Final Examination

Econ 103, Statistics for Economists

May 14th, 2019

You will have two hours 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: ______________________________

Signature: ______________________________

Student ID #: ______________ Recitation #: ______________

Question:	1	2	3	4	5	6	Total
Points:	60	25	25	20	20	50	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, a point will be deducted for each page on which you do not write your name and student ID.

1. Answer each of the following. For full credit, explain your answers clearly and succinctly.

$\boxed{10}$ (a) 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*. Suppose we randomly choose an orange from the supermarket and see that it is rotten. What is the probability that it is a *Tropicana*? In your answer, let R be the event that an orange is rotten and T be the event that it is a *Tropicana*.

$\boxed{7}$ (b) Let $X_1, X_2 \sim$ iid with mean μ and variance σ^2. Is $(0.1X_1 + 0.9X_2)$ is a more efficient estimator of μ than $(0.5X_1 + 0.5X_2)$?

$\boxed{7}$ (c) Let X and Y be RVs with $Var(X) = 2$, $Var(Y) = 1$, and $Cov(X, Y) = 0$. Calculate $Var(X - Y)$.

$\boxed{7}$ (d) Suppose that Alice and Bob each draw independent random samples of size $n = 100$ from a normal population with unknown mean μ and known variance $\sigma^2 = 9$. They both construct 95% confidence intervals for μ. Will the widths of Alice and Bob's intervals be the same? Will Alice and Bob's intervals be identical?

$\boxed{7}$ (e) In the "Pepsi Challenge" experiment from class there were four cups of Coke and four of Pepsi. In this question, consider a modified version of the experiment with *three* cups of each kind of soda. Everything else is unchanged. Calculate the probability that our test statistic, the number of cokes correctly identified, will equal two *under the null hypothesis*.

$\boxed{7}$ (f) Alice constructs a 95% CI for μ: $[-0.5, 0.3]$. Bob tests $H_0\colon \mu = 0$ vs. $H_1\colon \mu \neq 0$ with $\alpha = 0.01$ using the same dataset as Alice. Will he reject H_0?

Name: ____________________ Student ID #: ____________________

$\boxed{7}$ (g) Suppose that $X_1, \ldots, X_5 \sim$ iid $\text{N}(1, 4)$ independently of $Y_1, \ldots, Y_{20} \sim$ iid $\text{N}(-1, 24)$. Write a line of R code to calculate $P(\bar{X} - \bar{Y} > 0)$.

$\boxed{8}$ (h) The Fibonacci sequence is defined as follows: $F_1 = 1, F_2 = 1$, and $F_i = F_{i-1} + F_{i-2}$ for $i \geq 3$. In other words: $1, 1, 2, 3, 5, 8, 13, 21, 34, 55 \ldots$ and so on. Write R code to calculate the first 20 terms of the Fibonacci sequence $(F_1, F_2, \ldots, F_{20})$ and store them in a vector called `fib`.

2. Let X, Y, Z be iid discrete RVs with support set $\{-1, 1\}$ and probability mass function $p(-1) = 1 - p$, $p(1) = p$. Define $S = X + Y + Z$.

$\boxed{5}$ (a) Calculate $E[X]$.

$\boxed{5}$ (b) Calculate the variance of Z. How does it compare to that of a Bernoulli(p) RV?

$\boxed{5}$ (c) Calculate $Var(S)$.

$\boxed{10}$ (d) Calculate $P(S = 1)$.

3. Dr. Evil gives twenty quizzes in his Henchman Studies 103 course. Each quiz has a single question, drawn from a list of ten review questions. Each list of review questions contains seven *Easy* questions and three *Hard* questions. Dr. Evil claims to select quiz questions completely at random with no regard to their difficulty. He claims, for example, that the first quiz will contain one question drawn at random from the ten review questions for Lecture #1. Yvonne suspects that Dr. Evil is *lying* about choosing questions completely at random. Because 9 out of the 20 quiz questions during the semester were *Hard*, she thinks Dr. Evil took question difficulty into account when creating his quizzes. Let H be the total number of *Hard* questions that appear on quizzes during the semester.

$\boxed{5}$ (a) If Dr. Evil is telling the truth, what is $E[H]$?

$\boxed{5}$ (b) If Dr. Evil is telling the truth, what is $Var(H)$?

$\boxed{10}$ (c) Yvonne decides to test the null hypothesis that Dr. Evil is telling the truth against the alternative that Hard questions are disproportionately *likely* to appear on quizzes, using the approximation based on the CLT. Calculate her test statistic.

$\boxed{5}$ (d) Yvonne enters the R commands

```
x <- 0.9 + 0:10 / 200
y <- qnorm(x)
cbind(x,y)
```

and obtains the following output from the console:

```
          x        y
 [1,] 0.900 1.281552
 [2,] 0.905 1.310579
 [3,] 0.910 1.340755
 [4,] 0.915 1.372204
 [5,] 0.920 1.405072
 [6,] 0.925 1.439531
 [7,] 0.930 1.475791
 [8,] 0.935 1.514102
 [9,] 0.940 1.554774
[10,] 0.945 1.598193
[11,] 0.950 1.644854
```

Continuing from the preceding part, approximately what is the p-value for her test? Interpret her results.

$\boxed{20}$ 4. Write an R function called `myreg` to estimate β_0 and β_1 in the simple linear regression $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Your function should take two input arguments: a vector `y` of observed outcomes and a corresponding vector `x` of observed values for the predictor variable. You may assume that there are no missing values and that the lengths of `x` and `y` are the same. Your function should return a vector with two elements: the estimate of β_0 and the estimate of β_1 (in that order). In your answer you may use any R functions that you like *except* for `lm`.

$\boxed{20}$ 5. This problem is taken from the extensions. It has been re-worded slightly for clarity, but the solution is unchanged. Let Y and X be RVs. In this problem you will find the the constants β_0 and β_1 that solve

$$\min_{\beta_0,\beta_1} E[(Y-\beta_0-\beta_1 X)^2].$$

For the purposes of this question you may assume that expectation and differentiation can be interchanged, i.e. that $\frac{\partial}{\partial\theta}E[f(Z,\theta)] = E[\frac{\partial}{\partial\theta}f(Z,\theta)]$. You do not have to check the second order condition.

(a) Show that $\beta_0 = E[Y] - \beta_1 E[X]$.

(b) Using the preceding part, find β_1.

6. This question relies on an R dataframe called `kaiser` with data for 1174 babies born at the Kaiser Foundation hospital in Oakland California. Here are the first few rows:

```
  bwt gestation smoke
1 120       284     0
2 113       282     0
3 128       279     1
4 108       282     1
5 136       286     0
6 138       244     0
```

Each row in `kaiser` is a newborn baby: `bwt` gives the baby's birthweight in ounces, `gestation` gives the length of the pregnancy in days, and `smoke` is a dummy variable taking the value one if the baby's mother smoked during pregnancy. The last page of this exam contains results for five regression models estimated using `kaiser`. You may find it helpful to tear out the page of regression results for ease of reference.

$\boxed{5}$ (a) What is the sample mean of `bwt`?

$\boxed{5}$ (b) Approximately what is the sample variance of `bwt`?

$\boxed{5}$ (c) Explain why the R-squared of Regression #1 is exactly zero.

$\boxed{5}$ (d) Which mothers have heavier babies: those who smoke or those who do not? How large is the difference?

$\boxed{5}$ (e) Continuing from the preceding part, is there convincing evidence of a difference in the population, or could our estimate by explained by sampling variation?

$\boxed{5}$ (f) Continuing from the preceding part, does the `kaiser` dataset provide evidence that smoking during pregnancy has a *causal effect* on birthweight? Why or why not?

$\boxed{5}$ (g) About how accurately does a regression that uses *only* `gestation` predict birthweight?

$\boxed{5}$ (h) What is the approximate value of the correlation between `bwt` and `gestation`?

$\boxed{5}$ (i) Consider two mothers, both of whose pregnancies lasted exactly d days: Xanthippe smoked during pregnancy while Yvonne did not. Based on this information, whose baby would we predict will be heavier at birth? Does your answer depend on whether we use Regression #4 or #5 to make our prediction? Explain briefly.

$\boxed{5}$ (j) Is there convincing evidence of a different slope in relationship between `gestation` and `bwt` for smokers versus non-smokers? If so, what is the nature of the difference?

Regression #1

```
lm(formula = bwt ~ 1, data = kaiser)
            coef.est coef.se
(Intercept) 119.46     0.53
---
n = 1174, k = 1
residual sd = 18.33, R-Squared = 0.00
```

Regression #2

```
lm(formula = bwt ~ smoke, data = kaiser)
            coef.est coef.se
(Intercept) 123.09     0.66
smoke        -9.27     1.06
---
n = 1174, k = 2
residual sd = 17.77, R-Squared = 0.06
```

Regression #3

```
lm(formula = bwt ~ gestation, data = kaiser)
            coef.est coef.se
(Intercept) -10.75     8.54
gestation     0.47     0.03
---
n = 1174, k = 2
residual sd = 16.74, R-Squared = 0.17
```

Regression #4

```
lm(formula = bwt ~ smoke + gestation, data = kaiser)
            coef.est coef.se
(Intercept) -3.18     8.33
smoke       -8.37     0.97
gestation    0.45     0.03
---
n = 1174, k = 3
residual sd = 16.25, R-Squared = 0.22
```

Regression #5

```
lm(formula = bwt ~ smoke + gestation + smoke:gestation, data = kaiser)
                coef.est coef.se
(Intercept)      19.64    10.29
smoke           -72.69    17.23
gestation         0.37     0.04
smoke:gestation   0.23     0.06
---
n = 1174, k = 4
residual sd = 16.16, R-Squared = 0.22
```