Midterm 2 Solutions - RStudio

Problem 1

We are planning to obtain a simple random sample of size 50 from a population with known standard deviation $\sigma = 1$. We wish to perform a test of $H_0 : \mu = 7$ against $H_a : \mu > 7$ at the 8% significance level.

Part (A)

Find the sampling distribution of (all possible) sample means when the null hypothesis is true. You may assume the Central Limit Theorem holds.

$\bar{x} \sim N(\mu_0, \sigma/\sqrt{n}) = N(7, 1/\sqrt{50})$ or $\bar{x} \sim N(7, 0.14)$.

0.5 pts each for normal distribution, mean = 7, sd = 0.14

Part (B)

Suppose that we consider increasing our significance level from 8% to 10%. If this is the only change in the study methods, which of the following would also be guaranteed to increase?

Answer choices a, the probability of committing a Type I Error, and c, the power of the test, will increase.

Answer choices b, the probability of committing a Type II Error, and e, the critical value, would both decrease. Answer choice d, the p-value, would be unaffected by the significance level.

0.75 pts each for highlighting a and c, -0.5 pts for each wrong answer highlighted

Part (C)

Suppose that, instead, we want to compute a 95% (two-sided) confidence interval for the population mean. Can we find the margin of error for this confidence interval prior to obtaining sample data? If so, compute it. If not, explain why not.

0.25 pts yes, we can find the margin of error for this confidence interval.

1 pt for Margin of Error $= z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)\frac{1}{\sqrt{50}} = 0.277$

Partial credit for using a wrong z critical value or explaining that it is unknown

Partial credit for finding the margin of error but then identifying the confidence interval incorrectly as $\mu_0 \pm E = (6.723, 7.277)$.

Part (D)

Which of the following changes to our study design would result in a smaller margin of error for our confidence interval, assuming other factors remain the same?

Answer choices a, increasing the sample size to 100, and c, increasing our measurement precision so that the population standard deviation is lower, would result in a smaller margin of error.

Answer choice b, raising the confidence level to 99%, would result in a larger margin of error.

0.5 points each for highlighting a and c, -0.5 points for highlighting b

Part (E)

Select the best definition of 95% confidence.

We expect that 95% of all possible confidence intervals will contain our population mean

For parts F-K, assume that we have now actually collected our data (under the original study design) and performed our hypothesis test.

Part (F)

We obtain a z test statistic of 1.55 and a p-value of 0.06. Are these results statistically significant (at our 8% significance level)? Why or why not?

Yes, they are significant, because the p-value is less than our significance level.

Full credit if anyone actually computes the z critical value of 1.405 and compares it to $z_obs = 1.55$

Part (G)

Based on your answer to part (F), in which of the intervals below does the z critical value lie?

(0, 1.55)

Again, full credit if anyone actually computes the z critical value to be 1.405.

Part (H)

Based on your answer to part (F), what should we conclude about the null hypothesis?

Reject the null hypothesis

Part (I)

Oops! We accidentally compute a t test statistic instead of a z test statistic! Which of the following will change due to this mistake?

Answer choices b, the p-value, and c, the critical value, would change.

Answer choice a, the significance level, is set at whatever we want it to be before we compute the test statistic.

0.5 points each for highlighting b and c, -0.5 points for highlighting a

Part (J)

Explain why using a z test statistic would give us more accurate results for this hypothesis test than using a t test statistic.

Knowing the population standard deviation means that the test statistics will follow a normal distribution under H_0. Thus, the appropriate test statistic is the z-score corresponding to the observed sample mean under H_0.

Full credit as long as something is mentioned about the population standard deviation being known

Part (K)

If we continue to use a t-test, how many degrees of freedom should we use in our calculations?

49

Problem 2

A sample of 541 Australian students (12-15 years old) took the Short Mood and Feelings Questionnaire (SMFQ), a test commonly used to measure depressive symptoms in young adolescents. Higher scores on the questionnaire indicate greater severity of depressive symptoms.

The file SMFQ.csv contains the gender and SMFQ scores of the 541 students. Assuming this is a representative sample of all young adolescent students, is there evidence that one gender (either boys or girls) experiences greater depression symptoms than the other?

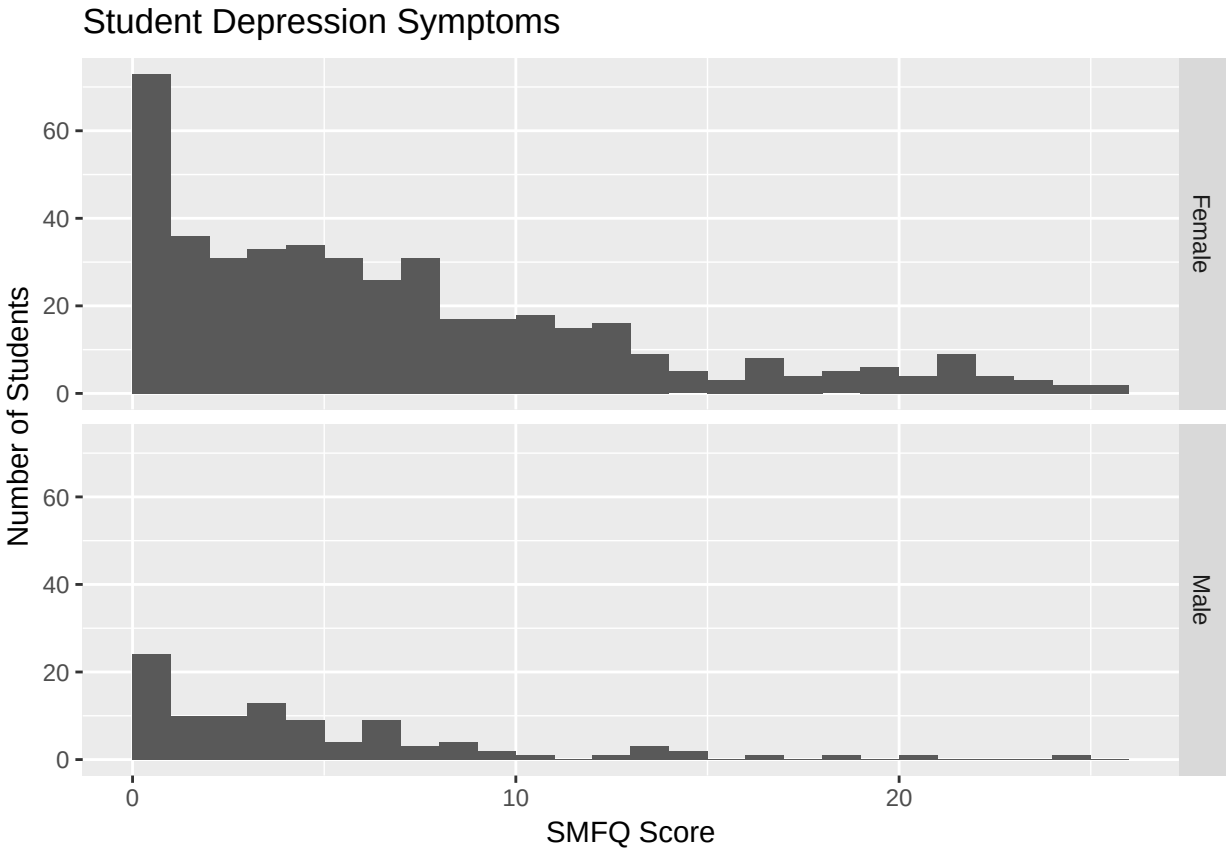
```
library(readr)
smfq <- read_csv("SMFQ.csv")
head(smfq)
```

```
## # A tibble: 6 × 2
##   Gender  SMFQ
##    <chr> <int>
## 1   Male     4
## 2 Female     0
## 3 Female     6
## 4 Female     6
## 5 Female    10
## 6 Female     6
```

A) The appropriate inference framework is two independent samples t for a difference of population means. 1 pt for either a two-sample t CI or a two-sample t test.

B) 0.5 pts each for 2 of the following 3 things: making the histograms (e.g., below), commenting on the extremely skewed distributions, commenting on the uneven sample size

```
library(ggplot2)
smfq_histogram <- ggplot(smfq, aes(x = SMFQ)) + geom_histogram(center = 0.5, binwidth = 1) +
  labs(x = "SMFQ Score", y = "Number of Students", title = "Student Depression Symptoms") +
  facet_grid(Gender~.)
print(smfq_histogram)
```



0.5 pts for indicating that since sample size is large (combined sample size is 541), t procedures are still okay to use

C) 1 pt for, at minimum, the following code and output:

```
t.test(SMFQ ~ Gender, data = smfq)
```

```
##
##  Welch Two Sample t-test
##
## data:  SMFQ by Gender
## t = 3.9763, df = 172.63, p-value = 0.0001028
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.152700 3.425023
## sample estimates:
## mean in group Female   mean in group Male
##             7.339367             5.050505
```

D) 1 pt for one of the following:

95% Confidence interval: (1.153, 3.425)

t = 3.976, p = 0.0001 $<$ 0.05 so reject H0

E) 1.5 pts for one of the following, or similar interpretation:

We are 95% confident that the mean SMFQ score for girls is between 1.153 and 3.425 points higher than the mean SMFQ score for boys. Since this interval does not include 0, we suspect that there is a gender difference; in particular, that girls experience greater depression symptoms than boys.

We reject our null hypothesis of no gender difference. Therefore, we did find evidence of a gender difference in depression symptoms.

Problem 3

A sample of 421 medical school students were asked if they would consider cheating on an exam or helping another student to cheat on an exam.

The file cheaters.csv contains the breakdown of students by gender (Male/Female) and willingness to consider cheating on the exam (Yes/No). The variable Num_Students represents the number of students at each combination of gender and willingness to cheat. Estimate how much more likely male medical students are to consider cheating than female medical students.

```
cheaters <- read_csv("cheaters.csv")
```

```
## Parsed with column specification:
## cols(
##   Gender = col_character(),
##   Would_Cheat = col_character(),
##   Num_Students = col_integer()
## )
```

```
head(cheaters)
```

```
## # A tibble: 4 × 3
##   Gender Would_Cheat Num_Students
##    <chr>       <chr>        <int>
## 1   Male         Yes           56
## 2   Male          No          210
## 3 Female         Yes           13
## 4 Female          No          142
```

A) The appropriate inference framework is for comparing a difference of population proportions. The question asks us to estimate the difference in proportions. 1 pt for a two-sample z CI.

B) 0.5 pts for checking the BINS assumptions in at least 1 population.

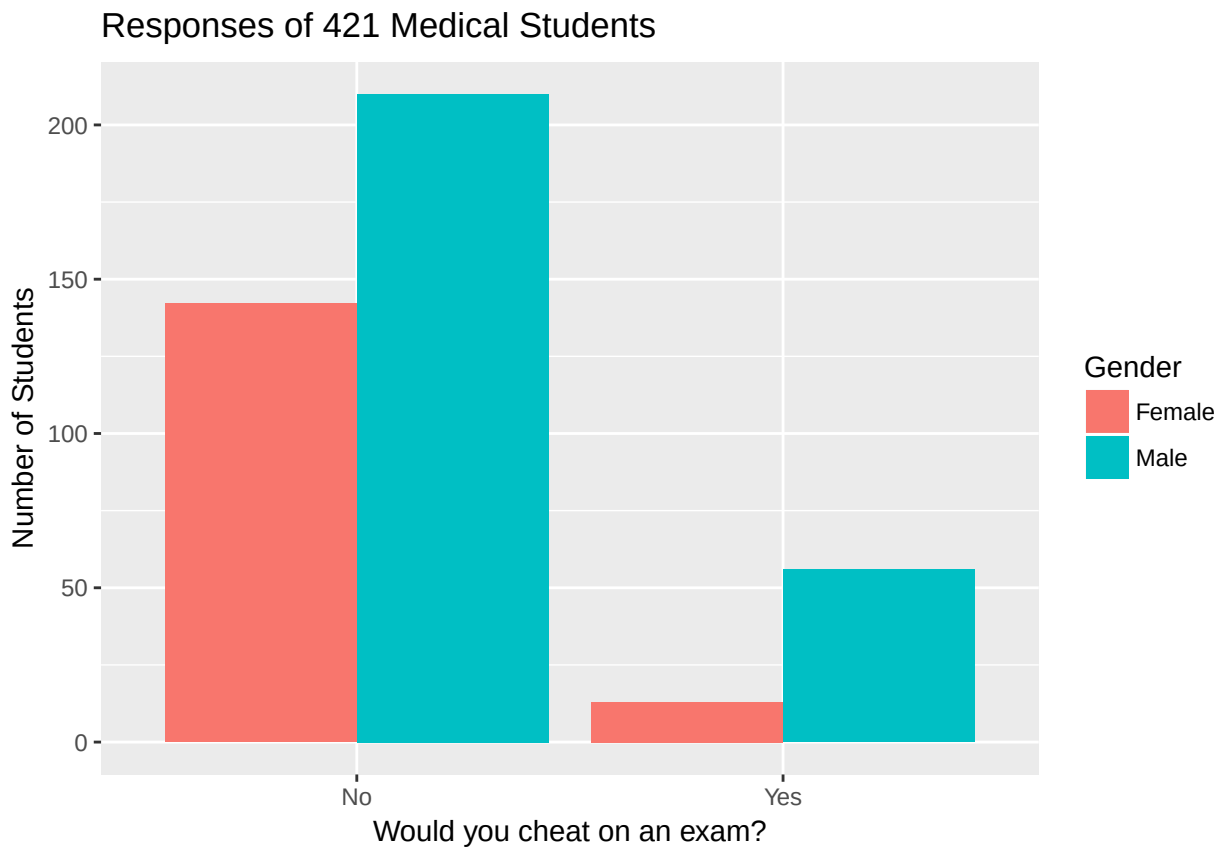
1 pt for checking the sample size assumptions: our smallest count is 13, so we have at least 10 successes and 10 failures in each sample, and so a z procedure is appropriate.

Output, such as the following table or bar graph, is nice but not strictly necessary.

```
xtabs(Num_Students ~ Gender + Would_Cheat, data = cheaters)
```

```
##         Would_Cheat
## Gender    No Yes
##   Female 142  13
##   Male   210  56
```

```
ggplot(cheaters, aes(x = Would_Cheat, y = Num_Students, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Would you cheat on an exam?", y = "Number of Students",
       title = "Responses of 421 Medical Students")
```



C) 1 pt for, at minimum, the following output or similar:

```
prop_matrix <- matrix(c(13, 142, 56, 210), nrow = 2, byrow = TRUE)
prop.test(prop_matrix, conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  prop_matrix
## X-squared = 10.559, df = 1, p-value = 0.001156
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.19736993 -0.05594077
## sample estimates:
##     prop 1     prop 2
## 0.08387097 0.21052632
```

D) 1 pt for either of the following 95% CIs, or equivalent with a different confidence level:

(-0.197, -0.056)

(0.056, 0.197)

E) 1.5 pts for the following, or similar interpretation:

We are 95% confident that male medical students are between 5.6 and 19.7 percentage points more likely than female medical students to consider cheating.

Problem 4

The file milk.csv contains the concentration of trace elements in human milk fed to a sample of infants in three different countries (Argentina, Poland, USA). All concentrations are in $\mu g/L$. All data from the USA were collected from a random sample of Boston-area mothers.

Suppose we believe that the mean arsenic (As) concentration in breast milk from Boston mothers is 3 $\mu g/L$, but we are concerned that it is increasing. Is a sample of 20 mothers (the sample size in the file) sufficient to detect an alternative mean concentration of 3.5 $\mu g/L$?

```
milk <- read_csv("milk.csv")
head(milk)
```

```
## # A tibble: 6 × 9
##   Country Infant_Sex        Zn       Ca       Fe        Cu        Pb
##     <chr>      <chr>     <dbl>    <dbl>    <dbl>     <dbl>     <dbl>
## 1     USA       male  398.6915 241951.2 1137.468 126.55923 0.5750046
## 2     USA     female  601.8943 242269.6 1142.999 116.81579 0.4795497
## 3     USA     female  858.5115 357404.6 1644.756 219.90390 0.4143093
## 4     USA     female 1614.4640 295684.8 1506.118 234.92810 1.0090874
## 5     USA       male  265.1879 293633.6 1385.782  95.17329 0.4309240
## 6     USA       male  426.0921 253291.6 1204.151 162.97591 1.1237631
## # ... with 2 more variables: Mn <dbl>, As <dbl>
```

A) 1 pt for a power analysis in the one mean t framework.

B) 1.5 pts for any reasonable explanation for why t power analysis is appropriate. It should include, at minimum, a description that we are in the process of study design and want to check whether our sample size is large enough (has enough power to detect our desired difference).

C) 1 pt for, at minimum, the following code and output. Note that we need to subset the data to get an appropriate estimate of the standard deviation:

```
library(dplyr)
milk %>% group_by(Country) %>% summarize(count = n(), sd = sd(As))
```

```
## # A tibble: 3 × 3
##     Country count        sd
##       <chr> <int>     <dbl>
## 1 Argentina    21 1.3435142
## 2    Poland    23 1.0038865
## 3       USA    20 0.8425908
```

```
power.t.test(n = 20, delta = 3.5 - 3, sd = 0.8425908,
             type = "one.sample", alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 20
##           delta = 0.5
##              sd = 0.8425908
##       sig.level = 0.05
##           power = 0.8192729
##     alternative = one.sided
```

D) 1 pt for reporting that the power is 0.819 or 81.9%.

E) 1.5 pts for comparing the power to the standard of 0.8 and finding that, indeed, a sample of 20 subjects is sufficiently large to detect the alternative mean of 3.5 $\mu g/L$.

Problem 5

Gastroenteritis is a viral or bacterial infection that spreads through contaminated food and water. Suppose that inspectors wish to determine if the proportion of public swimming pools nationwide that fail to meet disinfectant standards is different from 10.7%, which was the proportion of pools that failed the last time a comprehensive study was done, 2008.

A simple random sample of 30 public swimming pools was obtained nationwide. Tests conducted on these pools revealed that 26 of the 30 pools had the required pool disinfectant levels. Perform an appropriate statistical procedure to help the inspectors.

A) 1 pt for a one-proportion confidence interval or hypothesis test.

B) 0.5 pts for checking BINS assumptions.

1 pt for checking sample size assumptions. Define one of the following:

Success = having required levels, then we have 26 successes and 4 failures

Success = not having required levels, then we have 4 successes and 26 failures

Or, under the null hypothesis, we expect $(0.107)(30) = 3.21$ pools to fail to meet disinfectant standards and $(0.893)(30) = 26.79$ pools to meet disinfectant standards

We do not meet the sample size assumption for either a CI or HT and must use binomial exact procedures.

C) 1 pt for, at minimum, one of the following code and output pairs:

```
binom.test(x = 4, n = 30, p = 0.107)
```

```
##
##  Exact binomial test
##
## data:  4 and 30
## number of successes = 4, number of trials = 30, p-value = 0.5563
## alternative hypothesis: true probability of success is not equal to 0.107
## 95 percent confidence interval:
##  0.0375535 0.3072184
## sample estimates:
## probability of success
##              0.1333333
```

```
binom.test(x = 26, n = 30, p = 0.893)
```

```
##
##  Exact binomial test
##
## data:  26 and 30
## number of successes = 26, number of trials = 30, p-value = 0.5563
## alternative hypothesis: true probability of success is not equal to 0.893
## 95 percent confidence interval:
##  0.6927816 0.9624465
## sample estimates:
## probability of success
##              0.8666667
```

D) 1 pt for any of the following, or equivalent:

Our 95% confidence interval is (0.693, 0.962) if success = met disinfectant levels, and (0.038, 0.307) is success = failed to meet disinfectant levels

Our hypothesis test gives a p-value of 0.556 as long as the correct alternative hypothesis is given ($p \neq 0.893$ if success = met disinfectant levels, and $p \neq 0.107$ if success = failed to meet disinfectant levels)

E) 1.5 pts for either of the following interpretations, or similar:

We are 95% confident that between 3.8% and 30.7% of pools fail to meet required disinfectant levels (or between 69.3% and 96.2% of pools meet disinfectant levels). Since this interval includes 10.7% (or 89.3%), we do not believe there is a significant difference from the last time the study was done.

Since our p-value is greater than our significance level (0.05 or any other reasonable number), we fail to reject our null hypothesis. The proportion of pools that fail to meet required disinfectant levels is not significantly different from 10.7%.