Midterm Examination I

Econ 103, Statistics for Economists

September 30, 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	15	20	20	10	15	10	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, 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. (10 points) Saleem has gathered a large dataset comprising a random sample from the population of all US deaths between 2008 and 2012. For each death, Saleem's dataset contains detailed survey information listing the deceased's habits, lifestyle and age at death. Shortly after beginning his data analysis, Saleem uncovers a startling pattern: among those who died under the age of 25, a large majority were regular users of *Twitter*. In contrast, hardly anyone who died over the age of 75 was a regular *Twitter* user. Saleem concludes that *Twitter* use is an important cause of early death in the US. Do you agree with his analysis? Explain.

2. The *quartile deviation* is a measure of dispersion that we did not cover in class. Let Q_3 be the 75th percentile and Q_1 be the 25th percentile of the sample. Then the quartile deviation is given by the expression $(Q_3 - Q_1)/2$.

   (a) (5 points) The quartile deviation is related to another measure of dispersion that we did study in class. Which one and how?

   (b) (5 points) What are the units of the quartile deviation relative to those of the data?

   (c) (5 points) Would you expect the quartile deviation to be more or less sensitive to outliers than the standard deviation? Explain briefly.

3. Consider an R dataframe called `measurements` with two columns: `height` measures a student's height in inches while `handspan` measures her handspan in centimeters. The first few rows are as follows:

```
 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
```

Running a linear regression with $y =$ `height` and $x =$ `handspan` gives:

```
Coefficients:
(Intercept)     handspan
         42          1.2
```

(a) (4 points) What R command produces the regression results given above?

(b) (4 points) What are the units of the slope and intercept from this regression?

(c) (4 points) What height would we predict for someone with a handspan of 10 cm?

(d) (4 points) The sample mean of `height` for this dataset is 68 inches. What is the sample mean of `handspan`? Feel free to round to the nearest centimeter.

(e) (4 points) The sample standard deviation of `height` is approximately twice that of `handspan`. What is the correlation between `height` and `handspan`?

Name: _____ Student ID #: _____

4. Suppose we want to carry out a linear regression with a set equal to zero, that is we want to predict y from x according to $\widehat{y} = bx$.

   (a) (5 points) Write down the optimization problem we need to solve and explain it.

   (b) (10 points) Solve the optimization problem you wrote down for part (a) to find the formula for b.

   (c) (5 points) Suppose that we wanted to predict y from x using a *quadratic* relationship, namely $\widehat{y} = bx^2$. How would your answers to parts (a) and (b) change?

5. (10 points) What is the probability of getting *at least* one six when rolling a fair, six-sided die three times? Explain your answer.

6. Suppose I throw two fair, six-sided dice once. Define the following events:

$$\begin{aligned} E &= \text{The first die shows 5} \\ F &= \text{The sum of the two dice equals 7} \\ G &= \text{The sum of the two dice equals 10} \end{aligned}$$

   (a) (2 points) What is $P(F)$?

   (b) (2 points) What is $P(G)$?

   (c) (3 points) Calculate $P(F|E)$.

   (d) (3 points) Calculate $P(G|E)$.

   (e) (5 points) Suppose you wanted to bet on the outcome of this random experiment. If I revealed whether or not the first die was a 5, would this give you any relevant information for betting on F? What about for betting on G? Explain briefly.

7. (10 points) I have two six-sided dice in my pocket: one fair die and one loaded die. The fair die has the usual probabilities, but the probability of getting a 6 when rolling the loaded die is 1/2. Suppose I reach into my pocket and draw one of the two dice at random (both are equally likely to be drawn). I roll this randomly chosen die and get a 6. What is the probability that I drew the loaded die?

8. Helen wants to know whether Penn graduate admissions are biased against women. She has obtained data on 12,600 recent applicants indicating: sex (Male or Female), the field to which she applied (Arts or Sciences) and whether or not she was admitted (1 = Admitted, 0 = Rejected). Helen stores her data in an R dataframe called `admissions` with columns `Sex`, `Field` and `Admitted`. The first few rows are as follows:

```
     Sex    Field Admitted
1   Male Sciences        1
2 Female Sciences        0
3 Female     Arts        0
4   Male     Arts        1
5   Male     Arts        0
6 Female     Arts        0
```

   (a) (6 points) The first thing Helen does is calculate the admissions rates for men and women separately. Write R code to accomplish this. You may assume there are no missing observations.

(b) (8 points) Next Helen calculates the admission rates for men and women broken down by field of study. Write R code to accomplish this. Again, you may assume that there are no missing observations.

(c) (6 points) Finally, Helen arranges her results from part (a) into the following table

```
   Female      Male
0.3488372 0.4457831
```

and does the same for results from part (b)

```
       Field
Sex           Arts  Sciences
  Female 0.2812500 0.5454545
  Male   0.3043478 0.5000000
```

Interpret Helen's findings. How do her results from part (a) compare to those from part (b)? Has Helen found evidence of discrimination against women in Penn graduate admissions?

9. (a) (10 points) Write an R function called `wavg` that calculates a *weighted average*:

$$\sum_{i=1}^{n} w_i x_i$$

where $x_1, x_2, \ldots, x_n$ is a dataset, and $w_1, w_2, \ldots, w_n$ is a corresponding set of *weights* between zero and one that sum to one. Your function should take two arguments: `x` is a numeric vector containing the data to be averaged and `w` is a numeric vector containing the corresponding weights. You may assume that there are no missing values in either `x` or `w` and that both vectors have the same length. The output of your function should be a single number: the average of `x` computed using the weights `w`.

(b) (10 points) Write another R function called `mymean` that takes a single argument `x` and calculates the usual sample mean by calling the function `wavg` with appropriate weights.

Name: _____ Student ID #: _____