```
---
title: "Mid-term Exam #2"
subtitle: "Due date: 17 November"
name: "Clare Porter"
format:
  html:
    self-contained: true
toc: true
editor: visual
execute:
  echo: true
  message: false
  warning: false
---
```

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).

::: callout-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

::: callout-note
Please include all work (and computations) necessary to answer the questions.
:::

**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?

The 90% confidence interval for the difference between freshman and sophomore students hours spent studying is (27.65, 126.35).

```{r}
# CI = point estimate +- MOE
# MOE = t*SE = 1.645*30 = 49.35
# SE = sd/sqrt(n) = 30
# CI = (412-335) +- 49.35
# = (27.65, 126.35)
```

### 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?

Based on the results of your study described in question 1, yes, you can reject the null hypothesis of no statistically meaningful difference in the study habits of sophomore and freshman students with 90% confidence. We can confirm this conclusion because our confidence interval captures our sampled difference of means and does not include zero, and by calculating the t-statistic (t=77/30 =\~ 2.567) and finding a one-sided p-value of \~0.0051. Since our p-value (=0.0051) is less than 0.1 we can conclude statistical significance and reject the null with 90% confidence.

### 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.

You might prefer to commit a Type-II error over a Type-I error because failing to detect an effect in the population that truly exists is typically less harmful than falsely concluding that an effect exists when it does not. A type-1 error could lead to acting on a nonexistent relationship, like marketing vapes as products to help quit nicotine, which ultimately created and strengthened nicotine additions.

### 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

**A sample of 900 with a standard deviation of 15** would yield a sampling distribution with the lowest variability.

```{r}
# variance = sd/sqrt(n)
# A: 15/sqrt(900) = 1/2
# B: 10/sqrt(25) = 2
# C: 20/sqrt(625) = 4/5
# D: 6/sqrt(100) = 3/5
```

### 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.

**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)** is not independent because you are drawing cards without replacement.

### 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.**

**There is not enough information to answer this question** because we do not know the degrees of freedom.

### 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?

20% of sampled adults watched the NFL game last Monday. The complement to this probability is 80% of sampled adults 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

Of the options provided above, only **more observations** will make it more likely that a given sample statistic will be statistically different from zero.

### 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

The statement **a 95% confidence interval around the sample mean includes the value 40** is accurate (because 0.06 \> 0.04).

### 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?

There is a \~44.4% probability that both rolls will be odd or greater than four.

```{r}
# P(odd or >4) = 4/6
# (4/6)^2 = 4/9
```

#### 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?

There is a 14.4% probability that a student is in the 25-29 age group, given that the student is full time.

```{r}
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()

# Add up total students who are full-time
library(tibble)
students <- tibble(
  age = c("15 - 19", "20 - 24", "25 - 29", "30+"),
  full_time = c(155, 255, 75, 35),
  part_time = c(20, 55, 80, 95))
sum(students$full_time)

# P(25-29 | full-time) = 75/520 = 0.144
```

### 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.

The 95% confidence interval is (−0.0026, −0.0614). This means we are 95% confident that the proportion of people in the South who favor teaching sexual education in public schools is 0.26% to 6.14% lower than the proportion of people outside the South.

```{r}
# n= 1211, SE = 0.015
# Ha = pro sex education in South is 0.874, and outside South is 0.906. Difference=−0.032
# Let's provide an interval estimate with 95% confidence, meaning a z-statistic = 1.96
# MOE = z*SE = 1.96*0.015 =~ 0.0294
# CI = point estimate +− MOE
# −0.032 +− 0.0294 = (−0.0026, −0.0614)
```

#### Part B

*2 points*

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

Yes, the data shows support for your hypothesis that People in the South are less likely to favor teaching sexual education in publish school than people outside the South. Our 95% confidence interval is negative and does not include zero, which supports the directionality of the hypothesis and that the difference is not zero.

### 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?

We are 95% confident that the true average number of texts students send is between approximately 99 and 111.

```{r}
# CI = point estimate +− MOE
# MOE = t*SE = 1.96*3 = 5.88
# SE = sd/sqrt(n) = 120/sqrt(1600) = 120/40 = 3
# CI = 105 +− 5.88
# = (99.12, 110.88) =~ (99,111)
```

#### Part B

*2 points*

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

A test statistic of \~1.67 is associated with the sample statistic when testing the null hypothesis, which can be compared with a critical value to determine whether to reject the null hypothesis.

```{r}
```

```
# t = (sample mean - hypothesis mean)/SE
# t = (105-100)/3 =~ 1.67
```

#### 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?

A one-tailed test of the null hypothesis with a Type-I error rate of 0.05 is 1.645. Since our test statistic is greater than our critical value (1.67 \>1.645) we can reject the null hypothesis with 95% confidence and conclude that the data supports your hypothesis that the mean number of texts students send is greater than 100.

## Applied questions

::: callout-note
All datasets referenced below are available through the `poliscidata` package.
:::

```{r}
# packages needed
install.packages("caTools")
library(poliscidata)
library(dplyr)
library(modelsummary)
```

**Total points: 24**

### Question 1

```{r}

gss <- poliscidata::gss |>
  mutate(voted08 = as.numeric(voted08),
         conarmy = as.numeric(conarmy))

colnames(gss)
```

*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 2008 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 `vote08` (1 = did not vote; 2 = voted). 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).

Our cross table shows that people with a "great deal" of confidence voted at 68.5%, people with "only some" confidence voted at 67.8%, and people with "hardly any" confidence voted at 55%. This trend in the data suggests a positive association between confidence and voting, supporting your hypothesis. While the cross-table demonstrates a clear direction

in the relationship, a chi-square test would formally confirm whether there is a relationship.

```{r}
gss2 <- gss %>%
  filter(voted08 %in% c(1,2),
         conarmy %in% c(1,2,3)) %>%
  mutate(
    voted = factor(voted08, levels = c(1,2), labels = c("Did not vote", "Voted")),
    conarmy_f = factor(conarmy, levels = c(1,2,3),
                       labels = c("Great deal", "Only some", "Hardly any"))
  )

datasummary_crosstab(conarmy ~ voted08, data=gss)
```

#### 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.

As shown below, we calculate a chi-square statistic of \~6.035.

|                                            |
|--------------------------------------------|
| Expected counts: (row total)*(col total)/N |

| Confidence Level | Did not Vote | Voted | Total |
|---|---|---|---|
| 1 (Great Deal) | (670*316/1200) =176.4 | (670*884/1200) =493.6 | 670 |
| 2 (Only Some) | (435*316/1200) =114.6 | (435*884/1200) =320.4 | 135 |
| 3 (Hardly Any) | (95*316/1200) =25 | (95*884/1200) =70 | 95 |
| **Total** | 11 | 17 | 40 |

|                            |
|----------------------------|
| Chi-Square: (O - E)\^2 / E |

| Confidence Level | Did not Vote | Voted | Total |
|---|---|---|---|
| 1 (Great Deal) | (173-176.4)\^2/176.4 =0.066 | ((497-493.6)\^2/493.6 =0.024 | 0.09 |
| 2 (Only Some) | (108-114.6)\^2/114.6 =0.380 | (327-320.4)\^2/320.4 =0.135 | 0.515 |
| 3 (Hardly Any) | (35-25)\^2/25 | (60-70)\^2/70 | 5.43 |

| | =4.0 | =1.43 | |
|---|---|---|---|
| **Total** | 4.446 | 1.589 | 6.035 |

With X degrees of freedom \[(r-1)(c-1)=2*1 =2\]

Let's check my work:

```{r}
chisq.test(gss2$conarmy_f, gss2$voted)
```

#### 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?

Yes, we can reject the null hypothesis of no relationship between these two variables with 95% confidence and conclude there is a relationship between the confidence and voting (though, again, we cannot conclude the directionality of this relationship). As shown in the code above, our chi-squared test is associated with a p-value of 0.04961, which is below our 0.05 threshold.

### Question 2

```{r}

nes <- poliscidata::nes |>
  dplyr::mutate(envir_drill = factor(envir_drill, levels = c("1. Favor",
                                                             "3. Neither favor nor
oppose",
                                                             "2. Oppose")),
                envir_drill = as.numeric(envir_drill),
                pid_x = as.numeric(pid_x),
                relig_pray = as.numeric(relig_pray),
                voted2008 = as.numeric(voted2008))
```

*8 points total*

I hypothesize that citizens who do not support increased U.S. offshore drilling are more conservative than those who do not. Use data from the `nes` dataset to test my hypothesis (the unit of analysis is the individual survey respondent). Specifically, use the following variables: `envir_drill` (1 = “favor”; 2 = “neither favor nor oppose”; 3 = “oppose”); and `pid_x` (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 oppose drilling? What is the point estimate for the mean conservatism/liberalism score among those that are in favor of drilling?

The conservatism/liberalism score among those that oppose drilling is 2.93; the conservatism/liberalism score among those that favor drilling is 4.18.

```{r}
# Compute mean point estimate for each drilling position
```

```
nes %>%
  filter(!is.na(envir_drill), !is.na(pid_x)) %>%
  group_by(envir_drill) %>%
  summarise(mean_pid = mean(pid_x))
```

#### 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 drilling. 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).

By calculating the two sample t-test, and comparing with a critical value for alpha = 0.05 along a normal distribution since our sample size is large, we see that 20.93 (our two sample t-test) is much larger than our critical value two tailed test of 1.96. We can therefore reject our null hypothesis that there is no difference in the mean conservatism/liberalism score among those that approve vs. disapprove of drilling.

We cannot however, conclude, that the data supports your hypothesis that citizens who do not support increased U.S. offshore drilling are more conservative than those who do not. This is because our data actually shows the opposite (as clearly demonstrated in Part A): citizens who *do* support increased U.S. offshore drilling are *more* conservative than those who do not.

```{r}
# two-sample t-test: t= (mean point estimate1 - mean point estimate2)/SEdiff
# SEdiff = sqrt((sd1^2/n1) + (sd2^2/n1))

## We will use R to compute n, mean, and standard deviation for each group

nes %>%
  filter(!is.na(envir_drill), !is.na(pid_x)) %>%  # remove missing values
  filter(envir_drill %in% c(1,2)) %>%            # only Favor (1) and Oppose (2)
  group_by(envir_drill) %>%
  summarise(
    n = n(),
    mean_pid = mean(pid_x),
    sd_pid = sd(pid_x)
  )


# SEdiff = sqrt((2.186845^2/3069)+(1.814609^2/1648) =~ 0.0596
# t = (4.179537 - 2.931432)/0.0596 =~ 20.93
```

### 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 2008 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_pray`, which describes how often the respondent prays (1 = several times a day; 2 = once a day; 3 = a few times a week; 4 = once a week or less; 5 = never); and `voted2008` (1 = did not vote; 2= 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?

The point estimate for the turnout rate among citizens expressing that religion is not important is 72.4%. The point estimate for the turnout rate among citizens expressing that religion matters a great deal is 82.4%.

```{r}
nes3 <- nes %>%
  filter(!is.na(relig_pray), !is.na(voted2008))

# Compute turnout rate for extremes
nes3 %>%
  filter(relig_pray %in% c(1,5)) %>%
  group_by(relig_pray) %>%
  summarise(
    n = n(),
    voted = sum(voted2008 == 2),
    turnout_rate = voted / n
  )

```

#### 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).

Yes, the data supports your hypothesis that people who express that religion is important to them were more likely to turnout to vote in the 2008 presidential election. In part A, we saw an observable difference in voter turnout between citizens who expressed religion as being important and those who expressed it as being unimportant. By calculating the z test below (z=\~ 6.12 \> z* = 1.96 for a significance at alpha = 0.05), we can conclude that the difference is statistically significant with over 95% confidence.

```{r}
# z statistic = Difference in sampled proportions/SE
# Difference in sampled proportions = 0.8243012 - 0.7239199     = 0.1003813
# SE = sqrt(p(1-p)(1/n1 + 1/n2))
# p is pooled proportion here = (voted1 + voted5)/(n1 + n5)
# = (1445 + 687)/(1753+949) = 0.7896
# SE = sqrt(0.7896(1-0.7896)(1/1753 + 1/949) = sqrt(0.1661*0.001624)
# =~ 0.0164
# z = 0.1003813/0.0164 =~ 6.12
```