First Midterm Examination

Econ 103, Statistics for Economists

September 28th, 2016

You will have 90 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.

Please sign the back of your blue book.

Question:	1	2	3	4	5	6	7	Total
Points:	10	10	18	10	20	18	14	100
Score:								

Instructions: Answer all questions in your blue book. 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 in the back of your blue book. Make sure that you have all pages of the exam before starting.

Warning: If you continue writing after I call time, even if this is only to fill in your name, twenty-five points will be deducted from your final score. In addition, ten points will be deducted for not signing the back of your blue book.

Checklist before continuing:

☐ My name and Penn ID number are on my blue book

☐ I have signed the academic integrity statement in the back of the blue book

1. **True or False? If false, briefly explain why.**
   (a) [2 points] Nonsampling error will decrease as the sample size grows.
   (b) [2 points] Correlation implies causation.
   (c) [2 points] Statistics uses characteristics of the population to make predictions about the sample.
   (d) [2 points] Gender is an ordinal variable.
   (e) [2 points] Standard deviation is preferred to variance because it is unitless.

2. **Randomized trials**
   (a) [5 points] Describe a double-blind randomized controlled trial.
   (b) [5 points] Discuss why double-blind randomized controlled trials are superior to observational methods at determining effectiveness or causation.

3. **Summary Statistics:** Lake Wobegone is a "small rural town in central Minnesota where all the women are strong, all the men are good looking, and all the children are above average." Let's assume that Lake Wobegone has 100 children living there and children are scored on an integer scale ranging from 1 to 100.
   (a) [7 points] Is it possible that every child in Lake Wobegone is above the *national average*? Why or why not? If not, what is the maximum number of Wobegoner children that can be above the national average? Give an example supporting your solution.
   (b) [7 points] Is it possible that every child in Lake Wobegone is above the *average of children in Lake Wobegone*? Why or why not? If not, what is the maximum number of Wobegoner children that can be above the Lake Wobegone average? Give an example supporting your solution.
   (c) [4 points] Is it possible that every child in Lake Wobegone is above the *median of children in Lake Wobegone*? Why or why not? If not, what is the maximum number of Wobegoner children that can be above the Lake Wobegone median?

4. [10 points] **Regression:** Jay, an economist, is fitting a model of GDP growth and investment growth.
$$y_i = \beta * x_i^2 + e_i$$
where y is GDP growth and x is investment growth. Derive the formula for the OLS estimator β. Clearly state the optimization problem and then solve it.

5. **Presidential Election Predictions:** According to *The Washington Post*, Professor Allan Lichtman, a political historian at American University, has correctly predicted every presidential election since 1984 based on his "Keys to the White House" model.

Since the United States holds presidential elections every 4 years, that means he has correctly predicted 8 presidential elections.

(a) [2 points] If Professor Lichtman had been randomly guessing the winner of presidential elections, what is the probability that he would have gotten 8 right in a row? Assume there are only two candidates in each election and he flips a fair coin to decide his guess for the winner.

(b) [8 points] Across the nation, professors try to predict presidential outcomes. Let's say that 100 professors nationwide are trying to predict presidential elections. What is the chance that at least one professor has called the past eight elections correctly, assuming each adopts the same extensively-researched coin-flipping forecasting strategy? Computing the answer is not required for full credit, but your answer should be an expression that could then be evaluated on a calculator.

(c) [5 points] There are likely more than 100 professors predicting presidential elections (not counting private forecasting companies and individuals). If there are n professors trying to predict elections, what is the expression for how we would calculate the chance that at least one is right at predicting the past eight elections?

(d) [5 points] In reality, people do not use coin flips to make their predictions (or at least they do not admit to it). We have n people across the country making predictions for the election. Assuming each model has probability p of predicting each election correctly and each model is independent of the others, what is the probability that at least one model has predicted the past 8 elections correctly?

6. **R Programming and Statistical Analysis:** Sandy, an economist, is researching the gender-wage gap.

(a) [3 points] Before starting her analysis, she wants to make sure R is working properly. She types the following into R. What does she expect R to return?

```
x <- 2
y <- 1
z <- x * y
z + 2 == 4
```

(b) [3 points] She downloads the data and stores it in a data table called `wageData`. She then types `head(wageData)` and the following is displayed:

```
   Gender Wage Age
1:      F   20  25
2:      M   18  26
3:      F   10  27
4:      F   NA  28
```

```
5:      M   15  29
6:      F   18  30
```

She then types `wageData[, mean(Wage, na.rm = FALSE), by = Gender]`. What is she trying to compute?

(c) [3 points] What will R return and why? How should she modify her code to fix this?

(d) [2 points] She then runs `lm(data = wageData, Wage ~ Gender)`. What does the `lm()` command do? If you do not remember what this command does, I will also accept the R code that would pull up the documentation.

(e) [5 points] Sandy wants to estimate the following regression $Wage_i = a + b \cdot Age_i$. In this regression how are b and the correlation related? Do not simply state the formula. Start with one of the equations for b and demonstrate how to rearrange it to include the correlation.

(f) [2 points] She runs the following commands in R

```
> cov(wageData$Age, wageData$Wage, use = "complete.obs")
[1] -1.6
> cor(wageData$Age, wageData$Wage, use = "complete.obs")
[1] -0.1979083
> var(wageData$Age, na.rm = TRUE)
[1] 3.5
> var(wageData$Wage, na.rm = TRUE)
[1] 15.2
> mean(wageData$Age, na.rm = TRUE)
[1] 27.5
> mean(wageData$Wage, na.rm = TRUE)
[1] 16.2
```

If she were to estimate the following regression $Wage_i = a + b \cdot Age_i$, what would her coefficients be? You do not have to calculate the exact numbers, but you should have an expression that can be computed using a calculator.

7. **Probability:**

(a) [2 points] State the mathematical definition of conditional probability

(b) [2 points] State Bayes' rule (mathematically).

(c) [10 points] In "Silver Blaze," Sherlock Holmes is investigating the mysterious disappearance of a race horse and the murder of its trainer. Near the stables, there is a watchdog which will bark whenever a stranger approaches. However, at night, the dog is only 80% accurate. That is, if a stranger approaches, 80% of the time it

will bark and 20% of the time it will stay silent. If a familiar person approaches, the dog will stay silent 80% of the time and bark 20% of the time. Let's assume that 90% of the suspects are familiar to the dog and 10% are strangers.

One of the most famous Sherlock Holmes exchanges is below:

Gregory (Scotland Yard Detective): Is there any other point to which you wish to draw my attention?
Holmes: To the curious incident of the dog in the night-time.
Gregory: The dog did nothing in the night-time.
Holmes: That was the curious incident.

If a stranger had committed this crime, was this actually a curious incident? Compute the corresponding probability that a stranger committed the crime using the information that the dog did not bark.