

FIRST MIDTERM EXAMINATION  
ECON 103, STATISTICS FOR ECONOMISTS  
SEPTEMBER 28TH, 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	7	Total
Points:	30	25	10	10	15	20	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. Indicate whether each of the following statements is True or False. If False, provide a one sentence explanation. If True, no explanation is needed.

- 3 (a) The sample covariance between  $x$  and *itself* equals the sample variance of  $x$ .
- 3 (b) The sample variance of the z-scores always equals zero.
- 3 (c) If a variable has a positive skewness, then we would generally expect its mean to be *less than* its median.
- 3 (d) In a large population that is approximately bell-shaped, roughly 68% of the z-scores will be between -2 and 2.

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

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

- 3 (e) For any events  $A$  and  $B$ ,  $P(A \cap B) = P(B \cap A)$ .
- 3 (f) For any events  $A$  and  $B$ ,  $P(B) = P(A \cap B)/P(B|A)$
- 3 (g) A random variable is neither random, nor a variable.
- 3 (h) The expected value of a Bernoulli( $p$ ) random variable equals  $1/2$ .
- 3 (i) If  $X \sim \text{Bernoulli}(p)$  then  $E[X^2] = p$ .

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

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

- 3 (j) If  $X$  is a discrete RV with pmf  $p(x)$ , then  $\sum_{\text{all } x} xp(x) = 1$ .

2. Five friends want to lose weight so they sign up for a five-week program of fitness classes. At the end of the program, they note down how many weeks of classes they each attended,  $x$ , and the change in their weight in kilograms,  $y$ . The data appear in the table below. Since  $y$  denotes a *change* in weight,  $y_i = -1$  means that person  $i$  lost one kilogram.

**Note:** *If you get an early part to this question incorrect and have to use the results of this part later on, we will take account of this in grading and assign partial credit accordingly. We have left space in the table for your calculations.*

$x$	$y$	
3	0	
3	-1	
4	0	
5	-2	
5	-2	

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

- 6 (b) Calculate the sample variances,  $s_x^2$  and  $s_y^2$ .

- 4 (c) Calculate the sample covariance between  $x$  and  $y$ .
- 6 (d) Calculate the slope and intercept of a linear regression that uses the dataset for the five friends, given above, to predict weight change,  $y$ , from number of weeks of classes attended,  $x$ .
- 5 (e) Each of the five friends in our dataset above attended *more* than two weeks of fitness classes. Suppose we wanted to predict the change in weight for someone *not in our dataset* who attended only two weeks of classes. Based on the regression results from the preceding part, how much would you predict that this person's weight would change?

- 10 3. Use what you know about summation notation to prove the following equality:

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \left( \sum_{i=1}^n x_i y_i \right) - n\bar{x}\bar{y}$$

4. Suppose I flip a fair coin and roll a single fair die at the same time. Define the events  
 $A$  = the coin comes up tails  
 $B$  = the die shows a 3 or 5  
 $C$  = the die shows an *odd* number

- 3 (a) Calculate  $P(B|C)$ .

- 3 (b) Calculate  $P(A \cap B)$ .

- 4 (c) Calculate  $P(A \cup B)$ .

5. This question asks you to prove some results about probability that we did not cover in class. Let  $A$ ,  $B$  and  $C$  be three arbitrary events.

5 (a) Prove *Boole's Inequality*:  $P(A \cup B) \leq P(A) + P(B)$ .

5 (b) Does Boole's Inequality also apply when there are three events, i.e. is it true that  $P(A \cup B \cup C) \leq P(A) + P(B) + P(C)$ ? Explain why or why not.

5 (c) Prove *Bonferroni's Inequality*:  $P(A \cap B) \geq P(A) + P(B) - 1$ .

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

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

6. New England Patriots quarterback Tom Brady has a lucky coin, which he flips five times before every game. The coin is fair, so it has a probability  $1/2$  of coming up heads on any given flip.

5 (a) What is the probability that Tom gets the same outcome for all 5 flips of his lucky coin? Express your answer as a fraction.

15 (b) Tom is a superstitious person. If (and only if) he gets 5 heads or 5 tails he will cheat by deflating the balls used in the game. If Tom cheats, the Patriots win with probability  $\frac{3}{4}$ . But if Tom does not cheat, they win with probability  $\frac{2}{5}$ . Given that the Patriots won the game, what is the probability that Tom cheated?



7. This question asks you for the R commands needed to carry out tasks similar to those in the tutorials. The first two parts refer to the dataframe `jobsearch` that we created in R Tutorial 1. The entire dataframe is printed below.

```
##   location salary          title hours
## 1 New York  70000   Office Manager    50
## 2 Chicago  80000 Research Assistant    56
## 3 Boston   60000         Analyst     65
## 4 Boston   50000   Office Manager    40
## 5 New York  45000         Analyst     50
```

- 4 (a) Create a new dataframe called 'highpay' by taking a subset of the 'jobsearch' dataframe to include only jobs which pay more than \$55,000.
- 4 (b) Write code to calculate the number of observations in your new dataframe, `highpay`, and call this number `n`.

The next four parts of this question concerns the dataframe `survey` from R Tutorial 2. This dataframe has observations of six variables. The variable names, and the first six observations are printed below:

```
##      sex credits eye.color handedness height handspan
## 1  Male      5   Brown      1.0      67    20.0
## 2 Female      5   Brown      0.4      63    19.5
## 3 Female     NA   Brown      0.6      62    19.0
## 4 Female      5   Brown      0.6      65    19.5
## 5 Female      4   Brown      1.0      62    18.5
## 6 Female     NA   Brown      1.0      68    18.5
```

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

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

- 4 (c) Write the command that prints the first six rows of `survey`.
- 4 (d) Write the command to draw a scatter plot with handedness on the horizontal axis and height on the vertical axis.
- 4 (e) Create a dataframe called `numerical` that contains only the numerical variables in our dataframe above.

The final part of this question does not refer to any of the datasets from above: instead it asks to you demonstrate your knowledge of R functions.

- 10 (f) Write an R function called `myIQR` that calculates the interquartile range of a vector of data. Your function should take only one argument: the vector of data `x`. You may assume that `x` does not contain any missing values. You may use any R functions you like in your answer *except* the built-in function for calculating an interquartile range, namely `IQR`.