Midterm Examination I

Econ 103, Statistics for Economists

February 11, 2013

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	7	8	9	Total
Points:	10	10	15	20	20	10	15	20	20	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, fifteen points will be deducted from your final score. In addition, one point will be deducted for each page on which you do not write your name and student ID.

1. (10 points) A study conducted at the University of Missouri-Columbia ("Mizzou") found that students who had a fake ID were more likely to engage in underage drinking while at college. Suppose we were to give fake IDs to a randomly selected group of Mizzou freshmen. Based only on the evidence given in this question, do you think this would substantially increase the fraction of these students who drink underage? Explain why or why not.

2. (10 points) Let A and B be two mutually exclusive events, where both $P(A)$ and $P(B)$ are strictly greater than zero. Are A and B independent? Explain your answer and provide a proof. You may use any results we derived in class without proving them but be sure to name the results you use for full credit.

Name: _____ Student ID #: _____

3. Suppose that we want to predict y from x using the linear regression model $\widehat{y} = a + bx$.

   (a) (5 points) Write down (but do not solve) the optimization problem needed to calculate a and b from a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

   (b) (10 points) Using your answer to part (a), prove that the regression line goes through the means of the data, that is $\bar{y} = a + b\bar{x}$.

4. This question refers to the following dataset, containing 13 observations:

   $$-3 \quad -3 \quad -2 \quad -1 \quad -1 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 4 \quad 7 \quad 13$$

   (a) (5 points) Calculate the median of this dataset.

   (b) (5 points) Calculate the mean of this dataset.

(c) (5 points) Suppose that it turned out there was a mistake recording the dataset: the observation listed as 13 should actually be 130. How would the mean and median change?

(d) (5 points) Let f be a strictly increasing function, that is $x_1 < x_2 \Rightarrow f(x_1) < f(x_2)$. Suppose I apply f to the *original dataset* so that instead of $-3, -3, \ldots, 7, 13$ the data become $f(-3), f(-3), \ldots, f(7), f(13)$. What is the median of the transformed data? Explain your answer.

5. Consider a dataset with n observations on a variable x: $x_1, \ldots, x_n$. Define a new variable, y, as follows: for each x_i set $y_i = c + dx_i$ where c and d are constants and $d \neq 0$.

   (a) (5 points) How is s_y^2 related to s_x^2? Prove your answer.

(b) (5 points) How is s_{xy} related to s_x^2? Prove your answer.

(c) (5 points) What is the sample correlation between x and y? Prove your answer.

(d) (5 points) Suppose we were to carry out linear regression to predict y from x, namely $\hat{y} = a + bx$. What values would we find for a and b? Prove your answer. (You may use the regression formulas from class without proving them.)

6. (10 points) Let A and B be two events where $P(B) > 0$. Prove that

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

In your answer you may use any of the rules we learned in class except for the one you are being asked to prove. For full credit, provide the name of each rule that you use.

7. The Triangle is a neighborhood that once housed a chemical plant but has become a residential area. Two percent of the children in the city live in the Triangle, and fourteen percent of these children test positive for excessive presence of toxic metals in the tissue. For children in the city who do not live in the Triangle, the rate of positive tests is only one percent. Let T be the event that a child lives in the Triangle and M be the event that a child tests positive for excessive presence of toxic metals.

   (a) (5 points) If we randomly select a child who lives in the city, what is the probability that she both lives in the Triangle and tests positive?

   (b) (5 points) If we randomly select a child who lives in the city, what is the probability that she tests positive?

(c) (5 points) If we randomly select a child who lives in the city and she tests positive, what is the probability that she lives in the Triangle?

8. Let X be a random variable that takes on the values 1, 2, and 3 with equal probability.

(a) (4 points) Define the term *random variable*.

(b) (4 points) What is the support set of X?

(c) (4 points) What is the pmf of X? Write it out as a piecewise function.

(d) (4 points) What is the CDF of X? Write it out as a piecewise function.

(e) (4 points) Calculate the expected value of X.

9. This question refers to commands from the R statistical package that you have studied in recitation. Suppose I have a dataframe called `survey` with two columns. The first column is `height` and the second is `handspan`. Both contain numeric data. Here are the first few lines of the dataset:

```
  height handspan
1     67     20.0
2     63     19.5
3     62     19.0
4     65     19.5
5     62     18.5
6     68     18.5
```

(a) (4 points) What command would I use to display the column `handspan` only?

(b) (4 points) Suppose I wanted to display only those rows from `survey` corresponding to students whose height is greater than 60 inches. What command would I use?

(c) (4 points) Suppose I wanted to display only the 2nd and 9th rows of `survey`. What command would I use?

(d) (4 points) When looking at my data, I see that one of the values for handspan is `NA`. What does this mean?

(e) (4 points) Suppose that, for some strange reason, I wanted to rename the column `height` to `altitude` but leave the name of `handspan` unchanged. What command would I use?

Name: _____ Student ID #: _____