First Midterm Examination

Econ 103, Statistics for Economists

October 3, 2012

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	10	11	Total
Points:	10	15	10	10	20	10	10	10	10	20	15	140
Score:												

Instructions: Answer all questions in the space provided. Should you run out of space, continue on the back of the page. 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.

1. A survey collected the following information on new college graduates. Classify each variable as nominal, ordinal, or numerical. You do not need to explain your answers.

   (a) (2 points) Sex

   Solution: Nominal

   (b) (2 points) Age at graduation

   Solution: Numerical

   (c) (2 points) Future plans (graduate school, employment, indefinite)

   Solution: Nominal

   (d) (2 points) Class rank

   Solution: Ordinal

   (e) (2 points) Grade point average

   Solution: Numerical

2. A long time ago, the graduate school at a famous university admitted 4000 of their 8000 male applicants versus 1500 of their 4500 female applicants.

   (a) (3 points) Calculate the difference in admission rates between men and women. What does your calculation suggest?

   Solution: The rate for men is $4000/8000 = 50\%$ while that for women is $1500/4500 \approx 33\%$ so the difference is 17%. It appears that women are less likely to be accepted to the graduate school.

   (b) (7 points) To get a better sense of the situation, some researchers broke these data down by area of study. Here is what they found:

	Men		Women	
	# Applicants	# Admitted	# Applicants	# Admitted
Arts	2000	400	3600	900
Sciences	6000	3600	900	600
Totals	8000	4000	4500	1500

   Calculate the difference in admissions rates for men and women studying Arts. Do the same for Sciences.

Name: ____________ Student ID #: ____________

Solution: For Arts, the admission rate is 400/2000 = 20% for men versus 900/3600 = 25% for women. For Sciences 3600/6000 = 60% for men versus 600/900 ≈ 67% for women. In summary:

	Men	Women	Difference
Arts	20%	25%	-5%
Sciences	60%	67%	-7%
Overall	50%	33%	17%

(c) (5 points) Compare your results from part (a) to part (b). Explain the discrepancy using what you know about observational studies.

Solution: When we compare overall rates, women are less likely to be admitted than men. In each field of study, however, women are *more* likely to be admitted. In this example, field of study is a *confounder*: women are disproportionately applying to study Arts and Arts have much lower admissions rates than Sciences.

3. The 10-90 percentile range is a measure of dispersion that we did not cover in class. It is given by the difference $P_{90} - P_{10}$ where P_{10} is the 10th and P_{90} the 90th percentile of the sample.

(a) (5 points) Would you expect the 10-90 percentile range to be more or less sensitive to outliers than the range? Explain.

Solution: It will be less sensitive since it "trims off" the top and bottom 10% of the dataset, i.e. the most extreme observations. In contrast, the range is calculated directly from the maximum and the minimum and so is extremely sensitive to outliers.

(b) (5 points) Consider these three measures of dispersion: the 10-90 percentile range, the interquartile range, and the range. Which is the largest, and which is the smallest?

Solution: IQR $\leq (P_{90} - P_{10}) \leq$ Range

4. Let a be the intercept and b be the slope in the linear regression $y_i = a + bx_i + e_i$ where y is earnings in dollars and x is years of education.

(a) (2 points) Write down the formula for calculating b from the data.

Solution:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

(b) (2 points) Write down the formula for calculating a from the data and b.

Solution:

$$a = \bar{y} - b\bar{x}$$

(c) (3 points) What are the units of b? Explain briefly.

Solution: The units of s_{xy} are (dollars × years) while the units of s_x^2 are years^2 so that the units of b are dollars/year. Since y is measured in dollars, b has to be measured in dollars/year for the units to work out: multiplying x by b converts years into dollars.

(d) (3 points) What are the units of a? Explain briefly.

Solution: By the formula above, a must be in the same units as $\bar{y}$, dollars.

5. Suppose we want to estimate a linear, least squares regression with a set equal to zero, that is $y_i = bx_i + e_i$.

(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:

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

(c) (5 points) Suppose that we wanted to predict y from x using a *quadratic* relationship, namely $y_i = bx_i^2 + e_i$. 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.

6. (10 points) Major League Baseball honors its most outstanding first-year player with the title "Rookie of the Year." The overall major league batting average is around .260, compared to .290 for Rookies of the Year. In their second year, however, Rookies of the Year have an overall batting average of only .275 – a comparatively disappointing performance. This is sometimes called the "sophomore slump." It has been argued that the publicity, endorsement deals, and tv appearances that come from being named Rookie of the Year distract players from the game, and that this explains the "sophomore slump." Do you agree? Answer yes or no and explain briefly.

Solution: No. This is an example of regression to the mean. Someone who wins Rookie of the Year is both talented and lucky: the correlation between first and second-year performance is less than one. You will remain talented after your first year, but it is unlikely that you will be lucky twice in a row. Our best guess of their performance in the second year is closer to the mean than their performance in the first year.

7. Suppose A and B are events, i.e. $A, B \subseteq S$ where S is the sample space.

(a) (2 points) Write down the three axioms (aka postulates) of probability from class.

Solution:

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$

(b) (2 points) Are A and A^c collectively exhaustive?

Solution: Yes

(c) (2 points) Are A and A^c mutually exclusive?

Solution: Yes

(d) (4 points) Using parts (a)–(c), derive the Complement Rule.

Solution: By (a), the union of A and A^c is S, so $P(A \cup A^c) = P(S) = 1$. By (b), A and A^c are mutually exclusive, hence $P(A \cup A^c) = P(A) + P(A^c)$. Combining these, we have $P(A) + P(A^c) = 1$. Finally, rearranging, $P(A^c) = 1 - P(A)$.

8. Three percent of *Tropicana* brand oranges are already rotten when they arrive at the supermarket. In contrast, six percent of *Sunkist* brand oranges arrive rotten. A local supermarket buys forty percent of its oranges from *Tropicana* and the rest from *Sunkist*. Let R be the event that an orange is rotten and T be the event than it is a *Tropicana*.

(a) (3 points) What is the probability that a randomly chosen orange in the supermarket is a *Tropicana* and is rotten?

Solution: By the multiplication rule:

$$P(R \cap T) = P(R|T)P(T) = 0.03 \times 0.4 = 0.012$$

(b) (3 points) What is the probability that a randomly chosen orange is rotten?

Solution: By the law of total probability:

$$\begin{aligned} P(R) &= P(R|T)P(T) + P(R|T^c)P(T^c) \\ &= 0.03 \times 0.4 + 0.06 \times 0.6 \\ &= 0.012 + 0.036 \\ &= 0.048 \end{aligned}$$

(c) (4 points) Suppose we randomly choose an orange from the supermarket and see that it is rotten. What is the probability that it is a *Tropicana*?

Solution: By Bayes' Rule:

$$\begin{aligned} P(T|R) &= \frac{P(R|T)P(T)}{P(R)} \\ &= \frac{0.012}{0.048} = 1/4 = 0.25 \end{aligned}$$

9. Suppose a couple decides to have three children. Assume that the sex of each child is independent, and the probability of a girl is 0.48. (Both of these assumptions are approximately correct for US data.) Note that order matters in this example.

(a) (3 points) How many basic outcomes are there for this experiment?

Solution: There are two possible outcomes for each birth, so by the multiplication rule for counting, the total number of possibilities is $2 \times 2 \times 2 = 8$.

(b) (3 points) Are the basic outcomes equally likely? Why or why not?

Solution: No, because each child is more likely to be a boy than a girl. The outcome BBB is most likely, followed by outcomes with two boys, and then outcomes with one boy. The outcome GGG is least likely.

(c) (4 points) What is the probability that the couple has *at least one* girl?

Solution: Use the Complement Rule and independence to calculate the probability of no girls, i.e. all boys:

$$0.52 \times 0.52 \times 0.52 \approx 0.14$$

Hence, the probability of at least one girl is approximately $1 - 0.14 = 0.86$

10. You have been entered into a very strange tennis tournament. To get the \$10,000 Grand Prize you must win at least two sets *in a row* in a three-set series to be played against your Econ 103 professor and Venus Williams alternately: professor-Venus-professor or Venus-professor-Venus according to your choice. Let p be the probability that you win a set against your professor and v be the probability that you win a set against Venus. Naturally $p > v$ since Venus is much better than your professor! Assume that each set is independent.

(a) (5 points) Let W indicate win and L indicate lose, so that the sequence WWW means you win all three sets, WLW means you win the first and third set but lose the middle one, and so on. Which sequences of wins and losses land you the Grand Prize?

Solution: To get the prize, you have to win the middle set. Thus, the only possibilities are WWW, WWL, and LWW.

(b) (5 points) If you elect to play the middle set against Venus, what is the probability that you win the Grand Prize?

Solution: The probabilities of mutually exclusive events sum. Thus,

$$\begin{aligned} P(WWW) + P(LWW) + P(WWL) &= pvp + (1-p)vp + pv(1-p) \\ &= p^2v + pv - p^2v + pv - p^2v \\ &= 2pv - p^2v \\ &= pv(2-p) \end{aligned}$$

(c) (5 points) If you elect to play the middle set against your professor, what is the probability that you win the Grand prize?

Solution: Again, the probabilities of mutually exclusive events sum. Thus,

$$\begin{aligned} P(WWW) + P(LWW) + P(WWL) &= vpv + (1-v)pv + vp(1-v) \\ &= v^2p + vp - v^2p + vp - v^2p \\ &= 2pv - v^2p \\ &= pv(2-v) \end{aligned}$$

(d) (5 points) To maximize your chance of winning the prize, should you choose to play the middle set against Venus or your professor?

Solution: Manipulating the inequality,

$$\begin{aligned} p &> v \\ -p &< -v \\ 2-p &< 2-v \\ pv(2-p) &< pv(2-v) \end{aligned}$$

You can't get the prize without winning the middle set, so it turns out that it's better to face Venus twice rather than face her in the middle set. You should elect to play the middle set against your professor.

11. This question refers to commands from the R statistical package that you used in the homework. Suppose I have a dataframe called `gradebook` containing the following columns in order from left to right: `student.name`, `major`, and `GPA`. (That is, for each student, the dataframe contains the student's name, major and grade point average.)

(a) (3 points) Suppose I want to display the information from all columns for the student in row ten of the dataframe. What command should I use?

Solution: `gradebook[10,]`

(b) (3 points) Suppose I want to calculate the mean of the GPA column, which contains numeric data. What command should I use? (You may assume that there are no missing values in the dataframe.)

Solution: `mean(gradebook$GPA)`

(c) (3 points) When I enter the command `class(gradebook$major)`, I get `"factor"` as my result. What does this mean?

Solution: `major` is a categorical variable

(d) (3 points) Suppose I want to display names and grade point averages *only* for the students in rows two, seven, and thirteen of the dataframe. What command should I use?

Solution: `gradebook[c(2, 7, 13), c("student.name", "GPA")]` works, as does `gradebook[c(2, 7, 13), -2]`

(e) (3 points) What is the name of the command for calculating percentiles in R?

Solution: `quantile`

Name: ______________________ Student ID #: ______________________