

MIDTERM EXAMINATION I  
ECON 103, STATISTICS FOR ECONOMISTS

FEBRUARY 11TH, 2014

**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: \_\_\_\_\_

Student ID #: \_\_\_\_\_

Signature: \_\_\_\_\_

Question:	1	2	3	4	5	6	Total
Points:	15	30	15	35	15	30	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. Consider a dataset of  $n$  observations  $x_1, x_2, \dots, x_n$  with sample mean  $\bar{x}$  and sample variance  $s_x^2$ . Let  $z_i$  denote the sample z-score corresponding to the observation  $x_i$ .

(a) (2 points) Write down the formula for  $\bar{x}$ .

(b) (2 points) Write down the formula for  $s_x^2$ .

(c) (3 points) Write down the formula for  $z_i$ .

(d) (8 points) Prove that the sample mean of the z-scores is zero.

Name: \_\_\_\_\_

Student ID #: \_\_\_\_\_

2. In this question you will analyze a dataset containing *last semester's* final exam scores and math diagnostic test scores. Both scores are given in points out of 100. To answer the questions given below, you will need to consult the following table of sample statistics for the dataset:

	Diagnostic	Final Exam
1st Quartile	58	51
Median	68	66
Mean	68	65
3rd Quartile	80	78
Std. Dev.	16	17
Covariance	124	

- (a) (5 points) As you can see from the table, the first quartile for the diagnostic test was 58. Briefly explain what this means in terms that someone who has never taken Econ 103 would understand.
- (b) (5 points) Is there any evidence of skewness in the math diagnostic or final exam scores? Explain briefly.
- (c) (5 points) Were scores more variable on the final or the math diagnostic? Briefly discuss in terms of both the standard deviation and interquartile range.
- (d) (5 points) Calculate the sample correlation between scores on the math diagnostic test and those on the final exam.

Name: \_\_\_\_\_

Student ID #: \_\_\_\_\_

- (e) (5 points) Suppose you wanted to carry out a regression to predict a student's final exam score using her math pre-test score. Calculate the slope and intercept of the regression line.
- (f) (5 points) Lucy scored 80 on the math pre-test while Linus only scored 60 on the pre-test. Using your answer to the previous part, how much higher would we predict Lucy's score on the final exam will be than Linus' score on the final?
3. Let  $A$  be the event that it rains this Saturday,  $B$  be the event that it rains this Sunday and  $C$  be the event that it rains this weekend. In her weather forecast Molly, the local meteorologist, tells us that  $P(A) = 0.5$  and  $P(B) = 0.5$ .
- (a) (2 points) Express the event  $C$  in terms of the events  $A$  and  $B$  using set operations.
- (b) (2 points) In this example, what is the meaning of the event  $A \cap B$ ? Phrase it in a way that someone who has never taken Econ 103 would understand.
- (c) (3 points) Express  $P(C)$  in terms of  $P(A \cap B)$  using the addition rule.

- (d) (8 points) Adam, an anchorman for the local news, sees Molly's forecast and summarizes it as follows: "According to Molly we're in for a wet weekend. There's a 100% chance of rain this weekend: 50% on Saturday and 50% on Sunday." Is Adam correct? If so, briefly explain why; if not, point out the flaw in his reasoning.

4. On my desk I have 10 cups:  $N_B$  of them are *Blue Cups* and the remaining  $10 - N_B$  are *Red Cups*. Each cup contains five balls: *Blue Cups* contain 4 blue balls and 1 red ball while *Red Cups* contain 4 red balls and 1 blue ball. I chose a cup at random so that each cup was equally likely to be selected. I then drew three balls at random *with replacement* from the chosen cup. In order, the balls I drew were: red, red, blue. Let  $C_B$  be the event that I chose a *Blue Cup* and let  $RRB$  be the event that represents my three draws: a red ball, followed by another red ball, followed by a blue ball.

- (a) (15 points) Suppose  $N_B$  is 5. Calculate  $P(C_B|RRB)$ .

- (b) (15 points) Now suppose that we do *not* know the value of  $N_B$ . How large would  $N_B$  have to be for it to be more likely that I drew from a blue cup given that the event  $RRB$  has occurred? Prove your answer.

- (c) (5 points) Suppose that I made my draws *without* replacement. What is  $P(C_B|RRB)$  in this case? Briefly explain your answer.

Name: \_\_\_\_\_

Student ID #: \_\_\_\_\_

5. The so-called “Iris Dataset” comes pre-loaded in R in the dataframe `iris`. Here’s a description from the R documentation:

This famous (Fisher’s or Anderson’s) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

A *sepal* is a part of a flower, specifically one of the small leaves found behind the petals. Here are the first few rows of the dataset:

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1           3.5           1.4           0.2  setosa
2           4.9           3.0           1.4           0.2  setosa
3           4.7           3.2           1.3           0.2  setosa
4           4.6           3.1           1.5           0.2  setosa
5           5.0           3.6           1.4           0.2  setosa
6           5.4           3.9           1.7           0.4  setosa
```

Note that the column `Species` is a categorical variable, aka factor, that takes on three different values: `setosa`, `versicolor`, and `virginica`.

- (a) (5 points) Suppose you wanted to display only the columns `Sepal.Length` and `Petal.Width` of `iris`. What R command would you use?
- (b) (5 points) What R command would you use to extract data for only flowers of the species *Iris setosa* and store it in a dataframe called `setosa`?
- (c) (5 points) What R command would you use to separately calculate the sample mean `Sepal.Length` for *each species* of iris? Be sure to allow for the possibility of missing values.

6. (30 points) Write an R function called `reg.predict` that uses linear regression to predict missing  $y$ -values from their  $x$ -values. In your answer you may use any R functions you like *except* `lm`. The function should take three arguments: `x`, `y` and `x.new`. The vectors `x` and `y` contain the data for which we observe both the  $x$  and  $y$ -values, while `x.new` is a collection of  $x$ -values for which we do not know the corresponding  $y$ -values (e.g. the midterm grades example from lecture). Your function should carry out the following steps. First, use `x` and `y` to calculate the intercept  $a$  and the slope  $b$  of the regression line  $\hat{y} = a + bx$ . Second, calculate `y.pred`, a vector containing the  $y$ -values that we would *predict* for the  $x$ -values contained in `x.new`, using the regression slope and intercept from the first step. Finally, return `y.pred`. In your answer you may assume that `x` and `y` are of the same length and that there are no missing values. Again, you may use any R functions you like in your answer *other than* `lm`.

Name: \_\_\_\_\_

Student ID #: \_\_\_\_\_