# 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.

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

   (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.

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

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.

   (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.

   (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 $\bar{x} = 0.4$.

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 $\bar{y}$.

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

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

Name: _____ Student ID #: _____

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

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.

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

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

(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.

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?

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.

Name: _____ Student ID #: _____

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

(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.

(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.

(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.

(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.

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

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

(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.

(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
  Female 15.7 18.2
  Male   16.1 15.3
```

(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.