Final Examination

Econ 103, Statistics for Economists

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: ________________________________

Question:	1	2	3	4	5	6	7	Total
Points:	20	10	14	14	17	30	25	130
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. 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.

1. For each statement, indicate whether it is TRUE or FALSE and briefly explain why.

$\boxed{4}$ (a) For any two random variables X and Y, $E[XY] = Cov(X,Y) - E[X]E[Y]$.

> **Solution:** FALSE. The shortcut rule for covariance is $Cov(X,Y) = E[XY] - E[X]E[Y]$. Rearranging gives $E[XY] = Cov(X,Y) + E[X]E[Y]$. The expression in the problem statement has the wrong sign on $E[X]E[Y]$.

$\boxed{4}$ (b) Let $X_, \ldots, X_n \sim i.i.d.$ Bernoulli(p) and define $\widehat{p} = \sum_{i=1}^{n} X_i/n$. If sample size n is sufficiently large, the approximate sampling distribution of $\widehat{p}$ is chi-squared.

> **Solution:** FALSE. By the Central Limit Theorem, $\widehat{p}$ is approximately normally distributed.

$\boxed{4}$ (c) Suppose we have i.i.d. random sample from a normal distribution with mean μ and variance σ^2. A random variable Y is defined to be $\frac{(n-1)S^2}{\sigma^2}$ where n is sample size, and S^2 is sample variance. Then Y is Student t-distributed with degrees of freedom $n-1$.

> **Solution:** FALSE. $Y \sim \chi^2(n-1)$.

$\boxed{4}$ (d) If [3, 8] is a 95% CI for μ, we do not reject $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ with 5% significance level.

> **Solution:** FALSE: we would reject $\mu = 1$ since 1 lies outside the CI.

$\boxed{4}$ (e) The p-value for my test was 0.02. This means that if I had set $\alpha = 0.05$ I would have rejected the null hypothesis.

> **Solution:** TRUE. The p-value is the *minimum* significance level at which we would reject the null. Since $0.05 > 0.02$ we would have rejected at the 5% level.

2. Suppose I flip a fair coin and roll a single fair die at the same time. Define the events
$A =$ the coin comes up tails
$B =$ the die shows a 3 *or* 5
$C =$ the die shows an *odd* number

$\boxed{3}$ (a) Calculate $P(B|C)$.

Solution:

$$P(B|C) = P(B \cap C)/P(C) = (1/3)/(1/2) = 2/3$$

$\boxed{3}$ (b) Calculate $P(A \cap B)$.

Solution: Since the dice roll and coin flip are independent, we have $P(A \cap B) = P(A)P(B) = (1/2) \times (1/3) = 1/6$. You could also draw out the table with all 12 basic outcomes, all of which are equally likely, and count how many are in both A and B. This will give you $2/12 = 1/6$

$\boxed{4}$ (c) Calculate $P(A \cup B)$.

Solution: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/3 - 1/6 = 3/6 + 2/6 - 1/6 = 4/6 = 2/3$. Again, you could also draw out the table with all 12 basic outcomes, all of which are equally likely, and count how many are in either A, B or both A and B. This will give you $8/12 = 2/3$

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

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

$\boxed{2}$ (a) What is the support of X?

Solution: The support is the interval from zero to one $[0, 1]$ since the pdf is zero everywhere else.

$\boxed{5}$ (b) Calculate the cumulative distribution function of X.

Solution: $\int_{-\infty}^{x_0} f(x)dx = \int_0^{x_0} 3x^2 dx = x^3|_0^{x_0} = x_0^3$. Hence,

$$F(x_0) = \begin{cases} 0 & \text{for } x_0 < 0 \\ x_0^3 & \text{for } 0 \leq x_0 \leq 1 \\ 1 & \text{for } x_0 > 1 \end{cases}$$

$\boxed{3}$ (c) Calculate $E[X^2]$.

Solution:

$$E[X^2] \quad = \quad \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 3x^4 dx = \left.\frac{3x^5}{5}\right|_0^1 = 3/5$$

$\boxed{4}$ (d) Calculate $Var(X)$.

Solution:

$$Var(X) = E[X^2] - (E[X])^2 = 3/5 - 9/16 = 3/80$$

4. For each part, write your answer in the space provided. No explanation is needed.

$\boxed{3}$ (a) Write an R command to calculate the median of a $F(2, 1)$ random variable.

Solution: `qf(0.5, df1 = 2, df2 = 1)`

$\boxed{3}$ (b) Write a single R command to draw five numbers at random from the digits 0–9 *with replacement.*

Solution: `sample(0:9, size = 5, replace = TRUE)`

$\boxed{4}$ (c) Write R code to plot the cdf of a standard normal random variable between -3 and 3 using a grid of x-values with a step size of 0.01.

Solution:

```
x <- seq(from = -3, to = 3, by = 0.01)
plot(x, pnorm(x), type = 'l')
```

$\boxed{4}$ (d) Write an R function called `zscores2` that takes a vector `x` as its only input and generates the squared z-scores of `x` as its output. You may use any R functions that you like in your answer and may assume that there are no missing values.

Solution:

```
zscores2 <- function(x){
 return(((x - mean(x))/sd(x))^2)
}
```

5. Suppose that $X_1 \sim N(\mu, \sigma^2)$ independently of $X_2 \sim N(\mu, 3\sigma^2)$. Let $\bar{X} = (X_1 + X_2)/2$.

$\boxed{4}$ (a) Calculate the variance of $\bar{X}$.

Solution: In this example,

$$Var(\bar{X}) = Var\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}\left[Var(X_1) + Var(X_2)\right] = \frac{1}{4}\left(\sigma^2 + 3\sigma^2\right) = \sigma^2$$

$\boxed{4}$ (b) Let $\tilde{\mu} = \frac{3}{4}X_1 + (1 - \frac{3}{4})X_2$. Is $\tilde{\mu}$ an unbiased estimator of μ? Prove your answer.

Solution: Yes:

$$E[\tilde{\mu}] = E[\frac{3}{4}X_1 + (1 - \frac{3}{4})X_2] = \frac{3}{4}\mu + (1 - \frac{3}{4})\mu = \mu$$

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

Solution:

$$Var(\tilde{\mu}) = Var[\frac{3}{4}X_1 + (1 - \frac{3}{4})X_2] = \frac{9}{16}\sigma^2 + \frac{3}{16}\sigma^2 = \frac{3}{4}\sigma^2$$

$\boxed{4}$ (d) Is the sample mean $\bar{X}$ an efficient estimator of μ in this example? Explain.

Solution: Although X_2 gives us information about the mean μ this information is "three times as noisy" as the information contained in X_1. Hence, by giving X_2 a lower weight than X_1, we achieve an estimator with a lower variance. In this example the sample mean is NOT efficient because there is another unbiased estimator $\tilde{\mu}$ with a lower variance.

6. This question is based on a recent study that examines how "organic" labeling changes people's perceptions of different food products. Researchers recruited volunteers at a local mall in Brooklyn, 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 both samples contained exactly the same kind of yogurt, each contained the same number of calories. However, this was unknown to the volunteers.) To prevent confounding factors, 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.	0.7	
Sample Size	115	

$\boxed{4}$ (a) Give the units of each of the summary statistics from above:

Sample Mean ____________________
Sample Var. ____________________
Sample SD ____________________
Sample Corr. ____________________

Solution: calories, calories^2, calories, unitless.

$\boxed{4}$ (b) Sara thinks this experiment should be analyzed as independent samples data. Assume she is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the Central Limit Theorem (CLT).

Solution: The difference of means (regular minus organic) is 23 calories. Sara calculates her standard error assuming independent samples:

$$\sqrt{\sigma_r^2/n + \sigma_o^2/m} = \sqrt{3600/115 + 2916/115} = \sqrt{6516/115} \approx 7.5$$

so her confidence interval is approximately 23 ± 15, in other words $(8, 38)$.

$\boxed{8}$ (c) Sara conducts a hypothesis testing on the difference of means (`regular` - `organic`) with 5% significance level, treating the data as independent samples. Derive the power of this test. (*Hint:* Consider the null $H_0 : \mu_r = \mu_o$ against $H_1 : \mu_r \neq \mu_o$ where μ_r is for `regular` and μ_o is for `organic`. Denote the sample mean of each group to be $\bar{X}_r$ and $\bar{X}_o$ when constructing the test statistic. State the decision rule of hypothesis testing and how the test statistic is distributed under the alternative in order to express the power. The power expression will depend on $(\mu_r - \mu_o)$.)

Solution: The test statistic

$$T_n = \frac{\bar{X}_r - \bar{X}_o}{\sqrt{\sigma_r^2/n + \sigma_o^2/m}} = \frac{\bar{X}_r - \bar{X}_o}{7.5}$$

follows a standard normal distribution under the null. The decision rule with 5% significance level is to reject the null if $|T_n| \geq 2$. Under the alternative, however, the above test statistic does not follow a standard normal distribution. Instead,

$$\frac{(\bar{X}_r - \bar{X}_o) - (\mu_r - \mu_o)}{7.5} \sim N(0,1)$$

Therefore,

$$T_n = \frac{(\bar{X}_r - \bar{X}_o) - (\mu_r - \mu_o)}{7.5} + \frac{(\mu_r - \mu_o)}{7.5} \sim N(0,1)$$

That is, $T_n \sim N((\mu_r - \mu_o)/7.5, 1)$. Combining the decision rule with the distribution of the test statistic under the alternative, we calculate power as follows:

$$\begin{aligned} Power(\mu_r - \mu_o) &= P(\text{Reject } H_0 | \text{under } H_1) = P(|T_n| \geq 2) \\ &= P(T_n < -2) + P(T_n > 2) \\ &= P(Z + (\mu_r - \mu_o)/7.5 < -2) + P(Z + (\mu_r - \mu_o)/7.5 > 2) \\ &= P(Z < -2 - (\mu_r - \mu_o)/7.5) + P(Z > 2 - (\mu_r - \mu_o)/7.5) \\ &= \texttt{pnorm}(-2 - (\mu_r - \mu_o)/7.5) + \big(1 - \texttt{pnorm}(2 - (\mu_r - \mu_o)/7.5)\big) \end{aligned}$$

$\boxed{6}$ (d) 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.

Solution: Kevin takes into account the sample correlation between columns when calculating his standard error. He does this by using the sample statistics from the table to calculate the sample variance of the *differences*: regular minus organic. In particular, he calculates:

$$s_D^2 = 3600 + 2916 - 2 \cdot 0.7 \cdot 60 \cdot 54 = 1980$$

which gives a standard error of

$$\sqrt{s_D^2/n} = \sqrt{1980/115} \approx 4.1$$

This is the only difference between his procedure and Sara's. Hence, Kevin's confidence interval is approximately 23 ± 8.2, in other words $(14.8, 31.2)$.

$\boxed{4}$ (e) 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?

Solution: Kevin is right and Sara is wrong. This is matched pairs data because each row corresponds to a *single individual*. Unsurprisingly, we find a high sample correlation between the two columns: individuals who overestimate caloric content for one yogurt sample tend to do so for the other, as do individuals who underestimate. The only difference between Kevin and Sara's confidence intervals comes from how they calculated their standard errors. Both intervals are correctly centered, but Sara's is *too wide* because she calculated the standard error assuming independence between the two samples. When the sample correlation is positive this results in an *overestimate* of the standard error.

$\boxed{4}$ (f) 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?

Solution: Kevin's test statistic is the difference of means divided by the standard error, namely $23/4.1 \approx 5.6$. To calculate the p-value in R, he should use the command:

```
2 * (1 - pnorm(5.6))
```

The result will be less than 0.05 since the test statistic is larger than 2. Another way to see this is that his confidence interval does not include zero.

7. 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 on the last page. You may want to tear the page out for convenience.)

$\boxed{5}$ (a) 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 with 5% significance level. I can carry out this test directly from Regression 1. 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?

Solution: We should use the coefficient estimate for `male` since it is the difference of mean test scores between men and women (i.e. $\bar{x}_M - \bar{x}_W$). The test statistic is the absolute value of the ratio of this estimated coefficient divided by its estimated standard error: $|-0.22/2.91| \approx 0.08$. The approximate critical value for this test is 2, so we fail to reject the null hypothesis.

$\boxed{3}$ (b) What is the sample correlation between students' scores on the two midterms?

Solution: For a simple linear regression, the R^2 is the square of the sample correlation between x and y. Hence, $\sqrt{0.28} \approx 0.53$ is the sample correlation between scores on the first and second midterm.

$\boxed{5}$ (c) Explain the meaning of the coefficient estimate `midterm1` in Regression 2 and construct a 95% confidence interval for this parameter.

Solution: This estimate tells us the difference in score on midterm two that we would predict between two groups of students who differed by one point in their scores on midterm one: people who did one point better on the first exam do about 0.6 points better on the second exam. An approximate 95% confidence for this parameter is 0.6 ± 0.24, in other words $(0.36, 0.84)$. This interval does not include zero and is bounded substantially away from it. Our data strongly suggest that people who did better on the first exam continue to do better on the second.

$\boxed{3}$ (d) 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. Is this coefficient statistically significant at the 5% significance level?

Solution: We know immediately that we can reject at the 5% level since zero is not contained in the confidence interval from part (d). So the coefficient is statistically significant.

4 (e) Based on Regression 3, what is the predicted score in `midterm2` of a male student who scored 70 in `midterm1`?

Solution: The predicted value is 34.79 -0.31*1 + 0.6*70 = 76.48.

5 (f) Interpret the coefficients in Regression 4. (*Hint*: Consider intercept and slope for male and female, respectively.)

Solution: Regression 4 allows both the intercept and the slope to vary across sex. The coefficient `male` gives the difference of midterm2's mean between male and female holding everything else as fixed. The coefficient `midterm1` gives the difference in midterm 2 score that we would predict between two females who differed by one point in midterm 1. The summation of coefficient `midterm1` and coefficient `male:midterm1` gives the difference in midterm 2 score that we would predict between two males who differed by one point in midterm 1.

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     0.60     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     0.60     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
```