Midterm Examination #1

Econ 103, Statistics for Economists

September 29th, 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	Total
Points:	40	15	35	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. Mark each statement as True or False. If you mark a statement as False, provide a brief explanation. If you mark a statement as True, no explanation is needed.

   (a) (4 points) In a double-blind, randomized controlled trial, neither the patients participating in the study nor the statistician analyzing the results knows who was given the placebo and who was given the real drug.

   > **Solution:** FALSE: in a double-blind RCT it is the patients and *experimenters* who are blind, not the statistician. To find out if the treatment worked the statistician definitely needs to know which patients received it!

   (b) (4 points) In large populations that are approximately bell-shaped, roughly 95% of observations will lie within one standard deviation of the mean.

   > **Solution:** FALSE: roughly 68% of observations will lie within one standard deviation of the mean. Another way to correct this is to say that roughly 95% of observations will liw within *two* standard deviations of the mean.

   (c) (4 points) The average deviation of a data set from its mean is always zero.

   > **Solution:** TRUE. We showed in class that $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

   (d) (4 points) If the correlation between x and y is positive, then it must be smaller than the covariance between x and y.

   > **Solution:** FALSE: $r_{xy} = s_{xy}/(s_x s_y)$ so if, for example, s_x and s_y are both less than one, the correlation will be *larger* than the covariance.

   (e) (4 points) The complement rule is one of the axioms of probability.

   > **Solution:** FALSE: it is a *consequence* of the axioms, not an axiom itself.

   (f) (4 points) The intuition behind the addition rule is simply this: don't double-count $A \cap B$ when calculating the probability of $A \cup B$.

   > **Solution:** TRUE

   (g) (4 points) A random variable is neither random nor a variable: it is a fixed function.

Solution: TRUE. This is the definition from class.

(h) (4 points) If X is a random variable, the CDF $F(x_0)$ of X gives the probability that X exceeds a specified threshold x_0.

Solution: FALSE: it gives the probability that X *does not* exceed x_0, namely $P(X \leq x_0)$.

(i) (4 points) The support set of the Bernoulli random variable is $\{0, 1\}$.

Solution: TRUE

(j) (4 points) Let X be a random variable with support set $\{-1, 0, 1\}$ and probability mass function $p(-1) = 1/2, p(0) = 1/4, p(1) = 1/4$. Then $E[X] = 0$.

Solution: $1/2 \times -1 + 1/4 \times 0 + 1/4 \times 1 = -1/2 + 1/4 = -1/4 \neq 0$

2. Suppose I flip a fair coin three times. Let A be the event that I get a heads on the first toss and B be the event that I get tails on the third toss. When listing outcomes of the experiment, use the notation `[T/H][T/H][T/H]`. For example, `THT` indicates tails on the first toss, heads on the second, and tails on the third.

   (a) (3 points) How many basic outcomes are there in the sample space for this example?

   Solution: $2 \times 2 \times 2 = 8$ or you can write them all out and count by hand.

   (b) (3 points) Which basic outcomes make up the event $A \cap B$?

   Solution: `HHT, HTT`

   (c) (3 points) Which basic outcomes make up the event $A \cup B$?

   Solution: Everything *except* `TTH,THH`. In other words: `HHH, HTH, TTT, THT, HHT, HTT`.

   (d) (3 points) Which basic outcomes make up the event $(A \cup B) \cap (A \cap B)$?

   Solution: `HHT, HTT`

   (e) (3 points) Calculate the conditional probability of $A \cap B$ given $A \cup B$.

   Solution: $P\left[(A \cup B) \cap (A \cap B)\right] = P(A \cap B) = 2/8 = 1/4$ and $P(A \cup B) = 6/8 = 3/4$. Hence, the conditional probability is $(1/4)/(3/4) = 1/3$.

Name: ________________ Student ID #: ________________

3. An R dataframe called `height.data` records the annual earnings in US dollars, height in inches, and sex of 1192 individuals. In the sample, the mean earnings are $20,400 and the mean height is 67 inches. Here are the first few rows of the dataframe:

```
 earn height    sex
50000     74   male
60000     66 female
30000     64 female
50000     63 female
51000     63 female
 9000     64 female
```

(a) (4 points) Suppose I were to use a linear regression of the form $\widehat{y} = a + bx$ to predict `earn` from `height`. What would be the units of a? What would be the units of b?

Solution: The units of a would be dollars, and the units of b would be dollars per inch.

(b) (3 points) Write out the full R command you would use to calculate a and b from the previous part using the data contained in `height.data`.

Solution:

```
lm(earn ~ height, data = height.data)
```

(c) (4 points) The results from the preceding part are $\widehat{y} = -60000 + 1200x$. Who would you predict will earn more: someone who is 5 feet tall or someone who is 6 feet tall? What difference in earnings would you predict for these two individuals?

Solution: We would predict that the taller person would earn $12 \times 1200 = 14400$ dollars more per year.

(d) (8 points) Suppose I were to create an R vector called `height.center`, as follows
`height.center <- height.data$height - mean(height.data$height)`
and then run a linear regression predicting `earn` from `height.center`. What would be the regression intercept? Explain your answer.

Solution: The vector `height.center` is a *centered* version of `height`, constructed by subtracting the sample mean height from each observation. As we showed in clase, the average deviation of any dataset from its mean is zero, so the sample mean of `height.center` is zero. Thus, if we run a linear regression with `height.center` as the x-variable, we'll have $a = \bar{y} - b\bar{x} = \bar{y} - b \times 0 = \bar{y}$. Thus the intercept will simply be the sample mean of the y-variable, `earn`, which we were told in the problem statement is \$20,400.

(e) (8 points) Write R code to create two dataframes: `males` contains only the observations from `height.data` for which `sex` is `male`, and `females` contains only the observations from `height.data` for which `sex` is `female`. Then use these dataframes to calculate the average height and average earnings *separately* for each group.

Solution:

```
males = height.data[sex == 'male']
females = height.data[sex == 'female']
males[ , .(mean(height), mean(earn))]
females[ , .(mean(height), mean(earn))]
```

Or, in one line: `height.data[ , .(mean(height), mean(earn)), by = sex]`
If your data is not a `data.table` but a `data.frame`, you would use:

```
males <- subset(height.data, sex == "male")
females <- subset(height.data, sex == "female")
mean(males$height)
mean(females$height)
mean(males$earn)
mean(females$earn)
```

(f) (8 points) The results of the commands from the preceding part are as follows:

	females	males
mean `earn`	\$18000	\$30000
mean `height`	65 in	70 in

Based on all the results presented above, do you think there is a causal relationship between height and income? Why or why not? Explain briefly.

Solution: We can't tell simply from running a regression whether height causes income. Based on the results presented here, however, there is reason to be suspicious. Sex is clearly a confounder since women are, on average, shorter than men, and earn less. The relationship we found between height and income may be nothing more than a consequence of labor market discrimination against

Name: ____________________ Student ID #: ____________________

women.

4. (a) (15 points) Write a function called `tip.calculator` that calculates a restaurant tip. (Don't worry about taxes or rounding your results to the nearest cent.) Your function should take two inputs: `bill` is the restaurant bill *excluding tip* in dollars and cents, e.g. `34.50`, and `percent` is the desired tip in percentage points, e.g. `18` for 18%. Your function should return a dataframe with columns named `bill`, `percent`, `tip`, and `total`. The first two elements `bill` and `percent` are the function inputs while `tip` is the tip in dollars and cents and `total` is the total bill *including tip*. For example, if I input `45` for `bill` and `20` for `percent`, your function should return:

```
bill percent tip total
  45      20   9    54
```

Solution:

```
tip.calculator <- function(bill, percent){
  tip <- percent/100 * bill
  total <- bill + tip
  return(data.frame(bill, percent, tip, total))
}
```

(b) (5 points) After creating the `tip.calculator` function, suppose I entered the following commands at the R console:

```
x <- c(1, 10, 100)
y <- c(100, 10, 1)
tip.calculator(bill = x, percent = y)
```

Write out in full the output that R will generate from the last command, namely `tip.calculator(bill = x, percent = y)`.

Solution:

```
bill percent tip total
   1     100   1     2
  10      10   1    11
 100       1   1   101
```

The point is to recognize that: (1) all basic mathematical operations in R are vectorized, and (2) the names of the objects provided as arguments to a function are irrelevant.

5. Approximately 80% of all emails sent over the internet are spam. About 10% of spam

Name: ______________________ Student ID #: ______________________

emails contain the word "viagra" compared to 1% of non-spam emails. About 5% of spam emails contain the word "herbal" compared to 3% of non-spam.

(a) (20 points) Assume that the occurrences of words in emails are independent for both spam and non-spam. If an email contains both the words "viagra" and "herbal" what is the probability that it is spam?

Solution: Use Bayes' Rule:

$$P(\text{spam}|\text{herbal} \cap \text{viagra}) = \frac{P(\text{herbal} \cap \text{viagra}|\text{Spam})P(\text{spam})}{P(\text{herbal} \cap \text{viagra})}$$

By the Law of Total Probability,

$$\begin{aligned} P(\text{herbal} \cap \text{viagra}) &= P(\text{herbal} \cap \text{viagra}|\text{spam})P(\text{spam}) \\ &\quad + P(\text{herbal} \cap \text{viagra}|\text{non-spam})P(\text{non-spam}) \\ &= 1/10 \times 1/20 \times 4/5 + 1/100 \times 3/100 \times 1/5 \\ &= 1/1000 \times (4 + 3/50) \end{aligned}$$

Hence,

$$P(\text{spam}|\text{herbal} \cap \text{viagra}) = \frac{4}{4 + 3/50} \approx 99\%$$

(b) (10 points) After completing your calculations, you learn an additional piece of information: approximately 14.5% of spam emails contain the word "herbal" *or* "viagra." Does this new information support or contradict the assumption that words appear independently in emails? Explain.

Solution: This question is only for spam emails, so we won't explicitly write the conditioning on spam. Let H be the event that a spam email contains at least one occurrence of "herbal" and V be the event that it contains at least one occurrence of "viagra." We are given that $P(V) = 0.1$, $P(H) = 0.05$ and $P(V \cup H) = 0.145$. By the addition rule $P(V \cup H) = P(V) + P(H) - P(V \cap H)$ Substituting the known probabilities, we see that $0.145 = 0.15 - P(V \cap H)$. Hence, $P(V \cap H) = 0.15 - 0.145 = 0.005$. Independence requires that $P(H \cap H) = P(H)P(V) = 0.1 \times 0.05 = 0.005$ so this piece of information tells us that, at least for spam emails, the words "herbal" and "viagra" occur independently.

Name: ____________________ Student ID #: ____________________