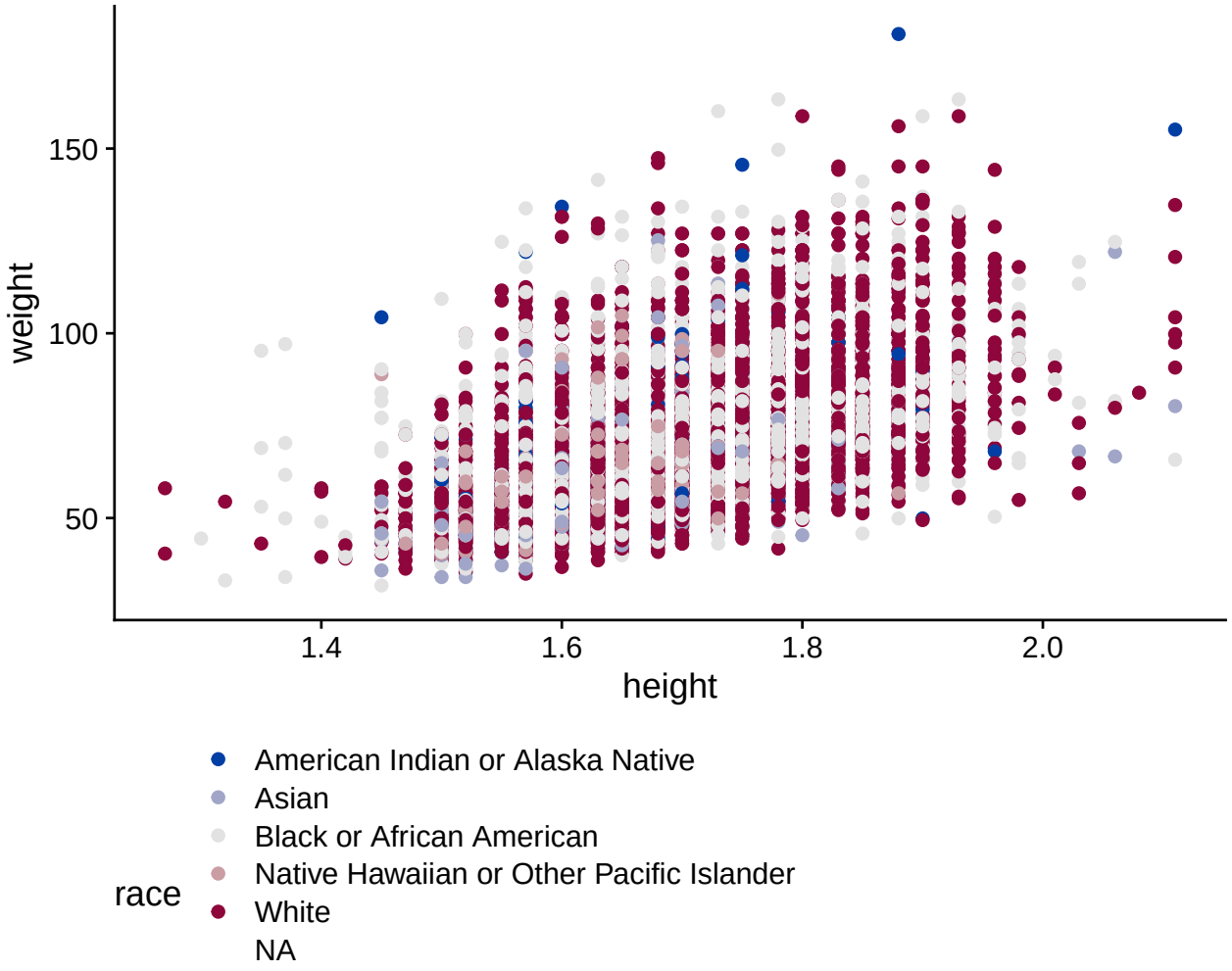
- Height has a very clear discreteness to it, indicating that it may have been rounded.
- There seems to be a positive correlation between height and weight (as expected).
- This correlation between height and weight doesn't seem to be influenced by race.

iii. [1] Which color palette should have been used and why?

A qualitative color palette should have been used here because the goal would be to distinguish between the different races. The diverging color palette that was used is not appropriate since there isn't a natural ordering to the races.

iv. [2.5] Provide a short critique of the figure. What would you have done differently?

- Change the color palette to qualitative
- Facet by race (i.e. 6 different height weight scatterplots) - it is too difficult to see any race effect because too many of the points overlap.
- Since height and weight are correlated, plot boxplots of BMI by race instead.



c) [2] Calculate the mean and standard deviation for weight.

```
mu <- mean(yrbss$weight)

# population sd should be divided by n
sigma <- sqrt(mean((yrbss$weight - mu)^2))
```

(This should be a complete sentence describing the population. Part marks for simply stating what the mean and sd are). The population mean weight (μ) of high-school aged youth is 67.9065 kg and with population standard deviation equal to 16.89754 kg.

d) [4] Can you calculate a 95% confidence interval for the mean calculated in c) ? If yes, calculate and interpret the interval and list your assumptions (if any). If not, explain why.

No, because we have the entire population, therefore there is no sampling error and a confidence interval doesn't have any meaning.

e) [2] Calculate the mean weight from a random sample of size 16 from the population. You can use the `dplyr::sample_n()` function to obtain a random sample from the population.

```
set.seed(123) # for reproducibility
sample.size <- 16
yrbss.sample <- sample_n(yrbss, size = sample.size, replace = FALSE)
sample.mean <- mean(yrbss.sample$weight)
```

The random sample of 16 high-school aged youth has a mean weight of 77.73562 kg.

f) [5] Can you calculate a 95% confidence interval for the mean calculated in e) ? If yes, calculate and interpret the interval and list your assumptions (if any). If not, explain why.

Yes it makes sense to calculate a 95% confidence interval since we have a sample from the population. Therefore we would prefer to provide some measure of uncertainty around our sample mean. We can use either the t-distribution if we estimated sigma from our sample and assume CLT has kicked in; we can use the z-distribution if we used the sigma from the population and assume CLT has kicked in; or we can use the bootstrap if we don't think the CLT has kicked in. All of these choices are valid as long as choice was appropriately justified.

Interpretation: we are 95% confident that the interval from 66.18 kg to 89.29 kg captures the true mean weight of high-school aged youth.

```
# using t-dist with +/- formula
sample.sd <- sd(yrbss.sample$weight)
sample.se <- sample.sd / sqrt(sample.size)
(sample.ci <- sample.mean + qt(p = c(0.025, 0.975), df = sample.size - 1) * sample.se)
```

```
## [1] 66.17678 89.29447
```

```
# using t-dist with canned function
t.test(yrbss.sample$weight, conf.level = 0.95)$conf.int
```

```
## [1] 66.17678 89.29447
## attr(,"conf.level")
## [1] 0.95
```

```
# using linear regression (which also uses the t-distribution)
confint(lm(weight ~ 1, data = yrbss.sample))
```

```
##                2.5 %   97.5 %
## (Intercept) 66.17678 89.29447
```

```
# using z-dist with population sigma
qnorm(p = c(0.025, 0.975), mean = sample.mean, sd = sigma / sqrt(sample.size))
```

```
## [1] 69.45598 86.01527
```

```
# using bootstrap since you can make the argument that CLT hasn't kicked at n=16
B <- 1000
R <- replicate(B, {
  yrbss.sample %>%
    sample_n(size = sample.size, replace = TRUE) %>%
    summarise(r = mean(weight)) %>%
    pull(r)
})
quantile(R, probs = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 67.86938 87.79948
```

g) One of your colleagues created the following figure. They tell you that it's based on 25 random samples of size 16. For each of the 25 samples, they calculated both an 80% and 96% confidence interval for the mean weight of high-school aged youth.

i. [3] Explain why the 96% confidence intervals are wider than the 80% confidence intervals.

Because 96% confidence intervals need to be large enough to capture the true mean weight of high-school aged youth 96 times out of 100, whereas 80% confidence intervals only need to capture the truth, 80 times out of 100 on average. In other words, you can afford to be off target more often with 80% CI compared to 96% CI.

ii. [5] Are you able to verify that these are in fact 80% and 96% confidence intervals? If yes, explain how. If not, explain what information you would need in order to perform this verification.

Yes we can do this verification since we have the population mean μ. We simply need to check how many of the 25 intervals contain the population mean weight of 67.9065 kg. We would expect about 20 out of the 25 80% confidence intervals to contain μ and about 24 out of 25 96% confidence intervals to contain μ.

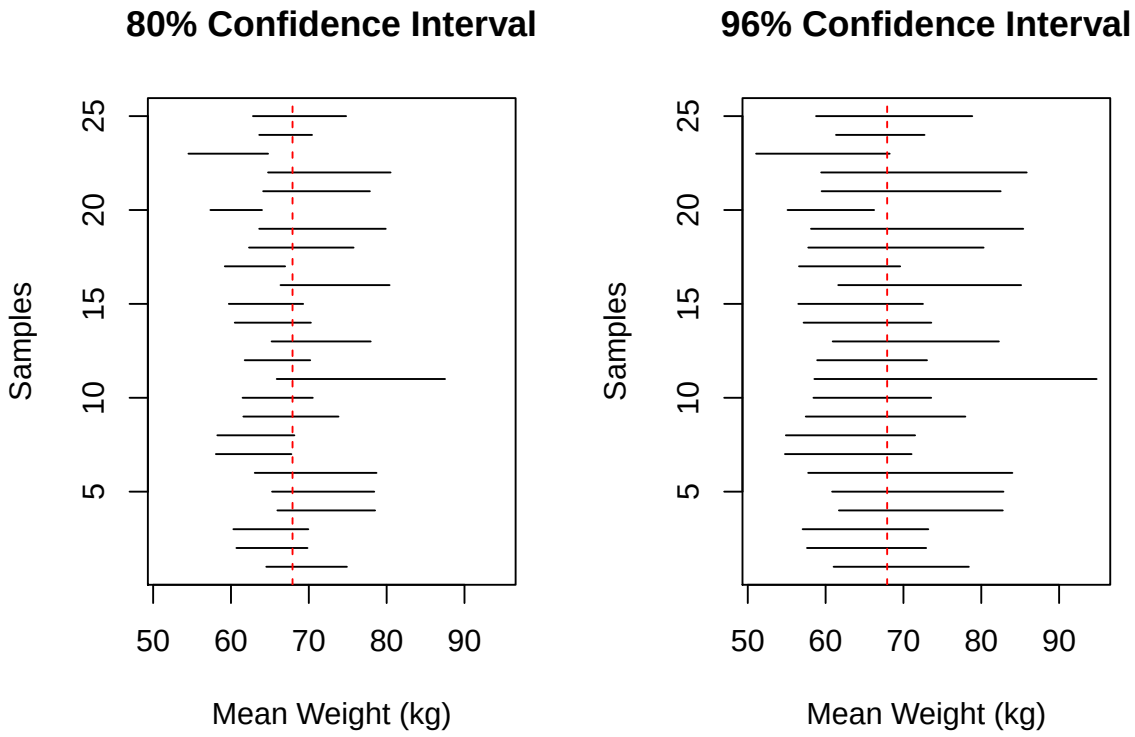
```r
# note that this code is shown only for pedagogical purposes.
# it was not necessary to actually perform the check for the midterm
set.seed(234)
number.samples <- 25

# Draw 25 samples of size 16 each and calculate 80%, 96% t-based confidence intervals
R.16.80.96 <- replicate(number.samples, {
  sample.16 <- yrbss %>%
    dplyr::sample_n(size = 16) %>%
    dplyr::pull(weight)

  c(t.test(sample.16, conf.level = .80)$conf.int,
    t.test(sample.16, conf.level = .96)$conf.int)
}) %>% t()

# check coverage for 80% interval
mean(sapply(1:number.samples, function(i){
  (mu >=  R.16.80.96[i,1]) & (mu <=  R.16.80.96[i,2])

}))
```

```
## [1] 0.84
```

```r
# check coverage for 96% interval
mean(sapply(1:number.samples, function(i){
  (mu >=  R.16.80.96[i,3]) & (mu <=  R.16.80.96[i,4])
}))
```

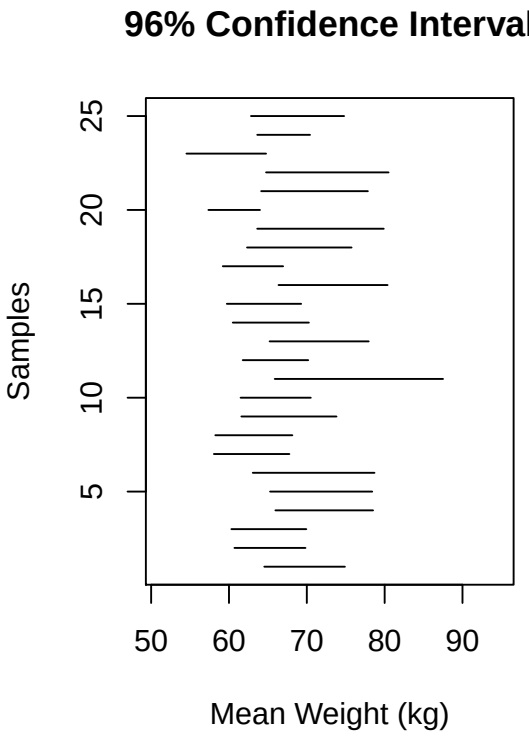
```
## [1] 0.96
```

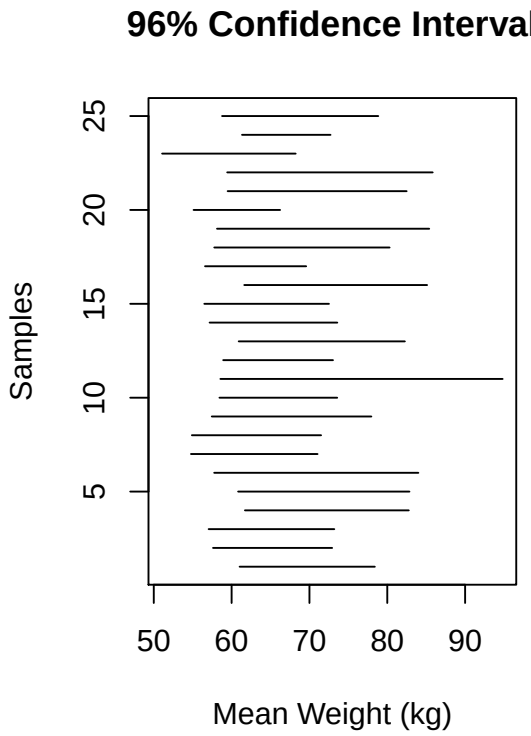
```r
# plot the confidence intervals with a vertical line for mu
par(mfrow = c(1,2))
plot(range(R.16.80.96), c(1, number.samples),
     col="black", typ="n",
     main = "80% Confidence Interval",
     ylab="Samples",
     xlab="Mean Weight (kg)")
segments(R.16.80.96[,1], 1:number.samples, R.16.80.96[,2], 1:number.samples)
abline(v = mu, lty = 2, col = "red")

plot(range(R.16.80.96), c(1, number.samples),
     col="black", typ="n",
     main = "96% Confidence Interval",
     ylab="Samples",
     xlab="Mean Weight (kg)")
segments(R.16.80.96[,3], 1:number.samples, R.16.80.96[,4], 1:number.samples)
abline(v = mu, lty = 2, col = "red")
```



h) [5] A few months later, your colleague informs you that they might have made some errors in their code. They think the left panel of the figure should also be titled *96% Confidence Interval*. They are confident that the right panel is correct and send you the updated figure below. What are the possible explanations for this, if any ? (i.e. how can it be that 96% confidence intervals in the figure below have **very** different widths?)

- By chance, or by cheating, your colleague selected samples that were not very variable, leading to a very small sample standard deviation, and thus, a narrow confidence interval
- Your colleague used a much larger sample size for the left panel than the n=16 used in the right panel


96% Confidence Interval
Samples
5
10
15
20
25
50
60
70
80
90
Mean Weight (kg)


96% Confidence Interval
Samples
5
10
15
20
25
50
60
70
80
90
Mean Weight (kg)

4 (30 points) Anemia

An individual is said to be anemic (or have anemia) if they do not have enough healthy red blood cells to transport oxygen to tissues in the body. According to the US National Heart, Lung and Blood Institute, anemia affects approximately 3 million Americans and is the most common blood disorder in the United States. Anemia can have many causes, including internal bleeding or nutritional deficiencies. Anemia is diagnosed by measuring the level of hemoglobin present in blood; hemoglobin is the protein in red blood cells responsible for oxygen transport.

Loss of energy and easy fatigue are common symptoms of anemia. Individuals with mild anemia may not recognize their symptoms. Detection of anemia typically occurs if a physician notices a change in energy level and requests a blood test, or if a routine blood test indicates low hemoglobin level. Anemia can be particularly serious in infants and young children, causing impairments in mental, physical, and social development.

Anemia is considered a severe public health concern, especially in under-resourced parts of the world where individuals have limited access to healthcare. In this problem, you will examine data from a study examining the health status of 120 children living in a slum area of a large city in Southeast Asia. The aim of the study was to estimate the prevalence of anemia and investigate possible predictors of anemia. For children in the age range of this study, hemoglobin level less than 10.5 g/dL constitutes a diagnosis of anemia. For low-income children in this age range in the United States, the prevalence of anemia is approximately 15%.

The study collected demographic information in addition to measuring health variables. Family income level was recorded as wealth quintile within the sample; for example, a value of `1` for the `wealth` variable indicates that the child's family income level was in the lowest 20% for families with children participating in the study. Iron level was recorded via an assay in which negative values indicate iron deficiency and positive values indicate adequate iron level.

Data from the study are in the file `anemia.csv` file which can be read into `R` as follows:

```
anemia <- readr::read_csv(file = "anemia.csv")
dplyr::glimpse(anemia)
```

```
## Rows: 120
## Columns: 7
## $ sex      <chr> "female", "male", "male", "female", "female", "male", "fem...
## $ wealth   <dbl> 3, 4, 4, 5, 4, 1, 3, 3, 1, 4, 1, 1, 4, 5, 3, 5, 3, 5, 1, 1...
## $ diarrhea <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "No", "No...
## $ whz      <dbl> -2.45, -0.85, -1.92, -0.33, -1.64, 1.20, -3.17, -1.31, -1....
## $ age      <dbl> 21.13, 21.72, 19.71, 17.08, 22.37, 17.12, 12.39, 17.31, 21...
## $ iron     <dbl> -5.68, 5.66, 0.30, -1.72, -1.92, 0.27, -7.80, 3.60, -0.12,...
## $ hb       <dbl> 7.6, 7.6, 7.7, 9.7, 8.5, 9.8, 7.4, 8.5, 11.5, 10.2, 9.8, 1...
```

The following table provides a list of the variables in the dataset and their descriptions.

Variable	Description
`sex`	sex, coded `female` for female and `male` for male
`wealth`	relative measure of family income level
`diarrhea`	coded `Yes` for at least one episode of diarrhea in the past two weeks, and `No` otherwise
`whz`	standardized weight-for-height z-score relative to the national population
`age`	age in months
`iron`	blood iron level, in milligrams per kilogram (mg/kg)
`hb`	blood hemoglobin level, in grams per deciliter (g/dL)

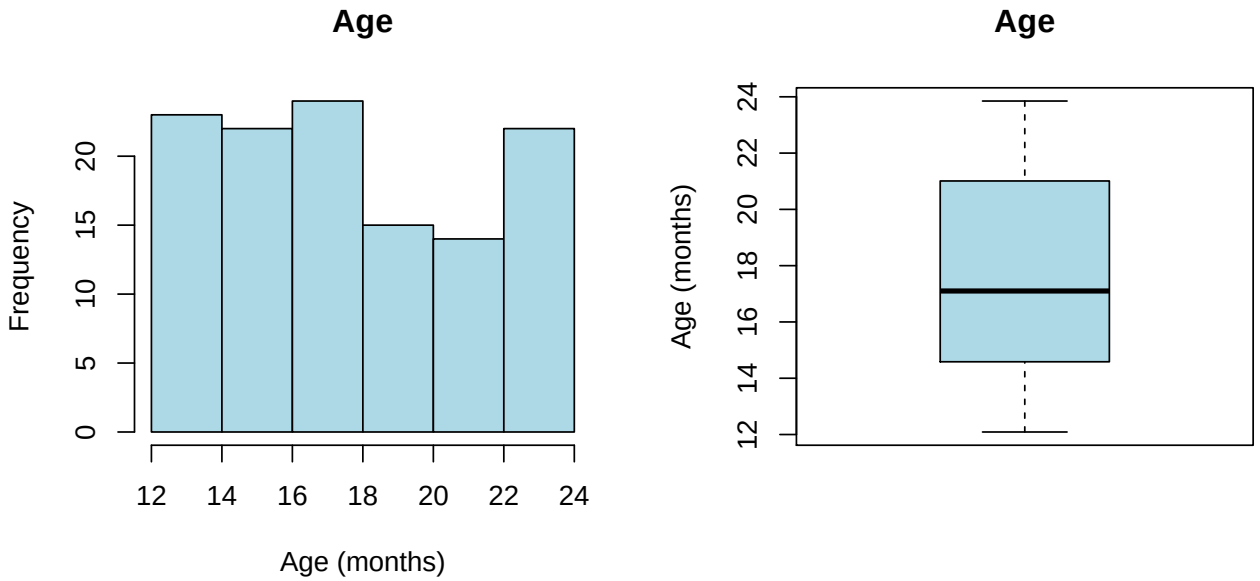
Use the data to answer the following questions.

a) [7] Describe the age and presence of anemia distribution of these children, with reference to appropriate numerical and graphical summaries.

The ages of the observed children range between about 12 months and 24 months, with a median age of

17 months. The mean age is also similar to the medain age. As shown by the boxplot and histogram, the distribution of ages is relatively symmetric and there are no children who are very young or much older relative to the other children in the data.

```
#age
par(mfrow = c(1, 2))
hist(anemia$age, main = "Age", xlab = "Age (months)", col = "lightblue")
boxplot(anemia$age, main = "Age", ylab = "Age (months)", col = "lightblue")
```




```
summary(anemia$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.09   14.62   17.10   17.71   21.00   23.85
```

Children in the age range of this study qualify as anemic with hemoglobin level less than 10.5 g/dL. 81 children in this study sample have hemoglobin level less than 10.5 g/dL; thus, anemia prevalence in the sampled children is 67.5%.

```
# create anemic binary indicator variable
anemia$anemic <- anemia$hb < 10.5
table(anemia$anemic)
```

```
##
## FALSE  TRUE
##    39    81
```

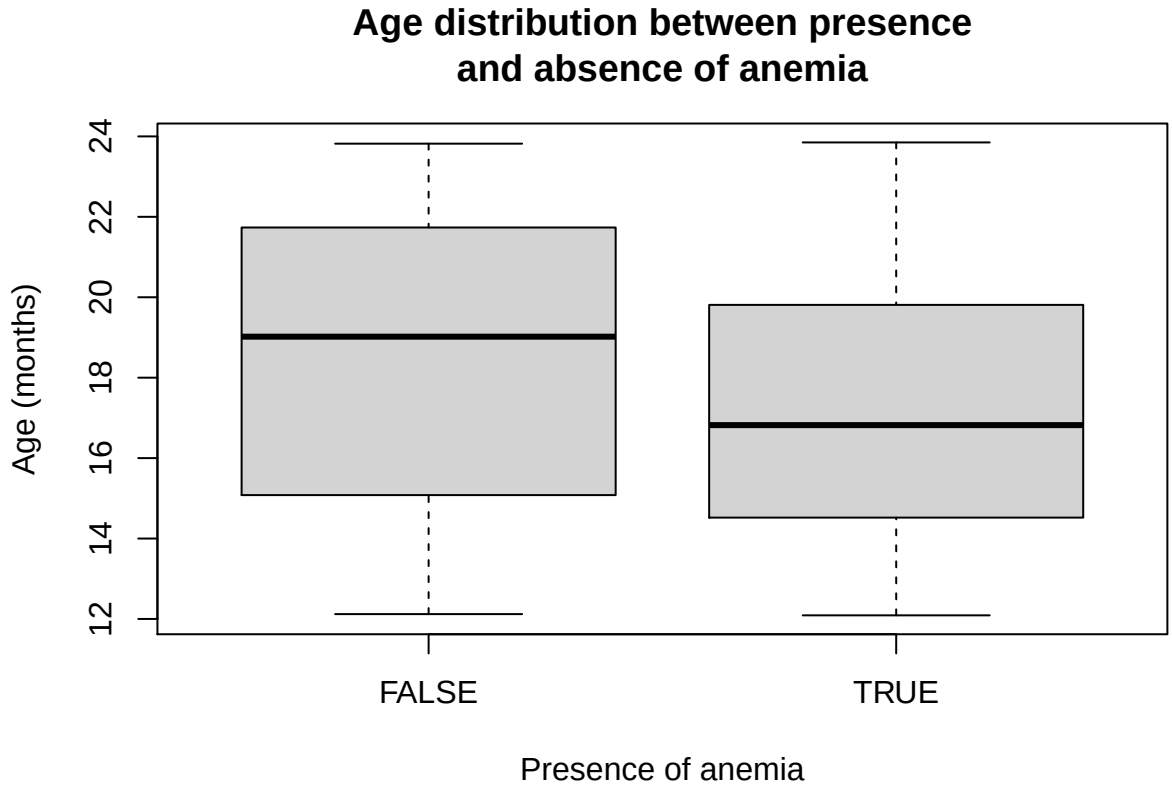
```
(pi.hat <- sum(anemia$anemic == TRUE)/nrow(anemia))
```

```
## [1] 0.675
```

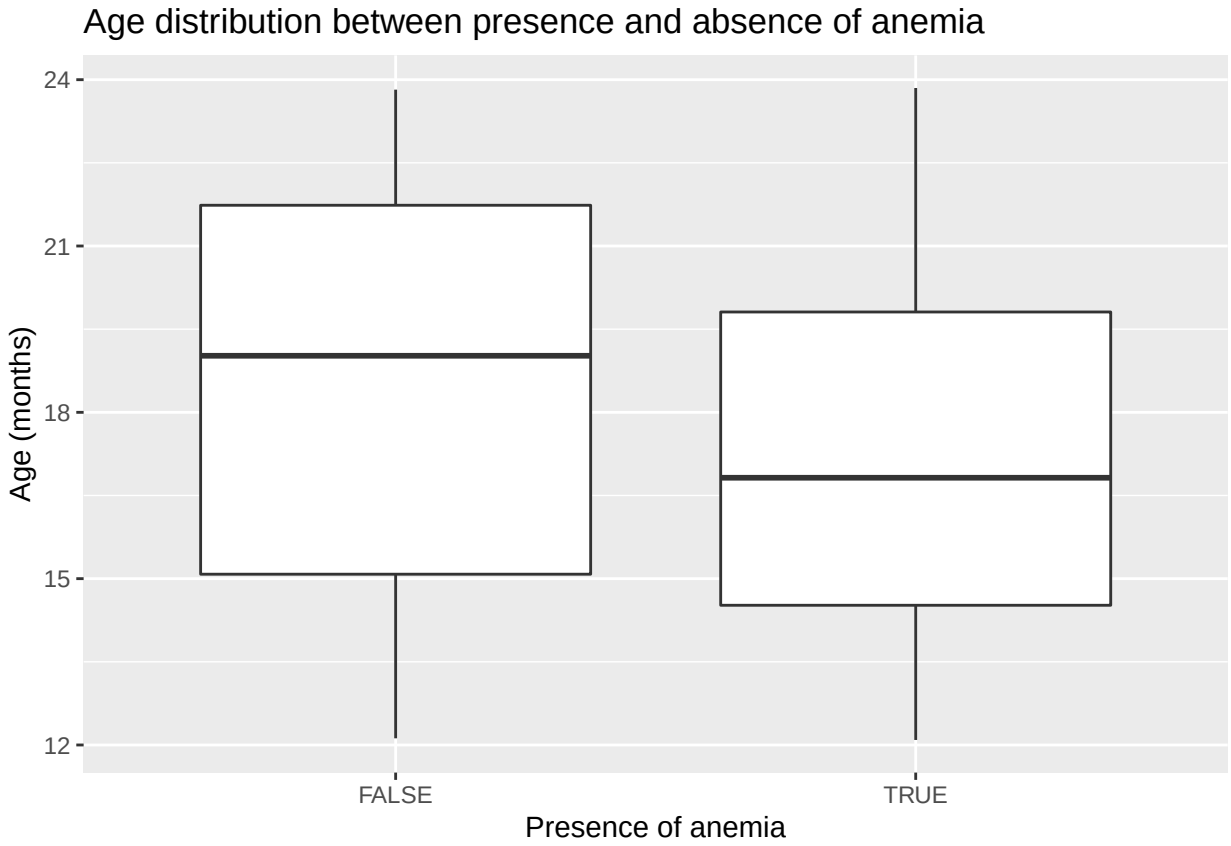
b) [6] Create a figure which compares the age distribution between presence and absence of anemia. Briefly comment on this figure.

We see that the anemic group is slightly younger than the non-anemic group. Both groups have similar minimum and maximum ages, while the median age is a bit higher for the non-anemic group.

```
# using base R
boxplot(age ~ anemic , data = anemia,
        xlab = "Presence of anemia", ylab = "Age (months)",
        main = "Age distribution between presence\nand absence of anemia")
```




```
# using ggplot
ggplot(anemia, aes(x = anemic, y = age)) +
  geom_boxplot() +
  labs(title = "Age distribution between presence and absence of anemia") +
  xlab("Presence of anemia") + ylab("Age (months)")
```



c) [6] Do the data support the claim that anemic children are younger? Briefly describe how you would perform this comparison. (Note: you are not required to actually do the comparison. Simply describe in words what you would do).

To perform this comparison I would use the bootstrap method to find a 95% confidence interval for the difference of median (or mean) ages between both groups. I would first create B bootstrap samples; then for each bootstrap sample I would calculate the median age for both the anemic and non-anemic group and then take their difference. This would result in B bootstrap statistics, and then I could take the 2.5th and 97.5th percentile as the 95% confidence interval. If this interval contained 0, then this would suggest there isn't enough evidence to conclude that there is a difference in ages between groups.

d) [3] How does the prevalence of anemia in the sampled children compare with anemia prevalence in low-income children in the United States?

The prevalence of anemia in the sampled children is 67.5%. This is substantially higher than anemia prevalence in low-income children in the United States (approximately 15%).

e) [8] Do the data support the claim that the proportion of anemic children in this slum area is different than 50%? Justify your answer and state your assumptions (if any).

We conduct a one-sample, two-sided, test of proportions for whether the proportion of anemic children in this area (π) is different from 0.50 at a α=5% level of significance:

$$H_0 : \pi = 0.5 \qquad H_A : \pi \neq 0.5$$

The p-value is 0.000158 with 95% confidence interval [0.583, 0.758]. There sufficient evidence at $\alpha = 0.05$ to reject the null hypothesis that half the children in this area are anemic. Thus, the data support the claim the proportion of anemic children in this region is different than 50%. It is also defensible to conduct a one-sided test of the alternative $H_A : \pi > 0.50$. The p-value is even smaller for the one-sided

test, which leads to rejecting H_0 at $\alpha = 0.05$. You can also use the normal approximation since np and $np(1-p)$ are both greater than 10, and you will also find that the 95% confidence interval does not contain 0.5, and therefore arrive at the same conclusion as above.

```
# two-sided test (Clopper-Pearson confidence interval)
binom.test(x = sum(anemia$anemic), n = nrow(anemia),
           p = 0.50, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  sum(anemia$anemic) and nrow(anemia)
## number of successes = 81, number of trials = 120, p-value = 0.000158
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5834728 0.7576549
## sample estimates:
## probability of success
##                  0.675
```

```
# one-sided test (Clopper-Pearson confidence interval)
binom.test(x = sum(anemia$anemic), n = nrow(anemia),
           p = 0.50, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  sum(anemia$anemic) and nrow(anemia)
## number of successes = 81, number of trials = 120, p-value = 7.902e-05
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.597764 1.000000
## sample estimates:
## probability of success
##                  0.675
```

```
# normal approximation
qnorm(p = c(0.025, 0.975), mean = pi.hat, sd = sqrt(pi.hat * (1-pi.hat) / nrow(anemia)))
```

```
## [1] 0.5911986 0.7588014
```

5 Code

```r
# Question 1 ------------------------------------------------------------
dbinom(20, 50, 0.371)
dbinom(30, 50, 1 - 0.371)
pbinom(10, 20, 0.579)
pbinom(10, 20, 0.579, lower.tail = FALSE)
# alternatively
1 - pbinom(10, 20, 0.579)

# Question 3 ------------------------------------------------------------

yrbss <- readr::read_csv(file = "~/git_repositories/EPIB607/midterm/datasets/yrbss.csv")
mu <- mean(yrbss$weight)
# population sd should be divided by n
sigma <- sqrt(mean((yrbss$weight - mu)^2))

set.seed(123) # for reproducibility
sample.size <- 16
yrbss.sample <- sample_n(yrbss, size = sample.size, replace = FALSE)
sample.mean <- mean(yrbss.sample$weight)

# using t-dist with canned function
t.test(yrbss.sample$weight, conf.level = 0.95)$conf.int


# Question 4 ------------------------------------------------------------
anemia <- readr::read_csv(file = "~/git_repositories/EPIB607/midterm/datasets/anemia.csv")

#age
par(mfrow = c(1, 2))
hist(anemia$age, main = "Age", xlab = "Age (months)", col = "lightblue")
boxplot(anemia$age, main = "Age", ylab = "Age (months)", col = "lightblue")
summary(anemia$age)

# create anemic binary indicator variable
anemia$anemic <- anemia$hb < 10.5
table(anemia$anemic)
pi.hat <- sum(anemia$anemic == TRUE)/nrow(anemia)

# age vs. anemic
boxplot(age ~ anemic , data = anemia,
        xlab = "Presence of anemia", ylab = "Age (months)",
        main = "Age distribution between presence\nand absence of anemia")


# two-sided test (Clopper-Pearson confidence interval)
binom.test(x = sum(anemia$anemic), n = nrow(anemia),
           p = 0.50, alternative = "two.sided")
```