Final Examination

Econ 103, Statistics for Economists

May 1, 2013

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	Total
Points:	10	20	10	25	15	20	30	10	30	30	200
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.

You may use the approximation `qnorm(0.975)` ≈ 2 to simplify your calculations on this exam.

1. You are about to bid on a poster autographed by the cast of *The Jersey Shore*. This is a one-of-a-kind item, not available in stores. The most you would be willing to pay is \$500 dollars. In an ordinary auction, you could start with a low bid, say \$10, and increase it only if someone outbid you. This auction, however, is a *sealed bid auction*. Each participant submits a single bid in a sealed envelope, so there is no way for you to know how much anyone else has bid. The highest bidder wins the auction and pays whatever bid she placed in her envelope; the losers pay nothing. Based on research you have carried out on other auctions of Jersey Shore memorabilia, you estimate that the probability p that you will win the auction as a function of your bid b is as follows:

$$p(b) = \begin{cases} b/600 & b \leq 600 \\ 1 & b > 600 \end{cases}$$

   (a) (6 points) Suppose you want to maximize your *expected payoff*. Losing the auction gives you a payoff of \$0; winning gives you a payoff of \$500 *minus your bid*, i.e. your consumer surplus. How much should you bid, and what is the probability that you will win the auction? Be sure to check the second order condition.

   (b) (4 points) Using what you know about expected value, briefly explain why someone might rationally decide *not* to follow the bidding strategy you derived in part (a).

Name: _____ Student ID #: _____

2. Suppose that X is a continuous random variable with probability density function

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) (2 points) What is the support of X?

(b) (3 points) Given your answer to (a), why *couldn't* the pdf be $4x^2$ in this example?

(c) (6 points) Calculate the cumulative distribution function of X.

(d) (3 points) Calculate $E[X]$.

(e) (3 points) Calculate $E[X^2]$.

(f) (3 points) Calculate $Var(X)$.

3. Let X and Y be two RVs with $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$ and $Corr(X, Y) = \rho$.
   (a) (2 points) Write down an expression for $Var(X + Y)$ in terms of σ_X, σ_Y and ρ.

   (b) (2 points) Write down an expression for $Var(X - Y)$ in terms of σ_X, σ_Y and ρ.

   (c) (6 points) Which is larger: $Var(X + Y)$ or $Var(X - Y)$? Explain briefly.

4. Suppose that $X_1 \sim N(\mu, \sigma^2)$ independently of $X_2 \sim N(\mu, 3\sigma^2)$. Let $\bar{X} = (X_1 + X_2)/2$.
   (a) (4 points) Calculate the variance of $\bar{X}$.

   (b) (4 points) Let $\tilde{\mu} = \omega X_1 + (1 - \omega)X_2$ where $\omega \in [0, 1]$. Is $\tilde{\mu}$ an unbiased estimator of μ? Prove your answer.

   (c) (5 points) Define $\tilde{\mu}$ as in part (b). Calculate the variance of $\tilde{\mu}$.

(d) (8 points) What value of ω minimizes $Var(\tilde{\mu})$? What is the minimum achievable variance? Be sure to check the second order condition.

(e) (4 points) Is the sample mean an efficient estimator of μ in this example? Explain.

5. This question asks you to supply R code to carry out a simple simulation experiment using what you learned in recitation.

(a) (5 points) The R command `runif(n)` returns a random sample of `n` Uniform$(0, 1)$ observations. Using this information, write an R function called `unif.sim` that takes a single argument `n` and carries out the following: (i) generate a sample of `n` iid Uniform$(0, 1)$ observations, (ii) return the sample mean of these observations.

(b) (5 points) Suppose you wanted to run your function from part (a) 10,000 times, each with a sample size of 10, and store the result in a vector called `sims`. How could you implement this in R?

(c) (5 points) Suppose you entered the command `mean(sims)` after carrying out (b). Approximately what value would you get? Why?

6. I wanted to find out how many fish are in the lake, so I sent Garth and Naijia out to catch a random sample of 100 fish, tag them, and then release them. A week later, I sent Garth and Naijia back to the lake to catch another random sample of 100 fish. Of the 100 fish they caught the second time, 20 had tags.

   (a) (5 points) Construct an approximate 95% confidence interval for the proportion of fish in the lake that have tags. Use the textbook CI to keep the calculations simple.

(b) (3 points) Using the fact that we know that exactly 100 fish in the lake have been tagged, what is our estimate of the *total number of fish in the lake*?

(c) (8 points) Using your answers from above, construct an approximate 95% confidence interval for the *total number of fish in the lake*.

(d) (4 points) Does the CI from part (c) take the form Estimate ± ME? Explain.

7. Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and suppose we know that $\sigma^2 = 1$.

(a) (2 points) Write down the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)$.

(b) (6 points) Using your answer to (a), derive the sampling distribution of $\sqrt{n}\bar{X}_n$.

(c) (6 points) Suppose we wanted to test the null hypothesis that $\mu = 0$ against the two-sided alternative at the 5% level. What test statistic should we use and what should our decision rule be?

(d) (6 points) How would your answer to (c) change if we tested instead against the one-sided alternative that $\mu > 0$?

(e) (10 points) Derive the power of the hypothesis test from (c) if $\sigma = 1$, $n = 25$ and $\mu = 1$. Your answer should be given in terms of the relevant R commands.

8. (10 points) A 1981 study published in the *New England Journal of Medicine* looked at the use of cigars, pipes, cigarettes, alcohol, tea, and coffee by patients with pancreatic cancer, concluding that there was "a strong association between coffee consumption and pancreatic cancer." This study was immediately criticized for carrying out multiple tests but only reporting the most statistically significant result. Subsequent studies failed to confirm an association between coffee drinking and pancreatic cancer. Suppose that six independent tests are conducted, in each case involving a product that is, in fact, unrelated to pancreatic cancer. What is the probability that at least one of these tests will find an association that is statistically significant at the 5% level?

9. This question is based on a recent paper examining how "organic" labeling changes people's perceptions of different food products. Researchers recruited volunteers at a local mall in Ithaca, New York and gave each two samples of yogurt to taste. Although both yogurts were in fact identical, the volunteers were *told* that one of them was organic while the other was not. After tasting both, each volunteer was asked to estimate how many calories each of the samples of yogurt contained. (Since, unknown to the volunteer, both samples contained exactly the same kind of yogurt, each in fact contained the same number of calories.) To prevent confounding from anchoring or other behavioral effects, the order in which a given volunteer tasted the two yogurts, i.e. "organic" first or "organic" second, was chosen at random. The results of this experiment are stored in an R dataframe called `yogurt`. Here are the first few rows:

```
> head(yogurt)
  regular organic
1      60       40
2       5        0
3     200      100
4      60       40
5     100      100
6      90       90
```

Each row in this dataframe corresponds to a single individual's guess of the number of calories contained in each of the two yogurts. For example, the values 60 and 40 in row 1 mean that volunteer number one guessed that the regular yogurt sample contained 60

calories and the organic sample contained 40. Summary statistics for the two columns are as follows:

	`regular`	`organic`
Sample Mean	113	90
Sample Var	3600	2916
Sample SD	60	54
Sample Corr.	0.8	
Sample Size	115	

(a) (4 points) Give the units of each of the summary statistics from above:

Sample Mean _____
Sample Var. _____
Sample SD _____
Sample Corr. _____

(b) (4 points) Sara thinks that this experiment should be analyzed as independent samples data. Assume that she is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

(c) (6 points) Kevin thinks that this experiment should be analyzed as matched pairs data. Assume that he is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

(d) (6 points) How do the confidence intervals constructed by Sara and Kevin differ? Explain the source of the discrepancy. Which of them has constructed the appropriate confidence interval for this example?

(e) (6 points) Suppose that Kevin wanted to carry out a two-sided test of the null hypothesis that organic labeling does not affect consumer's estimates of caloric content, on average. What is his test statistic? What R command should he use to calculate the p-value for his test? Will his result be greater or less than 0.05?

(f) (4 points) Using what you know about experiments, observational studies, hypothesis testing, and confidence intervals, what conclusions can we draw from this study?

10. This question refers to the four sets of regression results presented in Table 1, which appears on the final page of this exam. Each regression uses a data frame called `income.data`, the first few rows of which are as follows:

```
 head(income.data)
  income female   AFQT
1    5.5      1  6.841
2   65.0      0 99.393
3   19.0      0 47.412
4   36.0      1 44.022
5   65.0      0 59.683
6    8.0      0 72.313
```

Each row corresponds to an individual. The column `income` gives that individual's income (in thousands of dollars) in 2005. The column `female` is a dummy variable that takes the value 1 if a given individual is female. Finally, the column `AFQT` gives the individual's percentile score on the Armed Forces Qualifying Test.

(a) (3 points) Interpret the intercept in Regression 1.

(b) (3 points) Interpret the coefficient `AFQT` in Regression 1.

(c) (2 points) What is the sample correlation between income and percentile score on the AFQT?

(d) (2 points) Using the regression results, what is the average income of the men in the sample? What is the average income of the women in the sample?

Name: _____ Student ID #: _____

(e) (5 points) Using sex *alone* as a predictor, about how accurately can we predict an individual's income? How does this compare to using AFQT score alone?

(f) (5 points) Using the regression results, construct an approximate 95% confidence interval for the difference of mean income in the population (men - women). Interpret your results.

(g) (5 points) Regressions 2–4 each examine the relationship between sex and income. How do the regression models used in each differ? You don't need to discuss the results, just the models.

(h) (5 points) Is there evidence of a *different* relationship between intelligence as measured by AFQT and income for men and women? If so, explain the difference.

Table 1: Regression Results

Regression 1:

```
lm(formula = income ~ AFQT)
            coef.est coef.se
(Intercept) 21.18     1.93
AFQT         0.52     0.03
---
n = 2584, k = 2
residual sd = 44.46, R-Squared = 0.09
```

Regression 2:

```
lm(formula = income ~ female)
            coef.est coef.se
(Intercept)  63.32     1.23
female      -28.11     1.75
---
n = 2584, k = 2
residual sd = 44.57, R-Squared = 0.09
```

Regression 3:

```
lm(formula = income ~ female + AFQT)
            coef.est coef.se
(Intercept)  35.55     2.04
female      -27.09     1.67
AFQT          0.50     0.03
---
n = 2584, k = 3
residual sd = 42.36, R-Squared = 0.18
```

Regression 4:

```
lm(formula = income ~ female + AFQT + female:AFQT)
            coef.est coef.se
(Intercept)  27.47     2.55
female      -10.03     3.66
AFQT          0.65     0.04
female:AFQT  -0.31     0.06
---
n = 2584, k = 4
residual sd = 42.14, R-Squared = 0.19
```