Econ 3040 A01 - Midterm - Answer Key

Ryan T. Godwin

1. Players of Dungeons and Dragons use dice that have more than just 6 sides. Consider an 8-sided die that has the values $\{1, 2, 3, 4, 5, 6, 7, 8\}$.

   a) **[2 marks]** Find the expected value of an 8-sided die roll.

   $E[Y] = \left(\frac{1}{8} \times 1\right) + \left(\frac{1}{8} \times 2\right) + \cdots + \left(\frac{1}{8} \times 8\right) = 4.5$

   The variance of an 8-sided die roll is 5.25.

   ---

   b) **[3 marks]** Write down the probability function for an 8-sided die roll. What two important things does a probability function accomplish?

   The probability function is: $f(Y = y) = \frac{1}{8}$; $y = 1, 2, \ldots, 8$, where Y is the result of the die roll. The probability function accomplishes two important things: (i) it describes all possibilities, and (ii) assigns a probability to each possibility.

   c) **[3 marks]** What is the expected value and variance of the **sum** of two 8-sided dice?

   Let Y be the result of one of the dice, X the result of the other. Then:
   $E[Y + X] = E[Y] + E[X] = 4.5 + 4.5 = 9$, and
   $var[Y + X] = var[Y] + var[X] = 5.25 + 5.25 = 10.5$.

   d) **[3 marks]** When you roll two 8-sided dice, what is the correlation between the result of the first roll and the second roll? Write one sentence that carefully explains how you know what this correlation is.

   The correlation is zero because the two die rolls are independent.

   e) **[3 marks]** Imagine that you roll 200 8-sided dice all at once, and sum the results. What probability distribution will this sum follow, and why?

   The result of the sum of 200 10-sided dice is a random variable that will approximately follow the Normal distribution. This is due to the CLT which, loosely speaking, says that if we add up random variables, the resulting sum is Normal.

2. X and Y are two random variables, and their joint probability function is:

	$Y = -2$	$Y = 0$	$Y = 2$
$X = 2$	0.25	0	0.25
$X = -2$	0	0.5	0

The covariance between X and Y is zero.

a) **[3 marks]** Are X and Y independent? Why or why not?

> See Review Question #2, Chapter 2. Although the correlation is zero, X and Y are **not** independent. If we observe the value of Y, then we know with certainty, the value of X.

b) **[2 marks]** Does X *cause* Y? Explain in one sentence.

> Although these two variables are related to each other (not independent), one does not necessarily cause the other.

3. **[4 marks]** The estimators $\bar{y}$, b_0, and b_1 are random, and have *sampling distributions* (a special name for a probability function). Why is it so important to understand that estimators are random variables? (That is, what kind of results can we derive only by realizing that estimators are random variables?)

> By understanding that these estimators are random variables, we can derive their mean and variance. This allows us to determine that they are unbiased, efficient and consistent, and allows us to perform hypothesis tests.

4. The University of Mars claims that its students receive, on average, a scholarship of \$300. You collect a sample of scholarship amounts from 200 randomly selected students. You calculate the sample average scholarship to be \$293.5, and sample variance to be \$111063.1.

a) **[3 marks]** The number \$111063.1 was calculated using a formula with $(n-1)$ in the denominator. Why is the denominator $(n-1)$ and not, for example, n?

> s_y^2, a formula with $(n-1)$ in the denominator, provides an *unbiased* estimate of the population variance.

b) **[6 marks]** Test the null hypothesis that the population mean scholarship amount is \$300. Calculate the t-test statistic, and associated p-value. Decide whether to reject or fail to reject the null hypothesis.

> The null and alternative hypotheses are:
> $H_0 : \mu = 300$
> $H_A : \mu \neq 300$
> The t-test statistic is:
>
> $$t = \frac{293.5 - 300}{\sqrt{\frac{111063.1}{200}}} = -0.28$$
>
> The p-value associated with this t-test statistic (using Table 3.2) is $0.3897 \times 2 = 0.78$. We fail to reject the null hypothesis at any significance level.

c) **[3 marks]** Calculate the 95% confidence interval around the sample mean. Without doing any additional work, decide whether to reject or fail to reject the null hypothesis that the population mean scholarship is \$240, at the 5% significance level.

$\bar{y} \pm 1.96 \times \sqrt{\frac{s_y^2}{n}} = 293.5 \pm 1.96 \times \sqrt{\frac{111063.1}{200}} = [247.3, 339.7]$
Since 240 is **not** inside the 95% confidence interval, we would reject $H_0 : \mu = 240$ at 5% significance.

d) **[3 marks]** Hackers obtain the entire database of scholarship amounts for all 31,254 University of Mars students. The true population mean scholarship amount is actually \$290.93! Carefully explain, using one or two sentences, the result of your hypothesis test in light of the true population mean.

Since we failed to reject the null hypothesis, when it was in fact false, we had a *type II error* occur in part (a).

5. Use the following data: $Y = \{-1, -3, 3\} \quad ; \quad X = \{-2, 0, 1\}$.

a) **[3 marks]** Calculate the OLS estimates b_0 and b_1.

$$\bar{Y} = -0.33, \quad \bar{X} = -0.33$$
$$b_1 = \frac{(-1+0.33)(-2+0.33) + (-3+0.33)(0+0.33) + (3+0.33)(1+0.33)}{(-2+0.33)^2 + (0+0.33)^2 + (1+0.33)^2} = 1$$
$$b_0 = -.33 + 0.33 = 0$$

Alternatively, b_1 and b_0 may be found using the following R code:

```
y <- c(-1,-3,3)
x <- c(-2,0,1)
lm(y ~ x)
```

b) **[2 marks]** Calculate the OLS predicted value, and residual, for the 3rd observation.

$\hat{Y}_3 = b_0 + b_1(X_3) = 0 + 1(1) = 1$
$e_3 = Y_3 - \hat{Y}_3 = 3 - 1 = 2$

c) **[3 marks]** Where did the formulas for b_0 and b_1 come from?

The formulas are a solution to a calculus minimization problem. b_0 and b_1 are the values that minimize the sum of squared residuals.

6. **[4 marks]** Use the following **incomplete** data for this question: $Y = \{2, 3, ?\} \quad ; \quad X = \{1, 2, ?\}$. The OLS residuals associated with this data are: $e = \{0, 0, 0\}$. What are b_0 and b_1?

Since all of the residuals are zero, the estimated OLS line passes through each data point (there are no prediction errors). To get the estimated line, we can connect the two data points. The slope (rise over run) is 1, and intercept is 1. Alternatively, we can ignore the missing data point and find b_1 and b_0 using the following R code:

```
y <- c(2,3)
x <- c(1,2)
lm(y ~ x)
```

7. You have been hired by the government to determine how much less alcohol per capita will be consumed if taxes are increased by \$2. You collect data on the quantity consumed (Q, measured in litres per year) at different price levels (P).

a) **[3 marks]** Write down the population model that you need to estimate.

The population model is:
$Q = \beta_0 + \beta_1 P + \epsilon$
or just:
$Y = \beta_0 + \beta_1 X + \epsilon$

b) **[3 marks]** Why should you use OLS to estimate this model?

Because OLS is unbiased, efficient, and consistent.

c) **[3 marks]** What are some factors that might be in ϵ?

ϵ contains all factors that determine *quantity*, besides *price*. Some of these factors could be incomes, the prices of substitutes and complements, tastes and preferences, etc.

The estimated model is $\hat{Q} = 15.28 - 0.75P$.

---

d) **[3 marks]** What do you tell the government?

Quantity consumed is predicted to decrease by 1.5 L per capita per year ($-0.75 \times 2 = -1.5$).

e) **[3 marks]** How much alcohol will be consumed if the price is set at \$10?

The predicted quantity consumed when price equals \$10 is:
$\hat{Q} = 15.28 - 0.75 \times 10 = 7.78$

END