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.

Solution: There are many correct answers. Here is one possibility: "Saleem is mistaken. In order to die under the age of 25 during the years studied, you must have been born *after* 1983. In contrast, to die over the age of 75, you must have been born *before* 1938. People who were born before 1938 did not grow up with computers and the internet, and are hence much less likely to become users of Twitter. In contrast, people born after 1983 came of age just as the internet and personal computing became commonplace. These people are correspondingly much more likely to become users of Twitter. Hence *year of birth* is a confounder in this example: you have to be born after 1983 to "die young" and if you were born recently, you're also much more likely to be a Twitter user."

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?

   Solution: It equals the interquartile range (IQR) divided by two.

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

   Solution: It has the same units as the IQR, namely those of the data themselves.

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

Name: _____ Student ID #: _____

Solution: Like the IQR, the quartile deviation will be less sensitive to outliers than the standard deviation. This is because, unlike the standard deviation, the quartile deviation discards the most extreme 50% of the data before calculating the spread.

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?

Solution:

```
lm(height ~ handspan, data = measurements)
```

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

Solution: The intercept is measured in inches, and the slope is measured in inches per centimeter.

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

Solution: We predict using $\widehat{y} = a + bx$. Plugging in $x = 10$ cm along with the regression coefficients from above, we have $\widehat{y} = 54$ inches.

(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.

Solution: We use the fact that the regression line goes through the means of the data: $\bar{y} = a + b\bar{x}$. Rearranging, $\bar{x} = (\bar{y} - a)/b = (68 - 42)/(1.2) \approx 22$ cm.

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

Solution: We use the fact that $r = b(s_x/s_y)$. The correlation is $1.2/2 = 0.6$.

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.

Solution: We choose b to minimize the sum of squared vertical deviations:

$$\min_b \sum_{i=1}^{n} (y_i - bx_i)^2$$

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

Solution: Differentiating with respect to b gives the first order condition:

$$-2\sum_{i=1}^{n} (y_i - bx_i)x_i = 0$$

Rearranging:

$$\begin{aligned} \sum_{i=1}^{n} y_i x_i &= \sum_{i=1}^{n} bx_i^2 \\ b &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

(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?

Solution: Just replace x_i with x_i^2 in both the optimization problem statement and the first order condition:

$$b = \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}$$

We can do this because x and y are *constants* in the optimization problem rather than variables.

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.

Solution: Using the complement rule:

$$P(\text{At Least One Six}) = 1 - P(\text{No Sixes})$$

And by independence:

$$P(\text{No Sixes}) = 5/6 \times 5/6 \times 5/6 = 125/216$$

Hence,

$$P(\text{At Least One Six}) = 1 - 125/216 = 91/216 \approx 0.42$$

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 $P(F)$?

Solution: Of the 36 basic outcomes of the experiment, the pairs (1,6), (6,1), (2,5), (5,2), (3,4), and (4,3) sum to 7. Hence the probability is 1/6.

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

Solution: Of the 36 basic outcomes of this experiment, the pairs (5,5), (4,6), and (6,4) sum to 10. Hence the probability is $3/36 = 1/12$.

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

Solution: By the definition of conditional probability,

$$P(F|E) = \frac{P(F \cap E)}{P(E)}$$

We know that $P(E) = 1/6$. The only way that $F \cap E$ can occur is if we roll (5,7). Hence $P(F \cap E) = 1/36$. Thus, $P(F|E) = (1/36)/(1/6) = 6/36 = 1/6$.

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

Solution: Again, by the definition of conditional probability,

$$P(G|E) = \frac{P(G \cap E)}{P(E)}$$

As before, $P(E) = 1/6$. The only way for $G \cap E$ to occur is if we roll (5,5). Hence $P(G|E) = (1/36)/(1/6) = 6/36 = 1/6$.

(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.

Solution: Using the results of the preceding parts, we see that $P(F|E) = P(F)$, hence$F$ and E are statistically independent. This means that knowing whether or not E has occurred gives us no additional information about the probability of F. In contrast $P(G|E) \neq P(G)$, so G and E are *not* statistically independent. Indeed, if we know that E has occurred, this *doubles* the chance that G has occurred.

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?

Solution: Let L be the event that I draw the loaded die, F be the event that I draw the fair die and 6 be the event that I roll a six. By Bayes' rule, we have

$$P(L|6) = \frac{P(6|L)P(L)}{P(6)}$$

Name: _____ Student ID #: _____

Calculating the denominator by the Law of Total Probability, we have

$$\begin{aligned} P(6) &= P(6|L)P(L) + P(6|F)P(F) \\ &= 1/2 \times 1/2 + 1/6 \times 1/2 \\ &= 1/4 + 1/12 \\ &= 1/3 \end{aligned}$$

Hence,

$$P(L|6) = \frac{1/4}{1/3} = 3/4 = 0.75$$

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.

Solution: Perhaps the simplest way is as follows:

```
men <- subset(admissions, Sex == 'Male')
women <- subset(admissions, Sex == 'Female')
mean(men$Admitted)
mean(women$Admitted)
```

A more compact way is to use `by`

```
by(admissions$Admitted, admissions$Sex, mean)
```

(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.

Solution: The quickest way is

```
by(admissions$Admitted, admissions[,c("Sex", "Field")], mean)
```

To make this into a nice, clean table we can simply store the result and use `as.table`. It's also possible to solve this using subset to break the data into four blocks. For example, to calculate the admissions rate for men who applied to study Arts, we could use,

```
men.arts <- subset(admissions,  Sex == "Male" & Field == "Arts")
mean(men.arts$Admitted)
```

The other four categories are similar.

(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?

Solution: From part (a) it appears that women are admitted at a lower rate than men. However, once we break down by field of study in part (b), we see that women are admitted at comparable rates to men in Arts and higher rates than men in Sciences. The difference is explained by the fact that the admissions rate is lower for *both* men and women in Arts than it is in Science. Hence it must be the case that most women have applied to Arts while most Men have applied to Sciences. While this evidence doesn't rule out the possibility of bias against women in individual instances, there is no overall evidence of discrimination against women.

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`.

Solution: There are many correct answers. Here is one:

```
wavg <- function(x, w){
answer <- sum(w * x)
return(answer)
}
```

(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.

Solution: Again, there are many correct answers. The key is to realize what weights to use and how to compute them.

```
mymean <- function(x){
n <- length(x)
weights <- rep(1/n, n)
answer <- wavg(x, weights)
return(answer)
}
```

Name: _____ Student ID #: _____