First Midterm Examination

Econ 103, Statistics for Economists

February 10th, 2015

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	20	30	15	40	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, two points will be deducted for each page on which you do not write your name and student ID.

1. This question concerns the Anchoring Effect and the classroom exercise that we carried out to study it in our second class meeting. To jog your memory recall that in this exercise I asked you about the proportion of U.N. member states in Africa. In each of your answers to the following parts *write no more than three sentences.*

   (a) (4 points) Briefly explain the Anchoring Effect.

   Solution: The Anchoring Effect is a behavioral anomaly in which people's decisions are influenced by information they know to be irrelevant. In our classroom example, students from a past semester gave, on average, higher answers to the question "what fraction of all the countries in the UN are located in Africa" when they were shown a high random number (65) compared to a low random number (10).

   (b) (3 points) Was the data we collected in class to study the Anchoring Effect experimental or observational? Explain briefly.

   Solution: Experimental: each student was randomly assigned to either the High group or the Low group.

   (c) (4 points) Given your answer to the preceding part, is there any concern that our classroom results may have been influenced by a confounder? Explain briefly.

   Solution: Since we carried out a randomized experiment, we don't have to worry about confounding. On average, the students in the High group and the Low group were the same in all characteristics except the number they were shown. This allows us to attribute any differences that emerge to the Anchoring Effect.

   (d) (4 points) Did our classroom exercise to study the Anchoring Effect use a simple random sample? If so, from what population? Explain briefly.

   Solution: There are various possible correct answers. One could try to argue that the students taking Econ 103 during a particular semester constitute a random sample of Penn Economics majors. Alternatively, one could point out that students taking the course in one semester rather than the other might differ in important ways. We certainly wouldn't want to treat the students in this class as a random sample of Penn Undergrads or population more general than that!

2. Oleg and Julia observe a dataset of n students: $x_1, x_2, \ldots, x_n$. Observation x_i takes on

the value one if that student is male, and zero otherwise. Whenever asked to "explain briefly" below, write no more than three sentences.

(a) (4 points) Is x numerical, ordinal, or nominal? Explain briefly.

Solution: Although it is *represented* using the numbers zero and one, this variable is in fact nominal. The information that these numbers encode is simply Male versus Female, two categories that have no natural ordering.

(b) (6 points) Oleg wants to summarize this dataset so he calculates $\bar{x} = (\sum_{i=1}^{n} x_i)/n$ and gets a result of 0.4. Julia decides instead to calculate the sample *proportion*: she counts up the total number of ones in the dataset, and divides by n. Will Julia and Oleg's results be the same, or will they differ? Explain briefly.

Solution: As we saw in the Titanic exercise from R Tutorial #2, calculating the sample mean of a binary variable, one that only takes on the values one or zero, is *identical* to calculating the sample proportion. Even if you forgot this, you can easily see it from the formula for the sample mean: the only observations that contribute to the result are the ones. This means that Julia will get the same result as Oleg, namely 0.4. In words, this means that 40% of the students in the dataset are male.

(c) (10 points) Suppose that n is very large so that $n/(n-1) \approx 1$. Roughly what is the sample variance of this dataset? Hint: using the properties of summation notation and the fact that x_i can only take on the values zero and one there is a way to write s_x^2 *solely* in terms of $\bar{x}$ and $n/(n-1)$. Recall from above that $\bar{x} = 0.4$.

Solution: Expanding and splitting up the sum,

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \end{aligned}$$

All of the above calculations hold *regardless* of the values that x can take on. But in this case, x_i is either zero or one which means that $x_i = x_i^2$. Thus, the first term inside the square brackets equals $n\bar{x}$ in this special case! Simplifying, we find that when x is binary

$$s_x^2 = \frac{n}{n-1}\left(\bar{x} - \bar{x}^2\right)$$

Name: _____ Student ID #: _____

Since the sample mean that Oleg calculated was 0.4, and n is assumed to be large, the sample variance is approximately $0.4 - 0.16 = 0.24$.

3. Consider the following simple dataset with nine observations of two variables:

x	y
2	2
3	2
4	2
2	3
3	3
4	3
2	4
3	4
4	4

(a) (4 points) Calculate $\bar{x}$ and $\bar{y}$.

Solution: The sample mean is 3 for both x and y since $3 \times (2+3+4)/9 = 3$.

(b) (4 points) Calculate s_x^2 and s_y^2.

Solution: The calculation is the same for both:

$$3 \times \left[(2-3)^2 + (3-3)^2 + (4-3)^2\right]/8 = 3/4$$

(c) (6 points) Calculate r_{xy}.

Solution: Since $r_{xy} = s_{xy}/(s_x s_y)$, we first need to calculate the covariance.

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
-1	-1	1
0	-1	0
1	-1	-1
-1	0	0
0	0	0
1	0	0
-1	1	-1
0	1	0
1	1	1

Summing the third column and dividing by $n-1$ gives the covariance. Since

the sum is zero, so is the covariance. Since a fraction is zero if and only if its numerator is zero, we don't need to calculate anything further: $r_{xy} = 0$.

(d) (6 points) Calculate the slope and intercept of a linear regression model that uses this dataset to predict y from x.

Solution: The regression slope is s_{xy}/s_x^2. Since the covariance is zero, so is the regression slope. Since the regression line goes through the means of the data, $\bar{y} = a + b\bar{x}$ but since $b = 0$, we have $a = \bar{y} = 3$.

4. Rossa and Rodrigo are playing their favorite game: matching pennies. The game proceeds as follows. In each round, each player flips a penny. If the flips match (TT or HH) Rossa gets one point; if the flips do not match (TH or HT) Rodrigo gets one point. The game is best of three rounds: as soon as one of the players reaches two points, the game ends and that player is declared the winner. Since there's a lot of money on the line and graduate students aren't paid particularly well, Rossa secretly alters each of the pennies so that the probability of heads is 2/3 rather than 1/2. In spite of Rossa's cheating, the individual coin flips remain independent.

(a) (6 points) Calculate the probability that Rossa will win the first round of this game.

Solution: Rossa wins a given round if either of the two mutually exclusive outcomes HH or TT occurs. Thus:

$$P(\text{Rossa Wins}) = P(HH) + P(TT) = (2/3)^2 + (1/3)^2 = 5/9$$

(b) (6 points) Calculate the probability that the game will last for a full three rounds.

Solution: We need to calculate the probability of a tie after two rounds. There are two ways that a tie could occur: either Rossa wins the first round while Rodrigo wins the second, or Rodrigo wins the first round while Rossa wins the second. These two events are mutually exclusive and the probability of each is $5/9 \times 4/9 = 20/81$ since successive coin flips are independent. Thus, the desired probability is 40/81.

(c) (8 points) Calculate the probability that Rodrigo will win the game.

Solution: Rodrigo needs to win two rounds to win the game. There are three ways this can happen. First, Rodrigo could win both rounds 1 and 2, in which case no third round is played. The probability of this event is $4/9 \times 4/9 = 16/81$. Second Rodrigo could lose round 1 but win rounds 2 and 3. The probability of this event is $5/9 \times 4/9 \times 4/9 = 80/729$. Finally, Rodrigo could lose round 2 but win rounds 1 and 3. The probability of this event is $4/9 \times 5/9 \times 4/9 = 80/729$. Summing these probabilities, since their corresponding events are mutually exclusive, the probability that Rodrigo wins the game is $304/729 \approx 0.417$.

(d) (10 points) Yiwen is walking down the hallway and sees Rodrigo doing his victory dance: clearly Rossa has been defeated in spite of rigging the game. Given that Rodrigo won, calculate the probability that the game lasted for three rounds.

Solution: By the definition of conditional probability,

$$P(\text{3 Rounds}|\text{Rodrigo Won}) = \frac{P(\text{3 Rounds} \cap \text{Rodrigo Won})}{P(\text{Rodrigo Won})}$$

We already calculated the denominator in the preceding part: it equals 304/729. To calculate the numerator we simply add up the probabilities of the two mutually exclusive ways in which Rodrigo could win in three rounds: (Win, Lose, Win) and (Lose, Win, Win). We calculated these probabilities in the preceding part: both were 80/729 so the numerator is 160/729. Taking the ratio of these gives $160/304 \approx 0.526$. Given that Rodrigo won, it is slightly more likely than not that the game lasted for a full three rounds.

5. (15 points) Sherlock Holmes has gone away on vacation, instructing Dr. Watson to water the flowers in his absence. Unfortunately Watson has a rather poor memory: the probability that he will remember to water the flowers is only 2/3. The flowers weren't in the best shape when Holmes left: even if watered the probability that they will wither and die before Holmes returns is 1/2. If they aren't watered, the probability that they will wither and die increases to 3/4. Holmes returns to find that his flowers have died. What is the probability that Watson forgot to water them?

Solution: By Bayes' Rule:

$$P(\text{Forget}|\text{Die}) = \frac{P(\text{Die}|\text{Forget})P(\text{Forget})}{P(\text{Die})}$$

We calculate the denominator using the law of total probability as follows:

$$\begin{aligned} P(\text{Die}) &= P(\text{Die}|\text{Forget})P(\text{Forget}) + P(\text{Die}|\text{Remember})P(\text{Remember}) \\ &= 3/4 \times 1/3 + 1/2 \times 2/3 = 3/12 + 2/6 = 7/12 \end{aligned}$$

Thus $P(\text{Forget}|\text{Die}) = (3/12)/(7/12) = 3/7$. It is more likely than not that Watson forgot to water the flowers.

6. This question concerns an R dataframe called `tips` containing data collected by a waiter on the amount of money he recieved as tips and the characteristics of the tables he served at the restaurant. Here are the first few rows of the dataframe:

```
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner    4
6      25.29 4.71   Male     No Sun Dinner    4
```

Each row corresponds to a particular table that this waiter served and there are no missing values. The first two columns are both measured in US dollars: `total_bill` gives the total bill while `tip` gives the amount of the tip. To be clear: `total_bill` *does not include the tip*. The next four columns are categorical: `sex` is either `Female` or `Male` indicating the sex of the person in the party who paid the bill, `smoker` is either `Yes` or `No` indicating whether there were any smokers in the party, `day` indicates the day of the week when this party came to the restaurant (`Thurs`, `Fri`, `Sat`, or `Sun`), and `time` indicates whether the meal served was `Lunch` or `Dinner`. The final column, `size`, is a count of the number of diners in the party.

(a) (3 points) What R command did I use to display the first few rows of the `tips` dataframe above?

Solution: `head(tips)`

(b) (3 points) Write a line of R code to make a scatterplot with `total_bill` on the x-axis and `tip` on the y-axis.

Solution: `plot(tip ~ total_bill, tips)`

(c) (3 points) Write a line of R code that will create a vector called `percent` containing the tips left by each table in `tips` as a *percentage* of the total bill. Express the values as percentage points rather than decimals. For example, if a table left a tip of \$10 on a \$50 bill, the corresponding element of `percent` should be 20.

Solution: `percent <- 100 * tips$tip / tips$total_bill`

(d) (3 points) Write a line of R code that will create a new dataframe called `smokers` containing only those rows of `tips` corresponding to tables with smokers.

Solution: `smokers <- subset(tips, smoker == "Yes")`

(e) (3 points) Write a line of R code to carry out a linear regression where `tip` is the y-variable and `total_bill` is the x-variable.

Solution: `lm(tip ~ total_bill, tips)`

(f) (5 points) The results of the preceding regression are given below. Explain them.

```
Coefficients:
(Intercept)   total_bill
     0.9203       0.1050
```

Solution: The regression line is approximately $\hat{y} = 0.92 + 0.11x$. For each additional dollar on the bill, we predict about 11 cents of additional tip. The intercept is probably not meaningful: it means that we predict a tip of 92 cents on a bill of zero but a bill of zero means that you didn't eat in the restaurant!

(g) (5 points) Write R code to calculate the mean of `percent` broken down by `sex` and `smoker`. You can do this in one command or several: either is fine.

Solution: `as.table(by(percent, tips[,c("sex", "smoker")], mean))`

(h) (5 points) The results of running the command from the preceding part are given below. Explain these results in no more than three sentences.

```
          smoker
sex        No  Yes
```

Name: _____ Student ID #: _____

```
Female 15.7 18.2
Male   16.1 15.3
```

Solution: Various possibilities: here are some that jump out at me. Parties with smokers where a woman pays the bill leave the largest tips by far: about 18% on average. Tips left by other kinds of tables are fairly similar: between 15 and 16%. Among tables with smokers, women are more generous in tipping than men. The reverse is true for tables without smokers, although the difference is small.

(i) (10 points) As we will learn later in the course, we need to be careful when comparing sample means from different sub-groups. In particular, we need to take account of how accurately each sample mean estimates the corresponding population mean and this depends on the sample size of each group. The measure we will use to quantify this idea later in the semester is called the *standard error of the mean*. For a dataset $x_1, \ldots, x_n$, it is defined as $SE = s_x/\sqrt{n}$. Write an R function called `getSE` to calculate this quantity. Your function should accept a single input argument `x`, the vector of data for which we will calculate the standard error, and return SE as defined above. In your answer you may use any R functions you like *except* `var` and `sd`. You may assume that `x` does not contain any missing values.

Solution: There are many possible correct answers. Here's one:

```
getSE <- function(x){
  n <- length(x)
  var.x <- sum((x - mean(x))^2) / (n - 1)
  SE <- sqrt(var.x) / sqrt(n)
  return(SE)
}
```

Name: _____ Student ID #: _____