# Midterm #2

answers

**Instructions:** Each sub-question is worth 5 points. The exam is out of 90 points, so you get 10 points for free.

Refer to the provided probability tables as necessary. You may use a calculator, and one sheet of notes. You will never be penalized for showing work, but if what is asked for can be computed directly, points awarded will depend primarily on the correctness of your numerical answer. Good luck!

**Problem 1** You wish to estimate $\mu$, the average income of UK economics alumni. To do so, you use a university contact list to randomly sample 400 alumni to determine their income.

**a.** Explain, in detail, what you know about the sampling distribution of $\overline{x}$, the sample mean. The Central Limit Theorem states that $\overline{x}$ is a normal random variable, with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$, where $\sigma$ is the standard deviation of income, and $\sqrt{n} = 20$.

**b.** Suppose that income is known to have a standard deviation of $\sigma = 10,000$. What is the probability that your sample mean, $\overline{x}$, differs from $\mu$ by more than \$750?

\$750 is 1.5 standard deviations; the question is asking for the probability that a normal random variable differs from its mean by more than 1.5 standard deviations. The answer, from consulting a standard normal probability table, is .1336.

**c.** After conducting your sample, you determine that $\overline{x} = \$50,000$. Calculate a 95% confidence interval for $\mu$.

$[49,020, 50,980]$

**Problem 2** Dean Blackwell is concerned that Gatton professors are canceling too many classes. However, it is not feasible for the dean to personally monitor all of the hundreds of courses offered by Gatton each year, so he randomly chooses 25 of them, and checks each day to see if the class is being taught or not. The dean considers an acceptable number of canceled classes to be two or fewer per semester, per course. From past experience, the dean knows that the standard deviation of number of canceled classes is $\sigma = .5$.

**a.** State the appropriate null and alternative hypotheses for a test of whether or not too many classes are being canceled.

The null hypothesis should be the status quo outcome, or that the mean number of classes canceled per semester, per course is 2 or less. The alternative is that it is 2 or more.

**b.** For what values of the sample mean $\overline{x}$ do you reject the null hypothesis, with $\alpha = .05$? You reject $H_0$ if the test statistic $\frac{\overline{x}-2}{\frac{.5}{5}} > 1.64$, or if $\overline{x} > 2.164$.

**c.** Suppose in his sample of 25 courses, the dean finds that an average of 2.15 classes are canceled per course. What is the p-value of the test? The p-value is the probability of observing a test statistic at least as large as $\frac{2.15-2}{\frac{.5}{5}} = 1.5$ under $H_0$. Checking a standard normal probability table, we get that this probability is .0668.

**d.** Explain in words what the p-value you found in part c tells you. The p-value gives the lowest probability of making a type I error (incorrectly rejecting $H_0$) that would cause us to reject $H_0$.

**Problem 3** Consider the following regression equation:

$$y = \beta_0 + \beta_1 * X + \epsilon \tag{1}$$

**a.** Explain in plain English how to interpret the coefficient $\beta_1$.

$\beta_1$ is the amount by which $y$ increases given a one-unit increase in $x$.

**b.** Explain in plain English why the $\epsilon$ term is necessary.

The $\epsilon$ term represents noise, or luck. When we model the relationship between two variables using a linear regression model, we do not expect that there is an exact linear relationship between $y$ and $x$, but instead expect that the average value of $y$ has a linear relationship with $x$. The $\epsilon$ term is sometimes positive, sometimes negative, meaning that data will sometimes be above and sometimes below a regression line.

**c.** Suppose that $x$ is mileage and $y$ is the selling price of a certain type of used car. You estimate (1) using data provided by a local dealership. Do you expect the regression results to indicate that $\hat{\beta}_1 > 0$ or that $\hat{\beta}_1 < 0$? Why?

We would expect $\hat{\beta}_1 < 0$, as cars with greater mileage will, on average, sell for a lower dollar amount than cars with lower mileage.

**d.** Suppose that $x$ is the number of McDonald's restaurants in a country and $y$ is the GDP of a country. You estimate (1) using a dataset containing information on 150 countries, and find a large, positive, and statistically significant estimate $\hat{\beta}_1$. Does this mean that McDonald's restaurants cause a high GDP? Explain.

No, it means that there is a correlation between the number of McDonald's in a country and GDP, but it does not mean that the relationship is causal. For example, it may be that a high GDP is something McDonald's looks for when entering a country, and so the causality may even go from $y$ to $x$, and not the other way around.

**Problem 4** Consider the following regression model:

$$WAGE = \beta_0 + \beta_1 * EXPERIENCE + \epsilon \tag{2}$$

where $WAGE$ is hourly wage and $EXPERIENCE$ is years of full-time work experience. Suppose you estimate the regression parameters using a large Census dataset (a random sample of all U.S. residents). Your results are as follows:

| Regression | Statistics |
|---|---|
| Multiple R | .22358 |
| R Square | .24985 |
| Adjusted R Square | .24985 |
| Standard error | 5.2198 |
| Observations | 189,253 |

| | coefficients | Standard error | t stat | P-value | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Intercept | 7.11 | 2.00 | 3.56 | .000186 | 4.0924 | 14.1276 |
| EXPERIENCE | .96 | .31 | 3.097 | .000978 | .3524 | 1.5676 |

**a.** Consider the following hypothesis test:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Is there sufficient evidence to support rejecting $H_0$, with $\alpha = .05$?? How do you know?

Yes, we would reject $H_0$ in favor of $H_A$ here, as the p-value associated with the test is $.000978 < .05$.

**b.** Interpret the estimate of $\beta_1$ precisely.

The estimate of .96 means that one year of additional full-time work experience is associated with a 96 cent hourly pay increase, on average.

Differing levels of full-time experience is often suggested as one reason why women earn a lower wage than men. Suppose that, among 45-year old workers, women have an average of 17.7 years of full-time work experience, while men have an average of 22.6 years of full time work experience.

**c.** Give a point estimate for the wage of a worker with 17.7 years of experience. Give a point estimate for the wage of a worker with 22.6 years of experience.

Using the estimated regression line $WAGE = 7.11 + .96 * EXPERIENCE$, we get point estimates of $24.10 and $28.81, respectively.

**d.** Suppose that the average 45-year old man earns a wage which is $6.76 higher than the average wage of a 45-year old woman. What fraction of this difference is explained by differing levels of full-time work experience, according to the results?

The results of part c predict that differing levels of full-time work experience account for $4.70 of the difference between men and women, or fraction .696 of the difference.

**Problem 5** A criminologist is concerned about the number of crimes in a particular town that involve guns. He randomly chooses 600 crime reports and finds that 420 of the crimes involve a gun.

**a.** Develop a 95% confidence interval for the proportion of crimes that involve a gun.

[.6633, .7367]

**b .** What is the error of estimation for this interval estimate?

.0367

**c.** How many reports would the criminologist need to review in order to reduce the margin of error to .01?

We solve the equation $.01 = 1.96 * \sqrt{\frac{.7*.3}{n}}$. If the sample proportion stays the same, we would need a sample size of at least 8,068.