Ecn 102 - Analysis of Economic Data
University of California - Davis
Instructor: John Parman

April 23, 2009

Midterm 1 - Solutions

You have until 10:20am to complete this exam. Please remember to put your name, section and ID number on both your scantron sheet and the exam. Fill in test form A on the scantron sheet. Answer all multiple choice questions on your scantron sheet. Choose the single best answer for each multiple choice question. Answer the long answer questions directly on the exam. Keep your answers complete but concise. For the long answer questions, you must show your work for full credit.

Name: **ID Number:** **Section:**

(POTENTIALLY) USEFUL FORMULAS

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$CV = \frac{s}{\bar{x}}$

$skew = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^3$

$kurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$

$\mu = E(X)$

$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

$Pr[T_{n-1} > t_{\alpha,n-1}] = \alpha$

$Pr[|T_{n-1}| > t_{\frac{\alpha}{2},n-1}] = \alpha$

$\sum_{i=1}^{n} a = na$

$\sum_{i=1}^{n}(ax_i) = a\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

$s^2 = \bar{x}(1-\bar{x})$ for proportions data

$t_{\alpha,n-1} = TINV(2\alpha, n-1)$

$Pr(|T_{n-1}| \geq |t^*|) = TDIST(|t^*|, n-1, 2)$

$Pr(T_{n-1} \geq t^*) = TDIST(t^*, n-1, 1)$

(POTENTIALLY) USEFUL EXCEL OUPUT

TINV(.01,499)=2.59	TINV(.01,399)=2.59	TINV(.01,299)=2.59
TINV(.02,499)=2.33	TINV(.02,399)=2.34	TINV(.02,299)=2.34
TINV(.025,499)=2.25	TINV(.025,399)=2.25	TINV(.025,299)=2.26
TINV(.05,499)=1.96	TINV(.05,399)=1.97	TINV(.05,299)=1.97
TINV(.10,499)=1.65	TINV(.10,399)=1.65	TINV(.10,299)=1.65
TINV(.20,499)=1.28	TINV(.20,399)=1.28	TINV(.20,299)=1.28

SECTION I: MULTIPLE CHOICE (55 points)

1. Suppose you have annual observations of obesity rates for the past fifty years and you want a graph of the data that shows the trend in obesity over time. The best graph for this purpose would be:
   (a) A bar chart.
   (b) A pie chart.
   (c) A line chart.
   (d) A histogram.

   (c) A line chart would allow us to see how obesity rates are changing over time. The other types of charts would show us frequencies of particular values for obesity but wouldn't tell us which values were more frequent at different points in time.

2. The probability of a Type II error occurring depends on:
   (a) The true population mean.
   (b) The significance level, α.
   (c) The sample size.
   (d) All of the above.

   (d) The further the true population mean is from μ_0, the less likely we are to observae a sample mean that leads us to fail to reject the null hypothesis. The significance level determines the range of sample means over which we would fail to reject the null hypothesis so changing the significance level will change the probability of incorrectly failing to reject the null. The sample size influences the distribution of $\bar{X}$. As the sample size gets larger, the probability of getting a $\bar{x}$ close to the true mean goes up, so the likelihood of failing to reject the null hypothesis will go down.

3. Which of the following statements is false?
   (a) The interquartile range is always less than or equal to the range.
   (b) The coefficient of variation will be smaller when the sample mean is larger, all other things held constant.
   (c) The mode is always less than or equal to the median.
   (d) The mean will increase if the largest observation is doubled.

   (c) The mode is just the most frequent observation and can be less than the median, equal to the median or greater than the median. Think about the following set of observations: 0, 1, 2, 3, 4, 5, 5. For this sample, 5 is the mode which is clearly greater than the median.

4. The magnitude of the critical value for an upper one-tailed test at a 5% significance level will be:
   (a) Greater than the magnitude of the critical value for a two-tailed test using the same data at a 5% significance level.
   (b) Less than the magnitude of the critical value for a two-tailed test using the same data at a 5% significance level.

(c) Equal to the magnitude of the critical value for a two-tailed test using the same data at a 5% significance level.
(d) None of the above.

(b) For the two-tailed test, the probability of being to the right of the critical value is 2.5% while for the one-tailed test, the probability of being to the right of the critical value is 5%. Given that the distribution of t is centered at zero, this means that the critical value for the two-tailed test must be larger than the critical value for the one-tailed test.

5. A dataset containing the monthly food expenditures for 10 different families for an entire year (120 observations in total) is an example of:

(a) Time series data.
(b) Cross-section data.
(c) Panel data.
(d) Categorical data.

(c) We have observations of a cross-section of individuals at several points in time which is panel data. A single observation is identified both by the family it corresponds to and the month it corresponds to.

6. If X_a and X_b are two different consistent, unbiased estimators for the population mean, we would prefer to use the one with:

(a) The larger variance.
(b) The smaller variance.
(c) Whichever one gives us a larger value for the sample mean.
(d) Whichever one gives us a smaller value for the sample mean.

(b) The one with the smaller variance will allow us to be more certain about the value of μ. An easy way to see this is to think about the formula for the confidence interval. If s is smaller, the confidence interval for the mean will be narrower.

7. Which of the following is not a measure of central tendency?

(a) Interquartile range.
(b) Sample midrange.
(c) Median.
(d) Mean.

(a) The interquartile range is the difference between the 75th and 25th percentile. If we shifted the entire distribution to the right, the center of the distribution would clearly change but the interquartile range would stay exactly the same.

8. Which of the following would decrease the likelihood of both Type I and Type II errors?

(a) Increasing the significance level, α.
(b) Decreasing the significance level, α.
(c) Neither (a) nor (b) are correct.
(d) Both (a) and (b) are correct.

(c) Increasing α would decrease the likelihood of a Type II error but would increase the likelihood of a Type I error. Decreasing α would do the opposite, increasing the likelihood of a Type II error but decreasing the likelihood of a Type I error.

9. Increasing the width of the bins for a histogram will tend to:

   (a) Increase the absolute frequency of observations in each bin.
   (b) Decrease the absolute frequency of observations in each bin.
   (c) Lead to a greater proportion of bins with no observations.
   (d) Shift the center of the distribution to the right.

   (a) Increasing the width of the bins means fewer bins. Given that the total number of observations is still the same, this will mean more observations in each bin on average.

10. Which of the following is not a random variable?

   (a) The sample mean.
   (b) The sample standard deviation.
   (c) The population mean.
   (d) All of the above are random variables.

   (c) The population mean is a constant. The sample mean and sample standard deviation can take on different values depending on the sample we draw from the population.

11. We will reject the null hypothesis whenever:

   (a) $p < \alpha$.
   (b) $p > \alpha$.
   (c) $t^* > c$.
   (d) $t^* < c$.

   (a) We always reject the null hypothesis when $p < \alpha$. Our rejection rule when using a critical value depends on the type of hypothesis we are testing (two-tailed, lower one-tailed, upper one-tailed).

12. Suppose that we are interested in the mean level of income, μ, for the entire adult population of the United States. We take a sample of college graduates and calculate the mean income for the sample, $\bar{x}$. Which of the following statements is most likely true about our sample?

   (a) The sample mean would provide a biased estimate of the population mean because $E(\bar{X}) > \mu$ in this case.
   (b) The sample mean would provide a biased estimate of the population mean because $E(\bar{X}) < \mu$ in this case.
   (c) The sample mean would provide an unbiased estimate of the population mean.
   (d) The expected value of the sample mean would be equal to the expected value of the population mean.

   (a) College graduates will have higher incomes on average than people who did not graduate from college. This means that the average income in our sample will tend to be higher than the average income in the population which includes college grads and non-college grads.

13. Which of the following types of data would be the most likely to be represented with a pie chart?

    (a) Categorical, cross-sectional data.
    (b) Continuous, cross-sectional data.
    (c) Continuous, time series data.
    (d) Continuous, panel data.

    (a) Categorical data is more likely to be represented in a pie chart than continuous data. A pie chart is a simple depiction of the percentage of observations in the various categories.

14. Suppose that we reject the null hypothesis that $\mu = 58$ for a particular random variable X. We used a significance level of 10% and had a positive value for our test statistic t^*. Which of the following statements is definitely true?

    (a) We would reject the null hypothesis that $\mu \leq 58$ at a significance level of 5%.
    (b) We would reject the null hypothesis that $\mu \geq 58$ at a significance level of 5%.
    (c) We would reject the null hypothesis that $\mu = 58$ at a signifiance level of 5%.
    (d) None of the above.

    (a) If we rejected the null hypothesis that $\mu = 59$ at a 10% significance level and had a positive t^*, then it must have been the case that $t^* > t_{.05,n-1}$. This is the criteria for rejecting the null hypothesis that $\mu \leq 58$ at a significance level of 5%.

15. If a distribution is symmetric:

    (a) The mean will be equal to the median.
    (b) The 75th percentile will be exactly as far from the mean as the 25th percentile.
    (c) The value for the midrange will be equal to the mean.
    (d) All of the above.

    (d) If a distribution is symmetric, everything to the right of the mean mirrors everything to the left. So any percentile p we choose will be just as far from the mean as the $1 - p$ percentile.

16. On a line graph, the values on the horizontal axis typically correspond to ______ while the values on the horizontal axis of a histogram typically correspond to ______.

    (a) The realized values of the variable (x_i), the realized values of the variable (x_i).
    (b) The realized values of the variable (x_i), the values of the index for observations (i).
    (c) The values of the index for observations (i), the realized values of the variable (x_i).
    (d) The values of the index for observations (i), the values of the index for observations (i).

    (c) A line graph typically shows the values of x_i on the vertical axis and the values of i on the horizontal axis (think of a graph of GDP over time). A histogram shows ranges of the values of x_i on the horizontal axis and frequencies on the vertical axis.

17. Suppose we are interested in using census data to estimate the mean level of education for the US population. If a 5% sample of the census and a 10% sample of the US census give us the same sample mean and the same sample variance, which will give us a narrower 95% confidence interval for the population mean?

    (a) The 5% sample.
    (b) The 10% sample.
    (c) The confidence intervals will be the same.
    (d) Not enough information.

    (b) The width of the confidence interval depends on $\frac{s}{\sqrt{n}}$. As n gets larger, the width of the confidence interval will shrink. So the 10% sample, with twice as many observations as the 5% sample, will give us a narrower confidence interval.

18. A pollster reports that 60% of the public approves of President Obama's handling of the economy with a margin of error of 4%. What is the probability that the true percentage of the population that approves is greater than 64%?

    (a) 1%.
    (b) 2.5%.
    (c) 5%.
    (d) 10%.

    (b) The margin of error is based on a 95% confidence interval. So there is a 5% chance that the true population mean is either greater than 64% or less than 56%. The probability of the mean being greater than 64% will therefore be $\frac{1}{2}$5%, or 2.5%.

19. Suppose that the absolute frequency of people with a wage equal to $20 an hour in a sample of 200 people is 40. Which of the following is definitely true?

    (a) 20% of the population earns a wage of $20 an hour.
    (b) The percentage of people in the sample earning $40 an hour or less is greater than or equal to 20%.
    (c) The percentage of people in the sample earning $10 an hour or more is less than or equal to 80%.
    (d) The percentage of people in the population earning $20 an hour or less is greater than or equal to 20%.

    (b) If 20% of the sample earns $20 an hour, than at least 20% of people in the sample earn less than $40. We can't say anything for certain about the earnings at the population level (we only get to observe the sample).

20. If most apartment rents in Davis are between $600 and $800 but there is a long tail of more expensive apartments in the rent distribution, which of the following is most likely to be true?

    (a) The distribution of apartment rents is left skewed.
    (b) The mean apartment rent is less than the median apartment rent.
    (c) The value for the skewness of the rent distribution is positive.
    (d) The 50th percentile of the rent distribution is greater than the mean apartment rent.

(c) The question describes data with a long right tail. This would be right skewed data that would have a positive value for skewness and a mean that is greater than the median.

21. Which of the following would be a way of getting a random sample of independent observations of Davis students?
    (a) Sample students on the basis of the last digit of their student ID number.
    (b) Choose a class at random and use the students in the class as the sample.
    (c) Choose all students born in March of 1989 as the sample.
    (d) All of the above.

    (a) The last digit of the student ID number should be a random way to select people. If we chose a class at random, everyone in the class would share certain characteristics (they are all taking the same class) and the observations woudl not be independent. Choosing students born in a certain month of a single year would give us students all of a very similar age which would not be representative of the student body as a whole.

22. The distribution of the sample mean depends on:
    (a) The mean of the population.
    (b) The variance of the population.
    (c) The sample size.
    (d) All of the above.

    (d) The distribution of the sample mean will be centered at the population mean. The variance will depend on the population variance and the sample size.

23. A 90% confidence interval will:
    (a) Always be centered at the population mean.
    (b) Contain the population mean with a probability of 10%.
    (c) Be wider if the sample variance is larger.
    (d) Be wider than a 95% confidence interval.

    (c) The confidence interval will be centered at the sample mean. If will be wider if the sample variance is larger or if we choose a smaller value of α.

24. When testing the null hypothesis that $\mu = 40$ we get a p-value equal to .06. We can say that:
    (a) $\mu = 40$ at a 10% significance level.
    (b) $\mu > 40$ at a 5% significance level.
    (c) We reject the null hypothesis at the 5% significance level.
    (d) We fail to reject the null hypothesis at the 1% significance level.

    (d) We would reject the null hypothesis when $p < \alpha$. So at a signficance level of 1%, we would fail to reject the null hypothesis on the basis of our p-value of .06.

25. Suppose that we have a sample of heights measured in centimeters. If we switch to measuring heights in meters:

(a) The new sample mean will be 100 times the old one, the new sample standard deviation will be 100 times the old one.
(b) The new sample mean will be 100 times the old one, the new sample standard deviation will be 100^2 times the old one.
(c) The new sample mean will be $\frac{1}{100}$ times the old one, the new sample standard deviation will be $(\frac{1}{100})^2$ times the old one.
(d) The new sample mean will be $\frac{1}{100}$ times the old one, the new sample standard deviation will be $\frac{1}{100}$ times the old one.

(d) All of our original x_i values will now be $\frac{x_i}{100}$. This will lead to a new sample mean that is equal to the old sample mean divided by 100. The sample standard deviation, which is in the same units as the sample mean, will also be reduced by a factor of $\frac{1}{100}$.

SECTION II: SHORT ANSWER (45 points)

1. (15 points total) Suppose that you are interested in the percentage of Davis students who would favor switching to a semester system. You take a random sample of 500 students. 300 students in the sample say that they would favor switching to the semester system. 100 students in the sample say that they want to stay with the quarter system. The remaining 100 students say that they don't care.

   (a) Suppose that the administration decides they will switch to semesters only if the percentage of students who favor switching is greater than 55%. The administration would prefer to error on the side of not switching. That is to say, if it is hard to tell whether the percentage of students favoring semesters is greater than 55%, the administration would not switch. Write down the null hypothesis and alternative hypothesis the administration should use to test whether the switch should be made.

   $$H_o\text{: } \mu \leq .55$$
   $$H_a\text{: } \mu > .55$$

   We clearly want to use a one-tailed test here (the administration doesn't want to simply know whether the percentage is equal to .55 or not equal to .55, they want to know if it is greater than .55). The question that remains is whether it should be an upper one-tailed test or a lower one-tailed test. The administration wants to play it safe and only switch if they are very certain $\mu > .55$. If we used a lower one-tailed hypothesis test, our null hypothesis would be that $\mu \geq .55$. We would fail to reject this null hypothesis even in some cases where $\bar{x}$ turns out to be less than .55. The administration would prefer to use an upper one-tailed test. That way they would only conclude that $\mu > .55$ when $\bar{x}$ is sufficiently greater than .55 to make the administration confident.

   (b) Calculate the t statistic you would use to test the hypothesis you wrote down in part (a).

   First, we need to calculate the sample mean and sample standard deviation:

   $$\bar{x} = \frac{300}{500} = .6$$

   $$s = \sqrt{\bar{x}(1 - \bar{x})}$$
   $$s = \sqrt{.6(1 - .6)} = .49$$

   Now we can calculate the test statistic:

   $$t^* = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

   $$t^* = \frac{.6 - .55}{\frac{.49}{\sqrt{500}}}$$
   $$t^* = 2.28$$

   (c) Based on your answer in part (b), use the critical value approach to test the hypothesis from part (a) at a 10% significance level. Be clear about exactly what critical value you are using. Clearly state your conclusion and what it implies for the administration's decision.

We are doing an upper one-tailed test at a 10% significance level for a sample of 500 observations. So the relevant critical value would be $t_{.10,499}$. This is the value corresponding to TINV(.20,499) in Excel (remember that Excel is actually giving you $t_{\frac{\alpha}{2},n-1}$ when you enter TINV(α,n-1)). Using the TINV() values given on the first page of the exam, the critical value would be 1.28. Our value of t^* is larger than the critical value, so we will reject the null hypothesis that the percentage of students favoring a switch to semesters is less than or equal to .55. So the administration should switch to semesters.

(d) Calculate a 95% confidence interval for the percentage of students who favor the staying with the quarter system.

To calculate a 95% confidence interval, we need the value of $t_{.025,n-1}$ which is given by TINV(.05,n-1). Looking at the table of TINV() values and using 500 as our value of n, this gives us $t_{.025,499} = 1.96$. For $\bar{x}$, we need to calculate the fraction of students who favor staying with the quarter system which is $\frac{100}{500}$ or .2. The standard deviation is then $\sqrt{.2(1-.2)}$ or .4. Now we have everything we need to calculate the confidence interval:

$$\bar{x} \pm t_{.025,n-1}\frac{s}{\sqrt{n}}$$

$$.2 \pm 1.96\frac{.4}{\sqrt{500}}$$

$$.2 \pm .035$$

So our 95% confidence interval is (.165, .235).

2. (6 points total) You take a sample of the weights of 10 males in our ECN 102 class. The observed weights (in pounds) are:

   165, 170, 185, 180, 165, 180, 170, 175, 170, 190

   (a) Calculate two different measures of the central tendency of the data.

   There are several correct answers to this question. Here are a few measures of central tendency:

$$\bar{x} = \frac{1}{10}(165 + 170 + 185 + 180 + 165 + 180 + 170 + 175 + 170 + 190) = 175$$

$$midrange = \frac{1}{2}(165 + 190) = 177.5$$

$$mode = 170$$

$$median = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(170 + 175) = 172.5$$

   (b) Calculate two different measures of the dispersion of the data. One of your measures should be in the same units as weight. The other measure should be unitless.

   A measure in the same units as weight would be the standard deviation:

$$s = \sqrt{\frac{1}{9}(\sum_{i=1}^{10}(x_i - 175)^2)} = 8.50$$

   You could also have used the interquartile range. A measure of dispersion that is unitless would be the coefficient of variation:

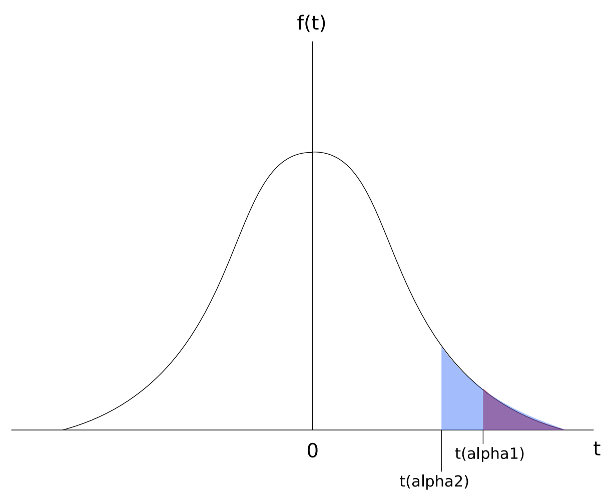
$$CV = \frac{s}{\bar{x}} = \frac{8.50}{175} = .049$$

3. (9 points total) For this question, consider a situation in which you are doing an upper one-tailed hypothesis test.

   (a) Describe a scenario in which a researcher would be very concerned about Type I error when doing an upper one-tailed hypothesis test.

   There are a wide range of possible answers. The key to a correct answer is that it should be situation in which rejecting the null hypothesis when it is actually true is a very costly mistake to make. For an upper one-tailed test, this would mean a situation where the population mean actually is less than or equal to a particular value but you reject this hypothesis in favor of the mean being greater than that value. An example where this would be bad is TV networks deciding to declare a winner in an election. They will declare candidate A the winner if the percentage of votes for candidate A is greater than .5 but will hold off on declaring a winner if the percentage of votes for candidate A is less than or equal to .5. In this scenario, it isn't very costly for the network to fail to reject the null hypothesis, it just means that they delay declaring a winner. However, if they incorrectly reject the null hypothesis, they incorrectly declare candidate A the winner which would make the network look very bad.

   (b) On a graph showing the distribution of t^*, show how the choice of a small value versus a large value of the significance level α influences the likelihood of a Type I error when doing an upper one-tailed hypothesis test.



   The graph above shows the critical values for an upper one-tailed test when using a small value of α (α_1) and a large value of α (α_2). The shaded areas give the probability of of Type I error. It is clear that the shaded region is larger for the larger value of α (the shaded area is exactly equal to α). When we choose a larger level for α we are increasing the range of t^* values for which we will reject the null hypothesis even though it is true.

   (c) Based on the situation you described in part (a), if the researcher switched to a lower one-tailed hypothesis test rather than an upper one-tailed hypothesis test, would the researcher be likely to change the choice of α? Explain your answer.

If the researcher switched to a lower one-tailed test, a Type I error would now be rejecting the null hypothesis that the population mean is greater than or equal to μ_0 when it actually is. In the case of calling an election cited above, this would be the equivalent of declaring that Candidate A had lost the election when he actually won. It is less clear in this case what the network would prefer. If the network is only concerned about incorrectly declaring somebody a winner, then they would not be as worried about Type I error as they would about Type II error in this case. So they would now select a larger value for α, increasing the probability of a Type I error but decreasing the probability of a Type II error.

4. (15 points total) You are given a dataset with annual observations of average real wage in the United States for the years 1950 to 1999. Write down the steps you would take in Excel to test whether the mean annual growth rate of real wages is equal to 5% using a 10% significance level. Be as specific as possible (explain all calculations you would make, what numbers you would use in your calculations, and how you would reach your conclusions).

   These are the general steps you would take:

   - The first step would be to create a variable measuring the annual growth rate of real wages. This could be done by creating a new variable in Excel using one of the two following formulas:

$$g_t = \frac{w_t - w_{t-1}}{w_{t-1}}$$

$$g_t = \ln(w_t) - \ln(w_{t-1})$$

   - Next, you would need to calculate the sample mean of the growth rates and the sample standard deviation of the growth rates using the Excel functions AVERAGE() and STDEV() and the series of growth rates.
   - Now you can calculate the test statistic using the following formula:

$$t^* = \frac{\bar{g} - .05}{\frac{s_g}{\sqrt{49}}}$$

     where $\bar{g}$ is the sample mean we calculated and s_g is the sample standard deviation we calculated. Notice that we are using 49 as our sample size even though we have wages for 50 years. We lose one observation when we do the growth rate calculation (we can't calculate a growth rate for the year 1950).
   - Now we need to calculate either our critical values or our p-value. Let's say we want to use critical values. We would calculate the critical values for a two-tailed test at the 10% significance level with 48 degrees of freedom. Excel will calculate this for us with TINV(.1,48).
   - We will reject the null hypothesis that the average growth rate in real wages is 5% if the absolute value of our t^* is greater than the absolute value of the critical value found in the previous part. If it is not, we fail to reject the null hypothesis.