Final Examination

Econ 103, Statistics for Economists

December 12, 2012

You will have 120 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	12	Total
Points:	10	10	15	25	15	15	25	20	25	30	15	35	240
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. (10 points) In an effort to reduce traffic fatalities, local authorities installed traffic cameras in particularly accident-prone locations in west London in 1992. In the preceding three years, there were 62 fatal accidents in these locations. In the three years following the introduction of the cameras, there were only 19 fatal accidents. Based on this evidence, local authorities concluded that traffic cameras reduce fatalities. Traffic cameras have since been installed in the whole of London. Using the information provided in this question *only*, do we have good evidence that traffic cameras reduce the frequency of fatal accidents? Explain why or why not.

2. (10 points) Of women who undergo regular mammograms, two percent have breast cancer. If a woman has breast cancer, there is a 90% chance that her mammogram will come back positive. If she does *not* have breast cancer there is a 10% chance that her mammogram will come back positive. Given that a woman's mammogram has come back positive, what is the probability that she has breast cancer?

3. (15 points) Explain the difference between μ, $\bar{X}_n$ and $\bar{x}$ using the concepts and terminology we have studied in this course.

4. Let X be a Uniform$(0, 1)$ random variable.

   (a) (5 points) Write down the probability density function (pdf) and support of X.

   (b) (5 points) Calculate the cumulative distribution function (cdf) of X.

   (c) (5 points) Calculate $\mathbb{E}[X]$

   (d) (5 points) Calculate $\mathbb{E}[X^2]$

(e) (5 points) Using the shortcut formula, calculate $Var(X)$.

5. (15 points) About as many boys as girls are born in hospitals. Many babies are born every week at City General. In Cornwall, a country town, there is a small hospital where only a few babies are born every week. A *normal week* is one where between 45% and 55% of the babies are female. An *unusual week* is one where more than 55% are girls, or more than 55% are boys. Do unusual weeks occur more often at City General or Cornwall? Explain your answer using what you learned about sampling distributions in this course.

6. This question concerns the quantity $\tilde{p}$ used to construct the "refined" confidence interval for a population proportion that I introduced in class. Recall that $\tilde{p} = (n\widehat{p}+2)/(n+4)$, where $\widehat{p} = \left(\sum_{i=1}^{n} X_i\right)/n$ and $X_1, \ldots, X_n \sim$ iid Bernoulli(p).

(a) (5 points) Is $\tilde{p}$ an unbiased estimator of p? If so, prove it. If not, calculate its bias.

(b) (5 points) Which has a higher variance: $\widehat{p}$ or $\tilde{p}$? Prove your answer.

(c) (5 points) Using your answers to (a) and (b), calculate the mean-squared error of $\tilde{p}$. Is $\tilde{p}$ a consistent estimator of p? Prove your answer.

7. Suppose that $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. Let $\bar{X}_n$ be the sample mean and S^2 be the sample variance. Except in part (d), you do not need to provide any justification for your answers.

(a) (5 points) What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$?

(b) (5 points) What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/S$?

(c) (5 points) What is the sampling distribution of $(n-1)S^2/\sigma^2$?

(d) (10 points) Using your answer to part (c), derive a $100 \times (1-\alpha)$ confidence interval for the variance of a normal population based on a random sample of n observations. Express the interval in terms of the appropriate R commands.

8. Suppose that $X_1, \ldots, X_n \sim$ iid Bernoulli(p) and define $\widehat{p} = \sum_{i=1}^{n} X_i/n$.

(a) (5 points) Calculate the standard deviation of the sampling distribution of $\widehat{p}$.

(b) (5 points) There is a technical term for the standard deviation of a sampling distribution. What is it?

(c) (5 points) What value of p maximizes the standard deviation of the sampling distribution of $\widehat{p}$? What is the maximized standard deviation? Be sure to check the second order condition.

(d) (5 points) Sarah is a graduate student in the political science department and is preparing to carry out a poll designed to estimate the proportion of college students who support gay marriage. Sarah plans to construct a 95% confidence interval based on her results and wants to gather enough data so that the margin of error of her poll is no more than 4%. Based on your answer to part (c) and the *textbook confidence interval* for a population proportion, what is the smallest sample size that Sarah should consider using in her poll?

9. This question concerns the following line of R code:

```
sims <- replicate(10000, quantile(rnorm(10), 0.5))
```

(a) (5 points) Explain what this command does. What is the point of running it?

(b) (5 points) After running this command I entered `mean(sims)` in the R console and got the result `-0.006093938`. What does this suggest about the sampling distribution of the sample median?

(c) (5 points) Next, I entered `hist(sims)` into the console. The resulting graph appears in Figure 1. Based on this histogram, what would you conjecture about the sampling distribution of the sample median?

(d) (5 points) Now suppose I were to run the following command

```
sims2 <- replicate(10000, mean(rnorm(10)))
```

If I entered `mean(sims2)` into the console, approximately what value would you expect R to return? What if entered `var(sims2)`?

(e) (5 points) Finally, I entered `var(sims)` at the console and got a result of `0.141154`. Based on this result and your answers to parts (b) and (d), which would you recommend as an estimator of the median of a normal population: the sample median or the sample mean?

b

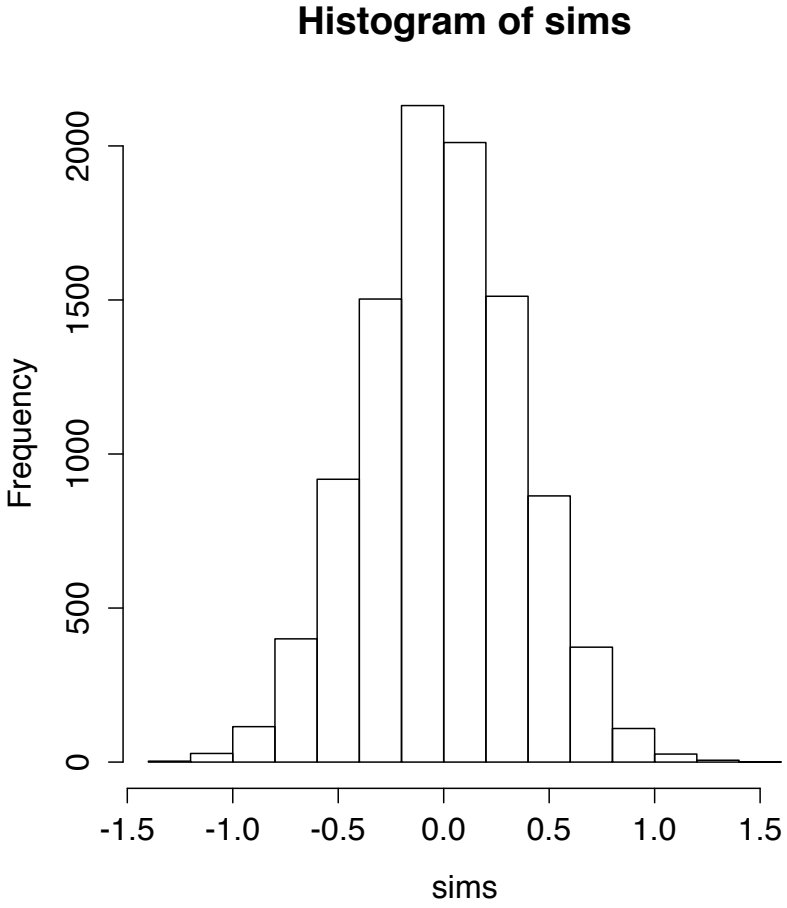

Figure 1: Histogram for Question 9

Name: _____ Student ID #: _____

10. This question concerns the "Pepsi Challenge" experiment. The setup is identical to the one we used in class except with five cups of each soda rather than four. The experimental procedure is as follows. We first fill five cups with Pepsi and five with Coke. Then we randomize the order of the cups and allow our expert to taste each. Finally, we ask the expert to identify which five cups contain Coke and record the number of cups that she correctly identifies. The idea is to determine whether our expert can really tell the difference between Coke and Pepsi based on the results of the experiment. This question asks to you formalize the experiment as a statistical hypothesis test. Our test statistic T is the number of Cokes that our expert correctly identified.

    (a) (5 points) What is our null hypothesis, H_0, in this experiment?

    (b) (5 points) What is our alternative hypothesis, H_1, in this experiment?

    (c) (5 points) Under H_0 what is the probability that $T = 5$?

    (d) (5 points) Under H_0 what is the probability that $T = 4$?

(e) (5 points) Suppose I wanted to test H_0 at the 5% significance level. What is my decision rule given the alternative you specified?

(f) (5 points) Suppose our expert identified four Cokes correctly. What is the p-value for our test against the alternative you specified?

11. Suppose there is a logarithmic relationship between random variables X and Y, namely $\ln Y_i = \beta X_i + \epsilon_i$ where $\epsilon_i \sim \text{iid}(0, \sigma^2)$ is an unobserved error term.

(a) (5 points) Suppose we want to estimate β by least squares regression using a random sample of observations $(y_1, x_1), \ldots, (y_n, x_n)$. What optimization problem should we solve?

(b) (10 points) Solve the problem from part (a) to derive an explicit formula for $\widehat{\beta}$. You do not need to check the second order condition.

Table 1: Regression Results

Regression 1:

```
lm(formula = midterm2 ~ male)
            coef.est coef.se
(Intercept) 81.82     2.12
male        -0.22     2.91
---
n = 70, k = 2
residual sd = 12.17, R-Squared = 0.00
```

Regression 2:

```
lm(formula = midterm2 ~ midterm1)
            coef.est coef.se
(Intercept) 34.63     9.32
midterm1     0.59     0.12
---
n = 70, k = 2
residual sd = 10.35, R-Squared = 0.28
```

Regression 3:

```
lm(formula = midterm2 ~ male + midterm1)
            coef.est coef.se
(Intercept) 34.79     9.47
male        -0.31     2.50
midterm1     0.59     0.12
---
n = 70, k = 3
residual sd = 10.43, R-Squared = 0.28
```

Regression 4:

```
lm(formula = midterm2 ~ male + male:midterm1 + midterm1)
              coef.est coef.se
(Intercept)   19.31    14.22
male          26.76    18.82
midterm1       0.78     0.18
male:midterm1 -0.34     0.23
---
n = 70, k = 4
residual sd = 10.34, R-Squared = 0.30
```

12. This question is based on an dataset containing observations on students in Econ 103: `male` takes the value 1 if a given student is male, zero otherwise; `midterm2` gives that student's score on the second midterm; and `midterm1` gives the student's score on the first midterm. Using this dataset, I estimated four regression models using R. The results appear in Table 1.

    (a) (5 points) Suppose I wanted to test the null hypothesis that men and women do just as well, on average, on the second midterm of Econ 103 against the two-sided alternative using a 5% significance level. I can carry out this test directly from the results present above. Which results should I use and how should I carry out the test? In particular, what is the appropriate test statistic, what is the appropriate critical value, and what is the outcome of the test?

    (b) (5 points) Explain the difference between the models used in Regression 2, versus 3 and 4. Do not comment on the results, only the *models themselves*.

(c) (5 points) Based on the above regression results, what is the sample correlation between students' scores on the two midterms?

(d) (5 points) Explain the meaning of the coefficient estimate `midterm1` in Regression 2 and construct a 95% confidence interval for this parameter. What would you conclude on the basis of this interval?

(e) (5 points) Instead of constructing a confidence interval, suppose I wanted to test the null hypothesis that the coefficient on `midterm1` in Regression 2 is zero against the two-sided alternative. What is my test statistic? Would I reject the null at the 5% level? What R command should I enter to find the p-value for this test?

(f) (10 points) Suppose we want to predict a student's score on the second midterm. Based on the results given above, do you think we should take into account that student's sex? If so, how should we use this information? Explain your reasoning.