Second Midterm Examination

Econ 103, Statistics for Economists

April 2nd, 2019

You will have 70 minutes to complete this exam. Graphing calculators, notes, and textbooks are not permitted.

I pledge that, in taking and preparing for this exam, I have abided by the University of Pennsylvania's Code of Academic Integrity. I am aware that any violations of the code will result in a failing grade for this course.

Name: _____

Signature: _____

Student ID #: _____ Recitation #: _____

Question:	1	2	3	4	5	6	Total
Points:	15	20	30	20	30	25	140
Score:							

Instructions: Answer all questions in the space provided, continuing on the back of the page if you run out of space. Show your work for full credit but be aware that writing down irrelevant information will not gain you points. Be sure to sign the academic integrity statement above and to write your name and student ID number on *each page* in the space provided. Make sure that you have all pages of the exam before starting.

Warning: If you continue writing after we call time, even if this is only to fill in your name, twenty-five points will be deducted from your final score. In addition, a point will be deducted for each page on which you do not write your name and student ID.

1. Assa made a cross-tab classifying all Penn undergraduates by their class standing in 2017–2018 (freshman, sophomore, junior, or senior) and the number of times that they visited the Penn museum during the academic year (never, once, or more than once).

		Museum Visits		
		Never	Once	More than Once
Class	Freshman	0.08	0.10	0.04
	Sophomore	0.04	0.10	0.04
	Junior	0.04	0.20	0.09
	Senior	0.02	0.15	0.10

For example, we see that 8% of Penn undergraduates in 2017–2018 were Freshmen who never visited the Museum, while 10% were seniors who visited more than once.

[5] (a) What fraction of undergraduates visited the museum more than once?

Solution: $P(\text{More than Once}) = 0.04 + 0.04 + 0.09 + 0.10 = 0.27$

[10] (b) Suppose we chose a random Penn undergraduate in 2017–2018. If she is a junior, what is the probability that she never visited the museum?

Solution: Since $P(\text{Junior}) = 0.04 + 0.20 + 0.09 = 0.33$,

$$P(\text{Never}|\text{Junior}) = \frac{P(\text{Never} \cap \text{Junior})}{P(\text{Junior})} = \frac{0.04}{0.33} = \frac{4}{33}$$

2. For each part, write a single line of R code to carry out the required task.

[4] (a) Find c such that $P(-c \leq Z \leq c) = 0.70$ if $Z \sim N(0, 1)$.

Solution: `qnorm(0.85)`

[4] (b) Find the 90th percentile of a Student-t RV with 8 degrees of freedom.

Solution: `qt(0.9, 8)`

[4] (c) Make 15 random draws with replacement from the set of numbers $\{1, 2, 3, 4\}$.

Solution: `sample(c(1, 2, 3, 4), 15, TRUE)`

[4] (d) Calculate $P(Z \geq 1.5)$ if $Z \sim N(0, 1)$.

Name: _____ Student ID #: _____

Solution: `1 - pnorm(1.5)`

4 (e) Calculate $P(-1 \leq X \leq 3)$ if $X \sim N(\mu = 2, \sigma^2 = 9)$.

Solution: `pnorm(1/3) - pnorm(-1)`

3. Write down the output generated by each chunk of R code. If the output is random, e.g. based on simulated random draws, give the approximate value of the output.

3 (a)
```
TRUE == 1
```

Solution: `TRUE`

3 (b)
```
x <- c(1, 5, 10)
y <- c(0, 6, 10)
x > y
```

Solution: `TRUE FALSE FALSE`

3 (c)
```
for(i in 1:3) {
  x <- 2 * i
  print(x)
}
```

Solution: `2 4 6`

3 (d)
```
x <- 5
y <- 7
!((x < 4) & !(y > 12))
```

Solution: `TRUE`

3 (e)
```
number <- 75
if (number < 10) {
  if (number < 5) {
    result <- "extra small"
  } else {
    result <- "small"
  }
} else if (number < 100) {
  result <- "medium"
```

```
} else {
  result <- "large"
}
print(result)
```

Solution: `medium`

5 (f)
```
flips <- rbinom(10000, 10, 0.5)
var(flips)
```

Solution: $\approx 10 \times 0.5 \times (1 - 0.5) = 2.5$

5 (g)
```
mean(rbinom(10000, 1, 0.2) == 0)
```

Solution: $\approx 1 - 0.2 = 0.8$

5 (h)
```
f <- function(x) {
  y <- sample(1:6, 1)
  return(y > x)
}
mean(replicate(10000, f(4)))
```

Solution: $\approx 1/3$

4. In each of the following parts X denotes a continuous RV.

5 (a) Let X have support set $[0, +\infty)$ and CDF $F(x) = 1 - e^{-x}$. Find the pdf of X.

Solution:

$$f(x) = F'(x) = e^{-x}$$

5 (b) Let $X \sim \text{Uniform}(0, 10)$. Calculate $E[X^2]$.

Solution:

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \frac{1}{10} \int_0^{10} x^2\, dx = \frac{1}{10} \left(\frac{x^3}{3} \right) \Bigg|_0^{10} = \frac{100}{3}$$

5 (c) Let X have support set $[-1, 1]$ and pdf $f(x) = (x + 1)/2$. Find the CDF of X.

Solution:

$$F(x_0) = \int_{-\infty}^{x_0} f(x)\,dx = \frac{1}{2}\int_{-1}^{x_0} (x+1)\,dx = \frac{1}{4}\left(x^2+2x\right)\Big|_{-1}^{x_0}$$
$$= \frac{1}{4}\left[\left(x_0^2+2x_0\right) - (1-2)\right] = \frac{1}{4}\left(x_0^2+2x_0+1\right)$$

Hence,

$$F(x_0) = \begin{cases} 0, & x_0 < -1 \\ \frac{1}{4}(x_0^2+2x_0+1), & -1 \le x \le 1 \\ 1, & x_0 > 1 \end{cases}$$

5 (d) Let X have support set $[0, c]$ and pdf $f(x) = \sqrt{x}$. Find c.

Solution:

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_0^c x^{1/2}\,dx = \frac{c^{3/2}}{3/2}$$

Hence $3/2 = c^{3/2}$. Squaring both sides, $9/4 = c^3$ so $c = \sqrt[3]{9/4} \approx 1.31$.

5. Suppose that we have a bag containing three different kinds of candies: "light" candies weigh 5 grams each, "medium" candies weigh 10 grams each, and "heavy" candies weigh 20 grams each. There are 100 candies in the bag: 50 light, 30 medium, and 20 heavy. To estimate the weight of the bag, we take a *random sample* of 5 candies from the bag and calculate their total weight. We then multiply the total weight by 20 to estimate the weight of the bag. Let $X_1, \ldots, X_5$ be the weights of the five randomly sampled candies, and $W = 20 \times (X_1 + \cdots + X_5)$ be our estimator of the weight of the bag.

Note: *Based on our experience from class, you may worry that it would be difficult to obtain a random sample in practice. Heavier candies might sink to the bottom of the bag, for example. This question asks you to disregard such concerns. In other words, assume that we have figured out a way to obtain a truly random sample from the bag of candies.*

6 (a) Calculate $E[X_1]$.

Solution: $E[X_1] = 0.5 \times 5 + 0.3 \times 10 + 0.2 \times 20 = 9.5$ grams.

8 (b) Calculate $Var(X_1)$.

Solution: We have

$$\begin{aligned} E[X_1^2] &= 0.5 \times 5^2 + 0.3 \times 10^2 + 0.2 \times 20^2 \\ &= 0.5 \times 25 + 0.3 \times 100 + 0.2 \times 400 \\ &= 122.5 \end{aligned}$$

Hence, by the shortcut rule $Var(X_1) = 122.5 - (9.5)^2 = 32.25$ grams^2.

8 (c) Is W an unbiased estimator of the weight of the bag? Explain briefly.

Solution: The true weight of the bag is $50 \times 5 + 30 \times 10 + 20 \times 20 = 950$ grams. By the linearity of expectation,

$$\begin{aligned} E[W] &= E[20 \times (X_1 + \cdots + X_5)] \\ &= 20 \times (E[X_1] + \cdots + E[X_5]) \\ &= 20 \times 5 \times 9.5 \\ &= 950 \text{ grams} \end{aligned}$$

Hence, W is an unbiased estimator of the weight of the bag.

8 (d) Calculate the standard error of W.

Solution: The RVs $X_1, X_2, \ldots, X_5$ are iid. Hence,

$$\begin{aligned} Var(W) &= Var\left[20 \times (X_1 + \cdots + X_5)\right] \\ &= 400 \times [Var(X_1) + \cdots + Var(X_5)] \\ &= 400 \times 5 \times 122.5 \\ &= 245000 \text{ grams}^2 \end{aligned}$$

Therefore, $SE(W) = \sqrt{Var(W)} \approx 495$ grams.

6. This question is based on a recent paper examining how "organic" labeling changes people's perceptions of different food products. Researchers recruited volunteers at a local mall in Ithaca, New York and gave each two samples of yogurt to taste. Although both yogurts were in fact identical, the volunteers were *told* that one of them was organic while the other was not. After tasting both, each volunteer was asked to estimate how many calories each of the samples of yogurt contained. (Since, unknown to the volunteer, both samples contained exactly the same kind of yogurt, each in fact contained the same number of calories.) To prevent confounding from anchoring or other behavioral effects,

the order in which a given volunteer tasted the two yogurts, i.e. "organic" first or "organic" second, was chosen at random. The results are stored in the dataframe `yogurt`:

```
> head(yogurt, 3)
  regular organic
1      60      40
2       5       0
3     200     100
```

Each row in this dataframe corresponds to a single individual's guess of the number of calories contained in each of the two yogurts. For example, the values 60 and 40 in row 1 mean that volunteer number one guessed that the regular yogurt sample contained 60 calories and the organic sample contained 40. Summary statistics are as follows:

	`regular`	`organic`
Sample Mean	113	90
Sample Var	3600	2916
Sample SD	60	54
Sample Corr.	0.8	
Sample Size	115	

10 (a) Sara thinks that this experiment should be analyzed as independent samples data. Assume that she is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

Solution: The difference of means (regular minus organic) is 23 calories. Sara calculates her standard error assuming independent samples:

$$\sqrt{\sigma_X^2/n + \sigma_Y^2/m} = \sqrt{3600/115 + 2916/115} = \sqrt{6516/115} \approx 7.5$$

so her confidence interval is approximately 23 ± 15, in other words $(8, 38)$.

10 (b) Kevin thinks that this experiment should be analyzed as matched pairs data. Assume that he is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

Solution: Kevin takes into account the sample correlation between columns when calculating his standard error. He does this by using the sample statistics from the table to calculate the sample variance of the *differences*: regular minus organic. In particular, he calculates:

$$s_D^2 = 3600 + 2916 - 2 \cdot 0.8 \cdot 60 \cdot 54 = 1332$$

which gives a standard error of

$$\sqrt{s_D^2/n} = \sqrt{1332/115} \approx 3.4$$

This is the only difference between his procedure and Sara's. Hence, Kevin's confidence interval is approximately 23 ± 6.8, in other words $(16.2, 29.8)$.

5 (c) How do the confidence intervals constructed by Sara and Kevin differ? Explain the source of the discrepancy. Which of them has constructed the appropriate confidence interval for this example?

Solution: Kevin is right and Sara is wrong. This is matched pairs data because each row corresponds to a *single individual*. Unsurprisingly, we find a high sample correlation between the two columns: individuals who overestimate caloric content for one yogurt sample tend to do so for the other, as do individuals who underestimate. The only difference between Kevin and Sara's confidence intervals comes from how they calculated their standard errors. Both intervals are correctly centered, but Sara's is *too wide* because she calculated the standard error assuming independence between the two samples. When the sample correlation is positive this results in an *overestimate* of the standard error.