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

January 27, 2010

Midterm 1

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 where appropriate 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

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

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

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

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(POTENTIALLY) USEFUL EXCEL OUPUT

TINV(.005,999)=2.81
TINV(.01,999)=2.58
TINV(.02,999)=2.33
TINV(.025,999)=2.24
TINV(.05,999)=1.96
TINV(.10,999)=1.65
TINV(.20,999)=1.28

TINV(.005,99)=2.87
TINV(.01,99)=2.63
TINV(.02,99)=2.36
TINV(.025,99)=2.28
TINV(.05,99)=1.98
TINV(.10,99)=1.66
TINV(.20,99)=1.29

SECTION I: MULTIPLE CHOICE (60 points)

1. The variance of the sample mean of X will always be:

   (a) Greater than or equal to the variance of X.
   (b) Less than or equal to the variance of X.
   (c) Equal to the variance of X.
   (d) None of the above.

   (b) The variance of the sample mean is equal to $\frac{\sigma^2}{n}$ where σ^2 is the variance of X. Since σ^2 is being divided by the sample size n (which is always greater than or equal to one) the variance of the sample mean will always be smaller than or equal to σ^2.

2. The nominal interest rate is equal to the real interest rate plus the inflation rate. If the mean of the real interest rate over the past year is positive and the mean of the inflation rate over the past year is positive, we can say for certain that:

   (a) The mean of the nominal interest rate over the past year is positive.
   (b) The variance of the nominal interest rate over the past year is greater than or equal to zero.
   (c) Neither (a) nor (b) is necessarily true.
   (d) Both (a) and (b) are true.

   (d) The mean will definitely be positive. The nominal interest rate is the sum of the real interest rate plus the inflation rate. So the mean of the nominal interest rate is equal to the mean of the real interest rate plus the mean of the inflation rate. These are both positive numbers so their sum will be positive. The variance of any variable is always greater than or equal to zero.

3. Which of the following is an example of panel data:

   (a) A dataset of income and educational attainment for 1,000 random individuals.
   (b) A dataset with ten years of monthly unemployment rates for each county in California.
   (c) A dataset giving current GDP for 100 different countries.
   (d) A dataset with daily caloric intake for one person over a six month period.

   (b) This dataset has observations for multiple time periods for multiple counties, making it panel data.

4. Suppose that X_t is a variable measuring monthly interest rates and $\tilde{X}_t$ is a five-month moving average of X_t. Which of the following would be the best way to look at just the high frequency fluctuations in the interest rates?

   (a) A line graph of X_t.
   (b) A line graph of $\tilde{X}_t$.
   (c) A line graph of $X_t + \tilde{X}_t$.
   (d) A line graph of $X_t - \tilde{X}_t$.

   (d) $\tilde{X}_t$ is capturing the long run trends in the interest rate. If we subtract this from X_t, we will be left with just the short run fluctuations.

5. The true population mean for the height of American males is 1.776 meters. Suppose that a researcher takes a random sample of 100 American males that has a sample mean is 1.655 meters and uses this information to reject the null hypothesis that the mean height in the population is greater than or equal to 1.700 meters. Which of the following can we say for certain:

   (a) The researcher has done the hypothesis testing incorrectly.
   (b) The researcher has committed a Type I error.
   (c) The researcher has committed a Type II error.
   (d) Both (a) and (b).

   (b) The researcher has rejected the null hypothesis when it is actually true. This is the definition of a Type I error. This can occur even if the researcher does all of the testing correctly if the sample mean just happened to be very low by chance.

6. Suppose that for a random variable Y, the mean of the distribution of Y is greater than the median of the distribution of Y. The distribution of the sample mean of Y will be:

   (a) Left skewed.
   (b) Right skewed.
   (c) Symmetric.
   (d) Either (a) or (b) could be true.

   (c) Whether or not the distribution of Y is skewed, the distribution of the sample mean of Y will be symmetric.

7. Which of the following would not affect the width of a confidence interval for the mean of X?

   (a) The standard deviation of the sample.
   (b) The mean of the sample.
   (c) The significance level chosen for the confidence interval.
   (d) All of the above would influence the width of the confidence interval.

   (b) The sample mean will determine where the confidence interval is centered but it will not affect the width of the interval. The significance level, sample size and sample standard deviation will all affect the width.

8. If you can reject the null hypothesis that $\mu = 10$ at the 5% signficance level, which of the following is definitely true?

   (a) You can reject the null hypothesis that $\mu = 10$ at the 1% significance level.
   (b) You can reject the null hypothesis that $\mu = 10$ at the 10% significance level.
   (c) You can reject the null hypothesis that $\mu \leq 10$ at the 10% significance level.
   (d) Both (b) and (c) are true.

   (b) If you can reject the null hypothesis at a particular significance level α you can reject that same null hypothesis at any significance level greater than α. Knowing that we can reject $\mu = 10$ at a 5% significance level does not tell us whether we can reject the hypothesis that $\mu \leq 10$ (if we rejected $\mu = 10$ because we had sample mean much smaller than 10 we definitely wouldn't reject the hypothesis that $\mu \leq 10$).

9. Suppose that the population growth rate is a constant 15% per year. Which of the following statements is true?

   (a) A graph with year on the horizontal axis and population on the vertical axis will have a slope of 15.
   (b) A graph with year on the horizontal axis and population on the vertical axis will have a slope of 0.15.
   (c) A graph with year on the horizontal axis and ln(population) on the vertical axis will have a slope of 15.
   (d) A graph with year on the horizontal axis and ln(population) on the vertical axis will have a slope of 0.15.

   (d) If a variable has a constant growth rate, a graph of the natural log of that variable will produce a straight line with a slope equal to the growth rate (as a decimal).

10. Suppose that we take a sample of incomes for 1,000 people. Which of the following is not realization of a random variable?

    (a) The mean income in the sample.
    (b) The sample size.
    (c) The variance of income in the sample.
    (d) All of the above are realizations of random variables.

    (b) We will get different numbers for the mean and variance if we draw different random samples. The sample size is just a constant in this scenario.

11. Suppose that we create a new variable Z that is equal to the variable X plus 10 (so $z_i = x_i + 10$ for observation i). Which of the following statements is definitely true?

    (a) The mean of Z is equal to the mean of X.
    (b) The standard deviation of Z is equal to the standard deviation of X.
    (c) The median of Z is equal to the median of X.
    (d) None of the above are definitely true.

    (b) Note that the distribution of Z will be identical to the distribution of X, just shifted to the right by 10 units. This will not change how spread out the data is; the distribution of Z will have the same standard deviation as the distribution of X.

12. Suppose that you are doing a univariate hypothesis test and calculate a p-value of 0.08. You would reject the null hypothesis if:

    (a) You are using a 5% significance level.
    (b) You are using a 10% significance level.
    (c) Both (a) and (b) are true if you are doing a two-tailed test.
    (d) Both (a) and (b) are true if your are doing a one-tailed test.

    (b) You reject the null hypothesis when the p-value is smaller than the significance level α. This is true whether you are doing a one-tailed or two-tailed test (the difference between the tests is in terms of how you calculate the p-value).

13. Suppose that 99% of the population in a country have annual incomes between \$0 and \$80,000. The other 1% of the population all have annual incomes over \$10,000,000. Which of the following is not true?

    (a) The mean income for the population is less than \$100,000.
    (b) The median income for the population is less than \$100,000.
    (c) The median income for the population is less than the mean income.
    (d) In a random sample of 100 people, the expected value of the sample mean is greater than \$100,000.

    (a) The mean income will definitely be greater than \$100,000. Even if 99% of the population earned \$0 and the rich 1% all earned exactly \$10,000,000, the mean income would be $0.99 \cdot \$0 + 0.01 \cdot \$10000000 = \$100000$. Given the the rich 1% earns even more than \$10,000,000 in annual income, the mean income will be even higher than \$100,000.

14. Increasing the significance level used for our hypothesis testing from 5% to 10% will:

    (a) Decrease the probability of a Type I error but have no effect on the probability of a Type II error.
    (b) Decrease the probability of a Type I error and increase the probability of a Type II error.
    (c) Decrease the probability of both a Type I and Type II error.
    (d) None of the above.

    (d) Increasing the significance level from 5% to 10% will double the probability of a Type I error.

15. Suppose that weight increases with age. We take a sample of 20-year-olds and use that sample to estimate the mean weight in the population of all people 20 years old and older. Which of the following statements is not true?

    (a) The sample mean will be an unbiased estimator of the population mean.
    (b) The sample variance will depend on the size of the sample we use.
    (c) The expected value of the sample mean will be smaller than the population mean.
    (d) The sample variance will be greater than or equal to zero.

    (a) The expected value of the sample mean will be equal to the population mean of 20-year-olds. But the population of interest is people 20 years or older. The mean of this population will be greater than the mean for just 20-year-olds. So the expected value of our estimator will not equal the population mean we are trying to estimate, making the estimator biased.

16. Which of the following would not decrease the probability of a Type II error?

    (a) A larger sample size.
    (b) A smaller variance of the variable of interest.
    (c) Using a smaller significance level.
    (d) All of the above would decrease the probability of a Type II error.

    (c) Using a smaller significance level would decrease the probability of a Type I error but would actually increase the probability of a Type II error by increasing the region over which we would fail to reject the null hypothesis.

17. The probability of getting a sample mean $\bar{x}$ that is greater than the median of x is:
    (a) Equal to 50%.
    (b) Greater than 50%.
    (c) Less than 50%.
    (d) Not enough information.

    (d) The distribution of the sample mean will be symmetric and centered at the mean of x. If the median of x is equal to the mean of x, then the probability of getting a sample mean greater than the median of x will be 50%. However, if the median of x is not equal to the mean of x, the probability of getting a sample mean greater than the median of x could be greater than or less than 50% (depending on whether the median of x is less than or greater than the mean of x).

18. A dataset of 1,000 voters that contains information on which presidential candidate people voted for in the last election is an example of:
    (a) Categorical data.
    (b) Cross-sectional data.
    (c) Both (a) and (b).
    (d) Neither (a) nor (b).

    (c) The data is cross sectional because it has observations for different individuals. It is categorical rather than numerical data (who somebody voted for does not have a natural numerical representation).

19. Suppose that we have a discrete random variable Z that can take on the values 1, 2, 3, 4 or 5. For a sample of 800 observations of Z, we plot a histogram with a separate bar for each of these five values. The value on the horizontal axis that corresponds to the highest point on this histogram is:
    (a) Equal to the mean of Z in the sample.
    (b) Equal to the median of Z in the sample.
    (c) Equal to the mode of Z in the sample.
    (d) None of the above are necessarily true.

    (c) The mode is the most frequent value in the dataset. The most frequent value will have be highest point on a histrogram whether the vertical axis is measuring absolute or relative frequencies.

20. Which of the following datasets would be the best choice for examining the differences in the standard of living over time in the United States?
    (a) Cross-sectional data on GDP.
    (b) Time series data on GDP.
    (c) Cross-sectional data on GDP per capita.
    (d) Time series data on GDP per capita.

    (d) Cross-sectional data would not be useful. It would show us variation across individuals but not over time. We would need time series data to look at changes over time. If we want to measure the standard of living, GDP per capita is a more relevant measure as it gives us a measure of income per person.

SECTION II: SHORT ANSWER (40 points)

1. (15 points) Suppose that there is a ballot initiative to overturn Proposition 13. In a sample of 1000 voters, 470 voters say they would vote to overturn Prop 13 and 530 voters say they would vote to keep Prop 13 intact.

   (a) Suppose we create a variable X that equals one if a person votes to overturn Proposition 13 and equals zero if the person votes to keep Proposition 13 intact. What is the mean and standard deviation of X in the 1000 person sample?

   The sample mean of X will be:

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

$$\bar{x} = \frac{1}{1000}(470 \cdot 1 + 530 \cdot 0)$$
$$\bar{x} = 0.47$$

   Since we are working with proportions data, we can use the formula for the standard deviation of proportions data to get s_x:

$$s_x = \sqrt{\bar{x}(1-\bar{x})}$$
$$s_x = \sqrt{0.47 \cdot (1 - 0.47)}$$
$$s_x = 0.499$$

   (b) Based on the information above, calculate a 90% confidence interval for the percentage of votes that will be cast on election day in favor of overturning Proposition 13.

   The formula for the confidence interval for the population mean of X is:

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$$

   Plugging in 0.10 for α and our values for $\bar{x}$, n and s from the previous part gives us:

$$0.47 \pm t_{0.05,999} \cdot \frac{0.499}{\sqrt{1000}}$$
$$0.47 \pm t_{0.05,999} \cdot 0.0158$$

   From the table of Excel output on the formula sheet, we can get $t_{0.05,999}$ by looking up $TINV(0.10, 999)$. This value is 1.646. Plugging in this last piece of information gives us the final confidence interval:

$$0.47 \pm 1.646 \cdot 0.0158$$

$$0.47 \pm 0.026$$

   So the 90% confidence interval for the percentage of votes that will be cast on election day in favor of overturning Proposition 13 is $(0.444, 0.496)$.

(c) Suppose that if the vote is tied, there will need to be another election. The California government needs to decide whether or not to print up a second set of ballots. Write down the set of hypotheses the government will use to determine whether they will need to print a second set of ballots. Your hypotheses should be written in term of X (or parameters relating to the distribution of X).

The government is trying to determine whether the election will be tied, in other words whether the population mean (μ) of X is 0.50. If the population mean is greater than or less than 0.50, new ballots will need to be printed. This implies that the government should use a two-tailed test (what is important is whether μ is equal to 0.50 or not equal to 0.50, it doesn't matter whether it's not equal because it is bigger or because it is smaller). So the null and alternative hypotheses will be the following:

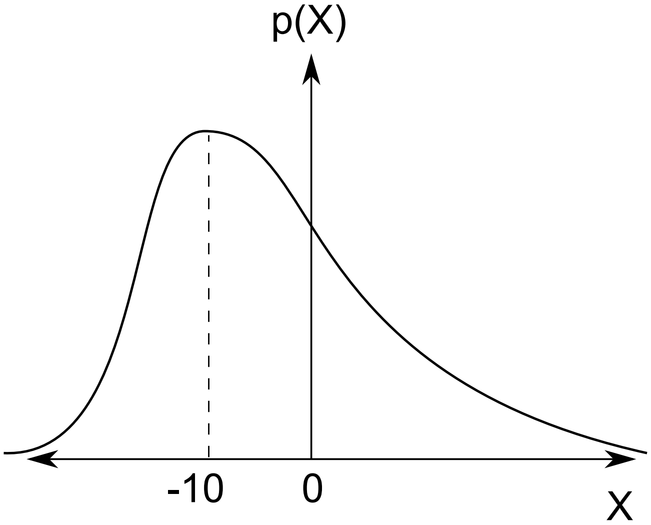
$$H_0\colon\ \mu = 0.50$$

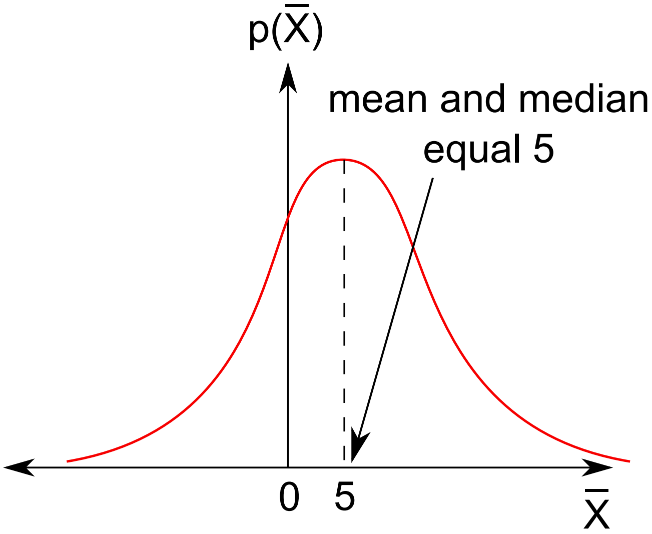
$$H_a\colon\ \mu \neq 0.50$$

(d) Use the information in the problem to test your hypotheses from part (c) using a 10% significance level. Be certain to show your work and clearly state your conclusions.

We have already found the 90% confidence interval for μ. This interval does not contain the value 0.50. This tells us that we can reject the null hypothesis that μ is equal to 0.50 at a 10% significance level. So we can reject the null hypothesis that the vote will be tied at a 10% significance level.

2. (10 points) The top graph below shows the distribution of the random variable X. The median of X is -5, the mean of X is 5 and the variance of X is 25. On the lower graph, draw the distribution for the sample mean of X. Label the mean and median of this distribution on the graph, including their numerical values if possible.





The sample mean has a normal distribution with a mean equal to the population mean of X (which is 5 in this case) and a variance equal to the population variance of X divided by the sample size. So you should have drawn a normal distribution centered at 5. Since the normal distribution is symmetric, the mean is equal to the median. So the median of the sample mean distribution is also 5.

3. (15 points) A researcher is testing for dangerous levels of lead in toys coming into the United States. A dangerous level of lead is considered any level greater than 600 parts per million (ppm). The researcher has a sample of 100 toys that she tests for lead levels and finds a mean level of 595 ppm and a standard deviation of 20 ppm.

   (a) Assume that the researcher wants to be cautious and error on the side of declaring toys dangerous. In other words, the researcher would rather assume the toys are dangerous and place the burden of proof on showing that the toys are safe. Write down the null and alternative hypotheses the researcher would use to test whether the toys are safe.

   The burden of proof falls on the alternative hypothesis. This researcher would rather fail to reject that the toys are dangerous than fail to reject that the toys are safe. So she will want to start with the null hypothesis that the toys are dangerous and see if the data can prove otherwise. This gives us the following set of hypotheses:

$$H_0\colon\ \mu \geq 600$$

$$H_a\colon\ \mu < 600$$

   (b) Given your hypotheses in part (a), which would the researcher consider worse, a Type I error or a Type II error? Explain your answer in no more than two sentences.

   In this case, a Type I error is worse. With a Type I error, we reject the null hypothesis even though the null hypothesis is true. In this case, a Type I error would mean declaring the toys safe even though the null hypothesis that they are dangerous is actually true.

   (c) Find the critical value the researcher will use to test the hypotheses from part (a) assuming she wants to use a 5% significance level.

   This is the critical value corresponding to a lower one-tailed test using a 5% significance level:

$$c = -t_{\alpha,n-1}$$

$$c = -t_{0.05,99}$$

   This value can be obtained from the Excel output on the formula sheet:

$$c = -t_{0.05,99} = -TINV(0.10, 99)$$

$$c = -1.66$$

   Note that we had to double our α when using the TINV function because we want all of the probability to be in one tail of the t distribution and Excel is automatically splitting it between the two tails of the distribution.

   (d) Use the critical value you found in part (c) and the information given in the problem to test your hypothesis from part (a), being certain to formally state your conclusions.

   Now that we have the critical value, all we need to do is calculate our test statistic and compare it to the critical value:

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

$$t^* = \frac{595 - 600}{\frac{20}{\sqrt{100}}}$$

$$t^* = -2.5$$

Our t statistic is to the left of the critical value, so we will reject the null hypothesis that the toys have dangerously high levels of lead in favor of the alternative hypothesis that the toys have a safe level of lead.