Mid-term Exam #2

Due date: 10 November

Please read all of the questions carefully and follow all instructions. Each question has an allotted point value. Be as thorough and specific as possible; extra calculations and incorrect information, even in the presence of correct information, will result in point deductions. Be sure to show all formulas and all necessary work to answer the questions. You may upload your completed exam to the Elms course site (attach to Midterm Exam #2).

> **Note**
>
> While this is an open-note exam, you are not to receive assistance from anyone else (as usual, the Honor Code applies).

Total points: 50 points

Conceptual questions

> **Note**
>
> Please include all work (and computations) necessary to answer the questions.

```
{.callout-note}
Please note that the dataset gss and nes are downloaded from the ELMS site and will not be au
```

Total points: 26

Question 1

2 points

Suppose I am interested in determining if freshman undergraduates at the University of Maryland spend more hours studying in the average week than sophomore undergraduates. I conduct a study in which I take a simple random sample (SRS) of 1200 freshman students and 1200 sophomore students. I find that the freshman students in my sample study for, on average, 412 minutes per week and the sophomore students in my sample study for, on average, 335 minutes per week. The standard error of the difference is 30. What is a 90% confidence interval for the difference between freshman and sophomore students?

Ans 1:

To construct a confidence interval for the difference between the mean study times of freshman and sophomore students at the University of Maryland, we can use the formula for the confidence interval of the difference between two means.

Since the standard error of the difference is provided, the formula simplifies the process.

We know that the confidence interval is calculated as:

$$Confidence = (\bar{X_1} - \bar{X_2}) + / - z * SE$$

Where:

- X1 is the mean of freshman students who study = 412
- X2 is the mean of sophomore students who study = 335
- SE or standard error = 30
- For a 90% confidence interval the z score is approx 1.645 with a double tailed test

Plugging these values in and calculating:

$$Confidence = (412 - 335) + / - 1.645 * 30$$

Hence CI is between [27.65, 126.35]

Conclusion:

The 90% confidence interval for the difference in the average weekly study time between freshman and sophomore students is between 27.65 minutes and 126.35 minutes. This means that we can be 90% confident that freshman students study between 27.65 and 126.35 minutes more per week on average than sophomore students at the University of Maryland.

Question 2

2 points

Based on the results of my study described in question 1, can you reject the null hypothesis of no statistically meaningful difference in the study habits of sophomore and freshman students? Why or why not?

Ans 2:

For this, let us assume:

The null hypothesis: No statistically meaningful difference in study habits between freshman and sophomore students at the University of Maryland.

Alternative Hypothesis: There is a statistical difference between the study habits

We can run a simple t test to test if we can reject the null,

We know that the t value can be defined as:

$$t = (\bar{X}_1 - \bar{X}_2)/SE$$

Plugging the values in we get,

$$t = (412 - 335)/30$$

Which means t is approx 2.57. Given that the t value is pretty high, it is likely that the p_value at a 90% interval will be not statistically significant. Hence we can reject the null and assume that there is significant difference between the study times.

As a simple alternative method, we know the 90% confidence measure ranges from [27.65, 126.35]. As the confidence interval does not contain zero, we can assume that there is a significant difference in the average study time without a separate statistical test.

Either way we have proved that there is significant statistical difference in values.

Question 3

2 points

If I am testing a null hypothesis that X has no effect on Y in the population (and thus my alternative hypothesis is that X does have an effect), why might I prefer to commit a Type-II error over a Type-I error (and, of course, this holds aside my first preference of making no error at all)? Answer in no more than two sentences.

Ans 3:

Preferring to commit a Type-II error (failing to reject a false null hypothesis) over a Type-I error (rejecting a true null hypothesis) typically reflects a situation where the consequences of falsely detecting an effect (Type-I) are considered more serious or costly than missing a potential effect (Type-II). This preference often arises in contexts where false positives can lead to unnecessary actions which are deemed more detrimental than the missed opportunity of identifying a true effect.

As an example, if we running a software method to identify individuals who might be part of a DDOS attack , it is better to miss a few individuals rather than degrade the user experience of the vast majority of customers.

Question 4

1 point

When conducting a difference-of-means test, which of the following samples would yield a sampling distribution with the lowest variability?

A. A sample of 900 with a standard deviation of 15
B. A sample of 25 with a standard deviation of 10
C. A sample of 625 with a standard deviation of 20
D. A sample of 100 with a standard deviation of 6

Ans4:

We know that the formula for the standard error (SE) of the mean is:

$$SEM = \sigma/\sqrt{n}$$

Where:

- Sigma is the standard deviation of the sample.
- n is the sample size.

Let's calculate the standard error for each option:

Sample of 900 with a standard deviation of 15:

$$SEM900 = 15/\sqrt{900}$$

Hence SEM900 = 15/30 = 0.5

Sample of 25 with a standard deviation of 10:

$$SEM25 = 10/\sqrt{25}$$

Hence SEM25 = 10/5 = 2

Sample of 625 with a standard deviation of 20:

$$SEM625 = 20/\sqrt{625}$$

Hence SEM25 = 20/25 = 0.8

Sample of 100 with a standard deviation of 6:

$$SEM100 = 6/\sqrt{100}$$

Hence SEM100 = 6/10 = 0.6

Based on these calculations, the sample of 900 with a standard deviation of 15 yields the lowest variability (standard error), as it has the smallest SE value of 0.5.

Question 5

1 point

Which of the following probabilities is not independent?

A. The probability that the roulette wheel will end up on 23 two times in a row.
B. The probability that three successive coin tosses will each turn up heads.
C. The probability that I draw an ace and then a king in a row from a deck of cards (when drawing a two-card hand).
D. The probability that I will get a 6 on three subsequent die rolls.
E. None of the above – all are independent probabilities.

Ans 5:

The probability that is not independent among the options provided is:

"The probability that I draw an ace and then a king in a row from a deck of cards (when drawing a two-card hand)."

This is because the outcome of the first draw affects the probability of the second draw. If you draw an ace first, without replacement, the composition of the deck changes, altering the probability of drawing a king next. In contrast, the outcomes of the other events listed (roulette wheel, coin tosses, die rolls) do not affect the probabilities of their subsequent trials, as each trial is independent of the previous ones.

Question 6

1 point

I conduct a two-tailed difference-of-means test and obtain a t-statistic of 2.10. Is my result statistically significant (with 95% confidence)?

A. Yes, at the 0.05 level ($p<0.05$)
B. Yes, but only at the 0.10 level ($p<0.10$)
C. Not at either the 0.05 or the 0.10 level
D. There is not enough information to answer this question.

Ans 6:

To determine whether your result is statistically significant at the 95% confidence level (which corresponds to a significance level of 0.05), we need to compare our t-statistic to the critical value from the t-distribution for our specific degrees of freedom at the two-tailed 0.05 level. However, since the degrees of freedom (which depend on the sample sizes) are not provided, it is not possible to definitively say whether the result is statistically significant at the 0.05 level based solely on the t-statistic of 2.10.

That said, for large enough sample sizes (which lead to a high degree of freedom), a t-statistic of 2.10 would typically be considered significant at the 0.05 level in a two-tailed test, as the critical values tend to be close to those of the standard normal distribution (around 1.96 for two-tailed tests at the 0.05 level). But without information on the degrees of freedom, this remains an assumption.

So option D: Not enough information to answer this question

Question 7

1 point

I take a sample of 1800 adults and find that 360 of them watched last Monday's NFL game. What probability represents the complement to the sample proportion of adults who watched the NFL game?

Ans 7:

The probability representing the complement to the sample proportion of adults who watched the NFL game is 0.8 or 80%. This means that 80% of the adults in the sample did not watch the game.

Question 8

1 point

Which of the following makes it more likely that a given sample statistic will be statistically different from zero (and thus allow you to reject the null hypothesis, all else equal)?

A. More observations
B. Greater variance in the sample
C. Using a two-tailed instead of a one-tailed significance test
D. A larger confidence interval around the test statistic
E. Both (a) and (b) (but not (c) or (d))
F. All of the above increase the chances of statistical significance
G. None of the above

Ans 8:

Null hypothesis rejection can be influenced by several factors. Let's check each one:

1. **More observations**: Increasing the sample size generally increases the likelihood of finding a statistically significant result, if there is an effect to be detected. A larger sample size reduces the standard error of the estimate, making it easier to detect a significant difference from zero.
2. **Greater variance in the sample**: Greater variance within the sample makes it harder, not easier, to find a statistically significant result. Increased variance typically increases the standard error, which can make the test statistic closer to zero and less likely to be considered statistically significant.

3. **Using a two-tailed instead of a one-tailed significance test**: Using a two-tailed test instead of a one-tailed test actually makes it harder to find significance, not easier, because the critical region for rejecting the null hypothesis is split between two tails of the distribution rather than concentrated in one.
4. **A larger confidence interval around the test statistic**: A larger confidence interval generally indicates greater uncertainty about the estimate, which does not necessarily make it more likely that a given sample statistic will be statistically different from zero.

So, the correct answer is **(a) More observations**, as this is the only option that increases the likelihood of rejecting the null hypothesis, assuming there is a true effect to be detected. The other options either make it less likely or do not directly affect the likelihood of statistical significance.

These are of course generalization and not hard and fast rules , but usually these rules will hold.

Question 9

1 point

The p-value for a two-tailed test of the null hypothesis $H_0 : \mu = 40$ is 0.06. Which of the following statements is accurate?

A. A 95% confidence interval around the sample mean includes the value 40
B. A 90% confidence interval around the sample mean includes the value 40
C. A 99% confidence interval around the sample mean does not include the value 40
D. None of the above statements are correct
E. All of the above statements are correct

Ans 9:

Given a p-value of 0.06 in a two-tailed test:

1. **A 95% confidence interval around the sample mean includes the value 40**: A 95% confidence level corresponds to a significance of 0.05. Since the p-value is 0.06, which is greater than 0.05, we cannot reject the null hypothesis at the 95% confidence level. This implies that a 95% confidence interval would include the null hypothesis value (40 in this case).
2. **A 90% confidence interval around the sample mean includes the value 40**: A 90% confidence interval corresponds to an alpha level of 0.10. Since the p-value of 0.06 is less than 0.10, this suggests that the null hypothesis value (40) might not be included in a 90% confidence interval, indicating potential rejection of the null hypothesis at this confidence level.

3. **A 99% confidence interval around the sample mean does not include the value 40**: A 99% confidence interval corresponds to an alpha level of 0.01. Since the p-value of 0.06 is much greater than 0.01, the null hypothesis value (40) would likely be included in a 99% confidence interval.

Based on these interpretations:

The most accurate choice would be:

Option A: A 95% confidence interval around the sample mean includes the value 40

Question 10

4 points total

Part A

2 points

If you roll a fair (six-sided) die twice, what is the probability that both rolls will be odd or greater than four?

Ans 10 A:

We can consider the two parts of the event separately and then combine them. The events are:

1. Both rolls are odd.
2. Both rolls are greater than four.

1. Both Rolls are Odd

The odd numbers on a die are {1,3,5} for both dice. Therefore, the probability of rolling an odd number on one roll is $3/6 = 1/2$. Since the rolls are independent, the probability that both rolls are odd is $1/2 * 1/2 = 1/4$.

2. Both Rolls are Greater than Four

The numbers greater than four on a die are {5,6} for both. The probability of rolling a number greater than four on one roll is $1/3$. Thus, the probability that both rolls are greater than four is $1/3 * 1/3 = 1/9$

Combining the Events

However, these two events are not mutually exclusive, as rolling a 5 falls into both categories. We must adjust for this overlap.

The probability of rolling a 5 twice is 1/6 * 1/6 = 1/36

To find the combined probability of both events happening, we add the probabilities of each event and subtract the probability of their intersection (rolling a 5 twice):

= 1/4 + 1/9 - 1/36

= 1/3

So, the probability that both rolls will be either odd or greater than four is 1/3

Part B

2 points

The following is a distribution of U.S. college students classified by their age and full- vs. part-time status. Based on these data, what is the probability that a student is in the 25-29 age group, given that (i.e., conditional on knowledge that) the student is full time?

```
tibble::tibble(
  age = c("15 - 19", "20 - 24", "25 - 29", "30+"),
  full_time = c(155, 255, 75, 35),
  part_time = c(20, 55, 80, 95)
) |>
  knitr::kable()
```

age	full_time	part_time
15 - 19	155	20
20 - 24	255	55
25 - 29	75	80
30+	35	95

Ans 10 B:

Determine the Total Number of Full-Time Students: To find the probability of a specific event among full-time students, we first need to know the total number of full-time students.

This is the sum of full-time students across all age groups.

Full-Time Students (15 - 19): 155

Full-Time Students (20 - 24): 255

Full-Time Students (25 - 29): 75

Full-Time Students (30+): 35

Total Full-Time Students = 155 + 255 + 75 + 35 = 520

The probability in this case will be a simple division of 75/520 = 0.144 which is about a 14.4% chance.

Alternative:

We can also do this using conditional probability formula:

P(A|B) = P(B(A)/ P(B)

Where A = event that the student is in the 25-29 age group

Where B = event that the student is full time

Hence P(B(A) = 75 from the table

P(B) = Full time student = 155 + 255 + 75 + 35 = 520

Hence the final probability will be 75/520 which is approx 14.4%

Question 11

4 points total

Part A

2 points

Using a SRS of 1211 people, I estimate that the proportion of people living in the South that favor teaching sexual education in public schools is 0.874 and the proportion of people outside of the south that favor teaching sexual education in public schools is 0.906. And, the standard error of the difference (in citizen views about teaching sexual education in public schools) between people living in the south and those not living in the south is 0.015. Give an interval estimate for the difference in the proportion of people favoring sex education in public schools between people who do, and do not, live in the south.

Ans 11 A:

For this we need to compare the proportions of certain populations:

We know that we can write this as:

$$Confidence = (\bar{p_2} - \bar{p_1}) + / - z * SE$$

Defining some values as:

Proportion of sample 1 p1 = 0.874

Proportion of sample 2 p2 = 0.906

Standard Error SE = 0.015

Z at 95% = approx 1.96

Calculating the confidence intervals:

$$Confidence = (0.906 - 0.874) + / - 1.96 * 0.015$$

Which means the intervals are around: [0.0026, 0.0614]

Therefore, the 95% confidence interval for the difference in the proportion of people favoring sex education in public schools between those living in the South and those not is approximately 0.0026 to 0.0614.

Part B

2 points

Do the data (i.e., estimates above) show support for my hypothesis? How do you know?

Ans 11 B:

Given that the hypothesis is that:

The estimate that the proportion of people living in the South that favor teaching sexual education in public schools is 0.874 and the proportion of people outside of the south that favor teaching sexual education in public schools is 0.906.

We can see that the interval values calculated above do not contain zero. This is a strong indicator that the proportion is statistically significant. Additionally, the direction of the interval is positive not negative, giving greater weight to that the fact that ratio is higher outside of the South.

Based on these two statements, I believe that there is enough data to support the hypothesis without moving to other statistical tests.

Question 12

6 points total

I am interested in estimating the average number of texts that University of Maryland undergraduate students send in a day. My hypothesis is that the average number of texts students send is greater than 100. I take a SRS of 1600 students and find that the mean number of texts they send is 105, and with a standard deviation of 120.

Part A

2 points

What is a 95% confidence interval around the sample statistic?

Ans 12 A:

To calculate the 95% confidence bounds around the mean, we know that confidence interval can be written as:

$$Confidence = \bar{x} + / - t * s/\sqrt{n}$$

Where we know that:

x is the sample mean = 105

Standard Deviation: SD = 120

Sample size: n = 1600

t = 1.96 at a 95% confidence interval

Plugging these values in,

$$Confidence = 105 + / - 1.96 * 120/\sqrt{1600}$$

Hence the confidence is around [99.12, 110.88] for a 95% confidence interval around the mean.

Part B

2 points

When testing the null hypothesis, what is the test statistic associated with the sample statistic?

Ans 12 B:

The test statistic associated with the sample statistic when testing the null hypothesis in this scenario is the t-statistic. The t-statistic is used in a one-sample t-test to determine whether there is a statistically significant difference between the sample mean and a known population mean under the null hypothesis.

In this case we can written the t statistic as:

$$t = (\bar{x} - \mu_0)/s/\sqrt{n}$$

Where we know that:

x is the sample mean = 105

mu = 100

s = 120

n = 1600

Plugging the values in we get,

$$t = (105 - 100)/120/\sqrt{1600}$$

Which is approx 1.67. This value is used to asses how strong the relationship is against the null hypothesis.

Although we will note that this is specific for this one case. Test statistics in general can be whatever metric is chosen. We can do t tests, z tests, two sample t tests, chi square tests, anovas etc depending on the type of data is being diagnosed.

Part C

2 points

If using a one-tailed test of the null hypothesis and you are willing to accept a Type-I error rate of 0.05, do the data support my hypothesis? Why or why not?

Ans 12 C:

For this let us assume that the hypothesis is:

Null Hypothesis: The true mean number of texts sent by students is 100 per day.

Alternative Hypothesis: The true mean number of texts sent by students is greater than 100 per day.

Let the error rate (Type -1) = 0.05

Our calculated t statistic is approx. 1.67

For fairly large sample sizes as we have with many degrees of freedom(>1000), the critical t statistic for the sample is 1.646. As our calculated t statistic is slightly higher, we can reject the null hypothesis in favor of the alternative hypothesis.

However this is not a very large difference and hence can do with some more investigation.

Applied questions

> **Note**
>
> All data sets referenced below are available through the course elms page. Do not use the `poliscidata` package for these questions, as the specific variables referenced are not all included in that package.

Ans:

Reading in the GSS and the NES data:

```
gss_data <- read.csv("~/Documents/GitHub/GVPT622_Problems/gss.csv")
nes_data <- read.csv("~/Documents/GitHub/GVPT622_Problems/nes.csv")
```

Total points: 24

Question 1

8 points total

I hypothesize that, among only those that were eligible to vote, people with greater confidence in the U.S. military were more likely to turnout to vote in the 2012 presidential election. Use data from the General Social Survey (i.e., the `gss` dataset) to test my hypothesis (the unit of analysis is the individual survey respondent). Specifically, use the following variables: `conarmy` (1 = a "great deal" of confidence; 2 = "only some" confidence; 3 = "hardly any"); and `vote12` (1 = voted; 2 = did not vote; 3 = ineligible). Answer the following questions.

Part A

3 points

Complete a cross-tab and interpret the results. Do the data support my hypothesis? Be sure to explain the nature of the relationship (or lack thereof, if relevant).

Ans 1 A:

```
# Crosstabulation visualization
library(modelsummary)
library(data.table)

#Reading the data in again to test
gss_data <- read.csv("~/Documents/GitHub/GVPT622_Problems/gss.csv")

#Doing some basic data cleaning
gss_data <- as.data.table(gss_data)

gss_data$conarmy<- as.numeric(gss_data$conarmy)
gss_data$vote12<- as.numeric(gss_data$vote12)

gss_data <- gss_data[complete.cases(gss_data[, .(conarmy, vote12)])]
gss_data <- gss_data[!is.na(gss_data$vote12) & !is.na(gss_data$conarmy)]

#Removing the vote12 = 3 population
gss_data <- gss_data[vote12 != 3]

datasummary_crosstab(conarmy ~ vote12, data = gss_data)
```

conarmy		1	2	All
1	N	642	278	920
	% row	69.8	30.2	100.0
2	N	498	208	706
	% row	70.5	29.5	100.0
3	N	75	51	126
	% row	59.5	40.5	100.0
All	N	1215	537	1752
	% row	69.3	30.7	100.0

```
#Running a statisical test:

# Creating a table
tab <- table(gss_data$conarmy, gss_data$vote12)

# Making a chi square test
chi_test <- chisq.test(tab)

# View the results
print(chi_test)
```

```
    Pearson's Chi-squared test

data:  tab
X-squared = 6.2734, df = 2, p-value = 0.04343
```

```
#tab
```

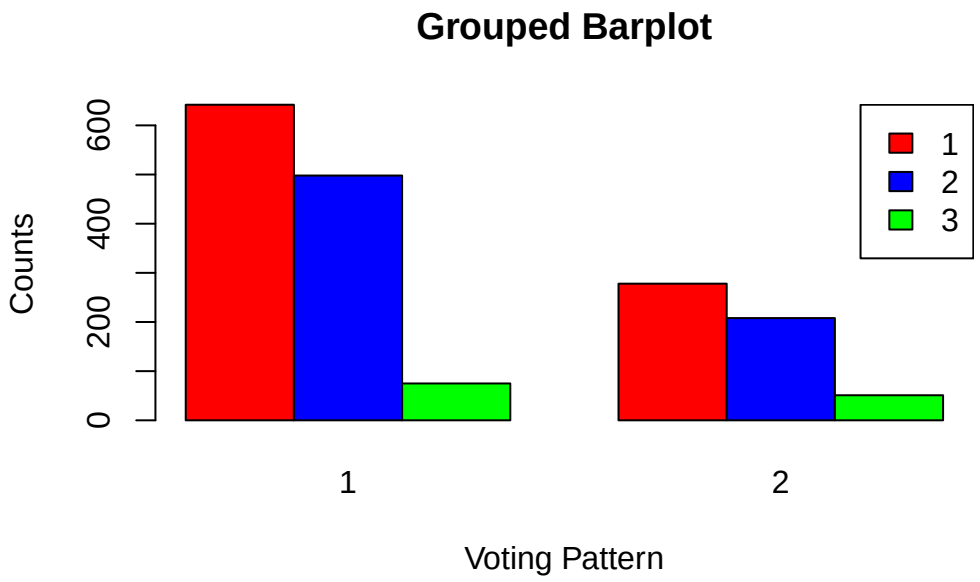
Based on the chi square test

- he Chi-square statistic suggests a relationship between confidence in the military and voting behavior, but the p-value of 0.043 indicates that null hypothesis can be rejected which indicates that there is a slight relationship between voting and confidence in the military.
- The cross tab is not a great way to see the relationship. There seems to be a slight increase in those who trust the military and those who vote. This is verified by the chi square test.

– This doesn't say anything about the direction of the relationship though, it only talks about the nature of the relationship that is present.

Let's use a bar chart to check the relationship visually:

```
barplot(tab, main="Grouped Barplot", xlab="Voting Pattern", ylab="Counts", beside=TRUE, co
legend("topright", legend=rownames(tab), fill=c("red", "blue", "green"))
```



Based on the graph there doesn't seem to be many conclusions drawn, this means that we need to trust the results of the statistical test and our visual inspection is not fine enough to make these differentiations.

Part B

3 points

Compute (by hand) the chi-square statistic to test the null hypothesis of no relationship between these two variables. Be sure to show your work.

Ans 1 B:

We know that the chi square test can be written as:

$$\chi^2 = \sum (O - E)^2 / E$$

Here O is the observed frequency and E is the expected frequency for all values

Expected Freqency = Row_Total * Column Total/Grand Total

Hence the values are:

- **For Group 1**
  - Observed in Category 1: 642
  - Expected in Category 1: 638.01
  - Observed in Category 2: 278
  - Expected in Category 2: 281.99
- **For Group 2**
  - Observed in Category 1: 498
  - Expected in Category 1: 489.61
  - Observed in Category 2: 208
  - Expected in Category 2: 216.39
- **For Group 3**
  - Observed in Category 1: 75
  - Expected in Category 1: 87.38
  - Observed in Category 2: 51
  - Expected in Category 2: 38.62

Hence calculating the final Chi Square value = 6.273 totally by plugging it into the above formula.

The degrees of Freedom for a chi square test is (R-1) * (C-1) = (3-1) * (2-1) = 2, where R and C are rows and columns respectively.

Hence taking the degrees of freedom into account, we can obtain a p_value of approx 0.043.

Thus based on the calculated p_value, we reject the null hypothesis and assume that there is a relationship between these two variables.

Part C

2 points

Using the chi-square statistic that you computed in question 1(b), can you reject the null hypothesis of no relationship between these two variables with 95% confidence? Why, or why not?

Ans 1 C:

Yes we can reject the null hypothesis for the above test due to the p_value being less than 95% confidence interval threshold. The Chi square value itself is interestingly high though.

Question 2

8 points total

I hypothesize that citizens who do not support fracking are less conservative than those who do support fracking. Use data from the `nes` dataset to test my hypothesis (the unit of analysis is the individual survey respondent). Specifically, use the following variables: `fracking` (1 = "approve" of fracking; 2 = "middle"; 3 = "disapprove"); and `libcon7` (higher values represent less liberalism, or more conservatism). Answer the following questions.

Part A

2 points

Using these data, what is the point estimate for the mean conservatism/liberalism score among those that disapprove of fracking? What is the point estimate for the mean conservatism/liberalism score among those that approve of fracking?

Ans 2 A:

Based on the data :

- The point estimate for the mean conservatism/liberalism score among those that disapprove of fracking (fracking = 3) is 3.5.
- The point estimate for the mean conservatism/liberalism score among those that approve of fracking (fracking = 1) is 5.10

These point estimates represent the average scores for liberalism/conservatism (with higher scores indicating more conservatism) within each group.

Part B

6 points

Evaluate the null hypothesis that there is no difference in the mean conservatism/liberalism score among those that approve vs. disapprove of fracking. Do the data support my hypothesis? Why or why not? Be sure to show all work necessary to answer the question by hand (i.e., you may only use R to the extent that is absolutely necessary to complete the question; otherwise, you must show how you would answer the question by hand).

Ans 2 B:

To do this we can use a two sample t test:

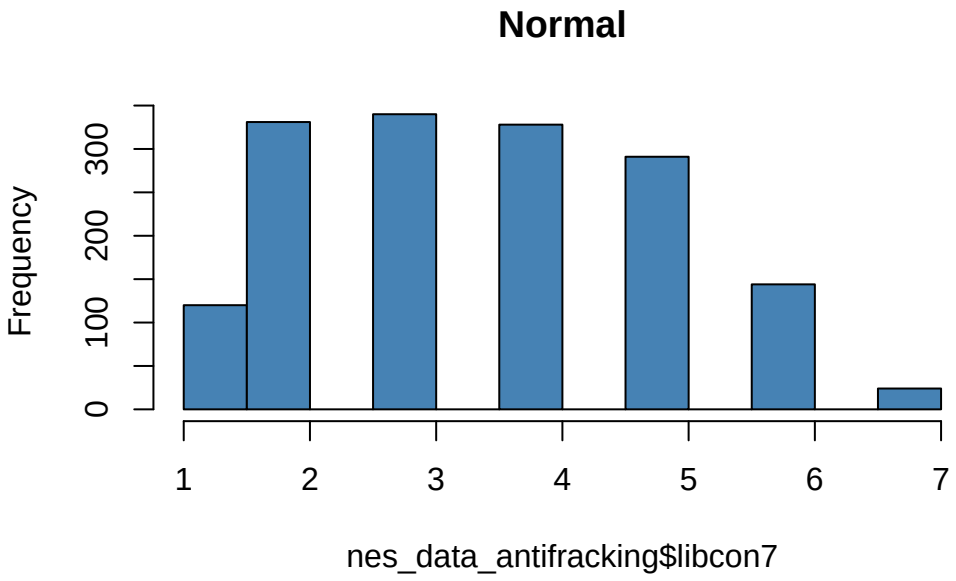
Defining the hypothesis:

Null Hypothesis: There is no difference in the mean conservatism/liberalism score among those that approve vs. disapprove of fracking. Or in this case the mean libcon7 score is the same for those who support fracking as those who do not support fracking.

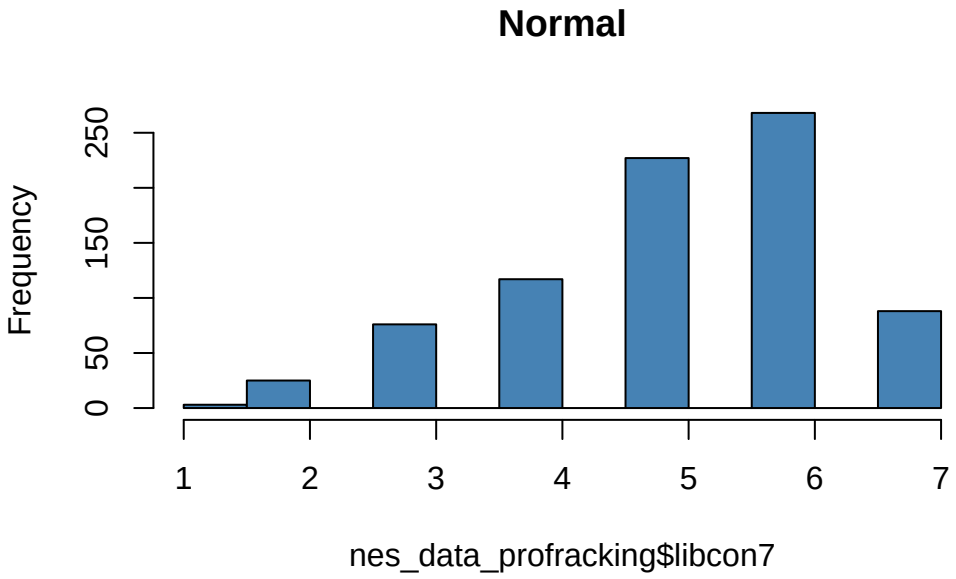
Alternative Hypothesis: The mean libcon7 score for both groups is not going to be equal.

```
#Cleaning the data, removing NA values and splitting it into wo groups
library(data.table)
nes_data <- read.csv("~/Documents/GitHub/GVPT622_Problems/nes.csv")
nes_data <- as.data.table(nes_data)

nes_data$fracking<- as.numeric(nes_data$fracking)
nes_data$libcon7<- as.numeric(nes_data$libcon7)

nes_data <- nes_data[complete.cases(nes_data[, .(fracking, libcon7)]), ]
nes_data <- nes_data[!is.na(nes_data$libcon7) & !is.na(nes_data$fracking)]

#Split into profracking and anti fracking datasets
nes_data_profracking <- nes_data[fracking == 1]
nes_data_antifracking <- nes_data[fracking == 3]
```

```
#Quick check for normaility
hist(nes_data_antifracking$libcon7, col='steelblue', main='Normal')
```




```r
hist(nes_data_profracking$libcon7, col='steelblue', main='Normal')
```



Little skewed but good enough for our process.

Now Let's use the t statistic to determine if there is a significant difference in the groups

We know that the t statistic can be written as:

$$t = (\bar{x_1} - \bar{x_2})/\sqrt{((s_1 2/n_1 + s_2 2/n_2)}$$

Where

x1 = mean of population 1 = 5.11

x2 = mean of population 2 = 3.55

n1 = sample size of group 1 = 1578

n2 = sample size of group 2 = 804

s1 = the sample variance of group 1 = 1.52

s2 = The sample variance of group 2 = 1.22

Now plugging the values in,

$$t = (5.11 - 3.55)/\sqrt{((1.62/804 + 2.22/1578)}$$

Which is approx: t = 26.64.

The p_value for this t statistic at a 95% confidence interval is approx 4.6 * 10^-135. With such a very small p value, this indicates that there is a high significance that these two groups are very different.

In conclusion, we reject the null hypothesis and assume that there is significant difference between these two groups as supported by the data.

Alternative Approach:

In this case we actually cannot use the T test as the assumption of normality is disregarded. I would have switched to a Mann-Whitney U test using ranking.

Question 3

8 points total

I hypothesize that people who express that religion is important to them were more likely to turnout to vote in the 2016 presidential election. Use data from the `nes` dataset to test my hypothesis (the unit of analysis is the individual survey respondent). Specifically, use the following variables: `Relig_imp` (0 = not important; 1 = somewhat important; 2 = quite a bit; 3 = a great deal); and `Voted_2016` (0 = did not vote; 1= voted). Answer the following questions.

Part A

2 points

Using these data, what is the point estimate for the proportion of respondents that voted (i.e., turnout rate), among citizens expressing that religion is not important? What is the point estimate for the proportion of respondents that voted, among citizens expressing that religion matters a great deal?

Ans 3 A:

We are creating a hypothesis that there is a relationship between the importance of religion and voting patterns.

Let's first create a summary of the variables to examine the relationship.

```
library(data.table)

nes_data <- read.csv("~/Documents/GitHub/GVPT622_Problems/nes.csv")
nes_data <- as.data.table(nes_data)

nes_data$Voted_2016<- as.numeric(nes_data$Voted_2016)
nes_data$Relig_imp<- as.numeric(nes_data$Relig_imp)

nes_data <- nes_data[complete.cases(nes_data[, .(Relig_imp, Voted_2016)]), ]
nes_data <- nes_data[!is.na(nes_data$Voted_2016) & !is.na(nes_data$Relig_imp)]

datasummary_crosstab(Voted_2016 ~ Relig_imp, data = nes_data)
```

Just looking at the cross tab results, we cannot really make an assumption based upon the percentage values.

Point means:

Voted_2016		0	1	2	3	All
0	N	297	106	141	234	778
	% row	38.2	13.6	18.1	30.1	100.0
1	N	971	367	509	995	2842
	% row	34.2	12.9	17.9	35.0	100.0
All	N	1268	473	650	1229	3620
	% row	35.0	13.1	18.0	34.0	100.0

1. **Proportion of Respondents That Voted Among Citizens Expressing Religion is Not Important (Group 0)**:
    - Total in Group 0 (Religion least important): 297+ 971 = 1268 respondents
    - Of these, the number who voted: 971 (from category '1').
    - The point estimate for the turnout rate is calculated as the proportion of voters to the total in Group 0: 971/1268
2. **Proportion of Respondents That Voted Among Citizens Expressing Religion Matters a Great Deal (Group 3)**:
    - Total in Group 3 (Religion most important): 234 + 995 = 1229
    - Of these, the number who voted: 995 (from category '1').
    - The point estimate for the turnout rate is calculated as the proportion of voters to the total in Group 3: 995/1229

Finalizing the proportions:

The point estimates for the turnout rates are as follows:

1. **Among Citizens Expressing Religion is Not Important (Group 0)**: The turnout rate is approximately 76.58%.
2. **Among Citizens Expressing Religion Matters a Great Deal (Group 3)**: The turnout rate is approximately 80.96%.

These percentages reflect the proportion of respondents who voted in each group, based on their views on the importance of religion.

Part B

6 points

Evaluate the null hypothesis that there is no difference in the proportion of voters (i.e., turnout rate) among citizens expressing that religion is not important vs. those reporting that religion matters a great deal. Do the data support my hypothesis? Why or why not? Be sure to show all work necessary to answer the question by hand (i.e., you may only use R to the extent that is absolutely necessary to complete the question; otherwise, you must show how you would answer the question by hand).

Ans 3 B:

1. **Null Hypothesis (H0)**: The proportion of voters among citizens expressing that religion is not important is equal to the proportion of voters among citizens reporting that religion matters a great deal.
2. **Alternative Hypothesis (H1)**: The proportions are not equal.

Testing this hypothesis we will use a two proportion z-test that allows us to see the difference in two population proportions.

We know that the formula is:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(1/n_1 + 1/n_2}}$$

where:

- Here p1 and p2 are the sample proportions in each group
- n1 and n2 are the group sizes
- p can be defined as the pooled sample proportion which can be written as:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Where here x1 and x2 are the counts of votes in each group.

Now noting down values:

1. **p1 (Proportion of voters among citizens expressing religion is not important)**:

   = Sample that votes and no low importance to religion/Total voting population where religion is low= 971/1268 = 76.5%

2. **p2 (Proportion of voters among citizens reporting religion matters a great deal)**:

   = Sample that votes and no low importance to religion/Total voting population = 995/1229 = 80.9%

3. **n1 (Sample size for citizens expressing religion is not important)**: = 1268

4. **n2 (Sample size for citizens reporting religion matters a great deal)**: = 1229

5. **x1 (Count of voters among citizens expressing religion is not important)**: = 971

6. **x2 (Count of voters among citizens reporting religion matters a great deal)**: = 995

7. **p (Pooled sample proportion)**: (971 + 995)/(1268 + 1229) = 78.7%

Hence calculating the final values:

$$z = \frac{0.765 - 0.809}{\sqrt{0.787(1 - 0.78)(1/1268 + 1/1229}}$$

z is approx -2.64. Using a two tailed test, we obtain a p_value of approx 0.008. If we used an alpha value of 0.05, we would have rejected the null hypothesis.

Conclusion:

Given the p-value is below 0.05, we reject the null hypothesis. This suggests that there is a statistically significant difference in the turnout rates between citizens expressing that religion is not important and those reporting that religion matters a great deal. So in summary there appears to be statistically significant difference in voting behavior based upon religion.