Ecn 102 - Analysis of Economic Data
University of California - Davis January 28, 2010
Instructor: John Parman

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 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

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

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

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

(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. Suppose we take a sample of 400 people and measure their heights. If we measure height in inches, the coefficient of variation will be:
   (a) Larger than if we measure height in meters.
   (b) Smaller than if we measure height in meters.
   (c) The same as when we measure height in meters.
   (d) (a), (b) or (c) could be true depending on the value of the sample mean.

   (c) Changing the unit of measurement will rescale the sample mean and the sample standard deviation by the same factor. When dividing the new sample mean by the new standard deviation, this scaling factor will cancel out and we will be left with the original coefficient of variation.

2. Which of the following is not a random variable?
   (a) The sample mean.
   (b) The sample variance.
   (c) The population variance.
   (d) All of the above are random variables.

   (c) The population variance is a constant parameter. The sample mean and sample variance can take on different values for different samples.

3. Suppose we can reject the null hypothesis that $\mu \geq 50$ at the 5% significance level. Which of the following statements is definitely true?
   (a) We can reject the null hypothesis that $\mu = 50$ at the 5% significance level.
   (b) We can reject the null hypothesis that $\mu \geq 50$ at the 10% significance level.
   (c) The value we obtained for t^* was positive.
   (d) The value we obtained for t^* was greater than the critical value for an upper one-tailed hypothesis test at the 5% significance level.

   (b) If we rejected the null hypothesis that $\mu \geq 50$ at the 5 significance level, our value for t^* must have been less than $-t_{.05,n-1}$. The magnitude of $t_{.1,n-1}$ is smaller than the magnitude of $t_{.05,n-1}$ so it will also be the case that $t^* < -t_{.1,n-1}$.

4. The sample variance of 100 observations of monthly unemployment rates will be _____ the sample variance of 100 observations of the unemployment rate covering the same time period for which each month's unemployment rate is an average of that month's rate and the previous two months' rates.
   (a) Less than or equal to.
   (b) Less than.
   (c) Equal to.
   (d) Greater than or equal to.

   (d) In general, using a moving average will tend to reduce the amount of variance in a data series. There are some special cases where it will leave the variance unchanged (for example, think about a sample with zero variance).

5. Which of the following is not a measure of central tendency?
   (a) The sample range.
   (b) The sample median.
   (c) The sample midrange.
   (d) The sample mean.

   (a) The sample range is a measure of dispersion.

6. Suppose we have data on the interest rate, i_t. Which of the following could we use to get the percent change in the interest rate from year t to year $t+1$ (expressed as a decimal)?
   (a) $i_{t+1} - i_t$.
   (b) $ln(i_{t+1} - i_t)$.
   (c) $\frac{i_{t+1}}{i_t}$.
   (d) $ln(i_{t+1}) - ln(i_t)$.

   (d) Percent changes in a variable can be approximated by a difference logs.

7. Increasing the sample size used for a two-tailed hypothesis test will tend to:
   (a) Decrease the probability of a Type I error.
   (b) Decrease the probability of a Type II error.
   (c) Increase the probability of a Type I error.
   (d) Increase the probability of a Type II error.

   (b) The probability of a Type I error is equal to the value of the significance level α. Increasing the sample size will change the critical values that correspond to the significance level of α and the range of sample means for which we would get a Type I error but the probability of a Type I error will still be α. The probability of a Type II error can be reduced by increasing the sample size which will decrease the variance of the distribution of the sample mean.

8. Suppose that we have data on coin flips. The variable X is equal to one if the coin flip is heads and zero if the coin flip is tails. In a sample of 100 coin flips, the sample mean of X turns out to be .6. The value of the skewness for the sample will be:
   (a) Positive.
   (b) Negative.
   (c) Zero.
   (d) Not enough information.

   (b) The sign of the skewness will depend on the sign of $\sum(x_i - \bar{x})^3$. Note that for 60 of our observations, the value of $(x_i - \bar{x})^3$ is $(1 - .6)^3$, or .064. For the other 40 observations, the value of $(x_i - \bar{x})^3$ is $(0 - .6)^3$, or -.216. So $\sum(x_i - \bar{x})^3$ is equal to $60 \cdot .064 + 40 \cdot -2.16$ which is -4.8.

9. Which of the following statements is not true?
   (a) Two random variables can have the same mean but different medians.
   (b) Two random variables can have the same mean but different variances.
   (c) The mean of the sum of two random variables is always equal to the sum of their means.

(d) The variance of the sum of two random variables is always equal to the sum of their variances.

(d) The variance of the sum of two random variables will be equal to the sum of their variances plus an additional term dependent on the covariance of the two variables.

10. If the distribution of X is symmetric and reaches its highest point at the exact middle of the distribution:

(a) The mode of X will be equal to the median of X.
(b) The mean of X will be equal to the median of X.
(c) The median of X will be at the exact middle of the distribution.
(d) All of the above are true.

(d) If the distribution is symmetric, the mean and median will both be at the exact middle of the distribution. If the highest point of the distribution is at the exact middle, then the mode is at the middle of the distribution.

11. Which of the following would increase the probability of a Type I error?

(a) Decreasing the sample size.
(b) Increasing the sample size.
(c) Decreasing the significance level.
(d) Increasing the significance level.

(d) The probability of a Type I error is equal to the significance level. Increasing the significance level will increase the probability of Type I error.

12. Suppose that the 95% confidence interval for the population mean hours of studying per week is $(6.5, 6.9)$. Which of the following statements is true?

(a) You would reject the null hypothesis that $\mu \geq 6.5$ at the 10% significance level.
(b) You would reject the null hypothesis that $\mu = 7.1$ at the 5% significance level.
(c) You would reject the null hypothesis that $\mu \leq 6.9$ at the 5% significance level.
(d) All of the above are true.

(b) Note that 7.1 falls outside of the 95% confidence interval. That tells us that we would reject the null hypothesis that $\mu = 7.1$ at a 5% significance level.

13. Suppose you have a dataset of the unemployement rate for 100 different cities for the month of December, with one observation per city. These are:

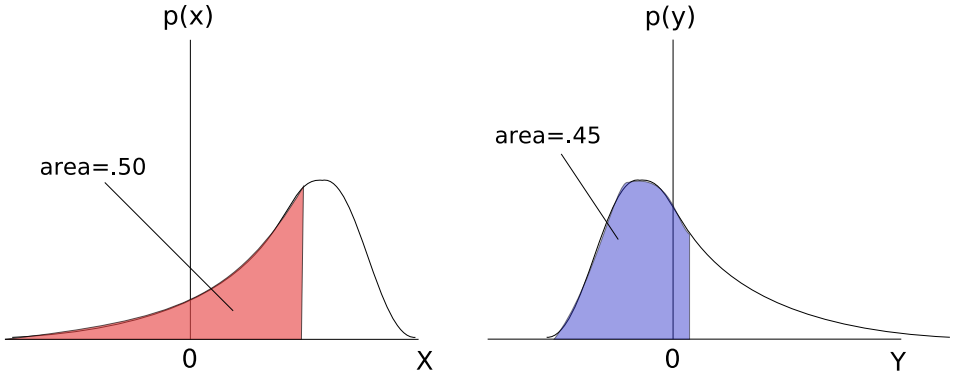
(a) Cross-sectional data.
(b) Panel data.
(c) Time-series data.
(d) Both (a) and (c).

(a) These are cross-sectional data. We are observing a cross-section of cities at a single point in time.

14. Suppose that rather than the sample mean, we use another statistic that we'll call $\tilde{X}$ to estimate the population mean. If the distribution of $\tilde{X}$ is symmetric, what must be true for $\tilde{X}$ to be an unbiased estimator for the population mean?

    (a) The distribution of $\tilde{X}$ must get narrower as the sample size increases.
    (b) The variance of $\tilde{X}$ should not depend on the sample size.
    (c) The distribution of $\tilde{X}$ must be centered at the population mean.
    (d) The value of $\tilde{X}$ should approach the population mean as the sample size goes to infinity.

    (c) To be an unbiased estimator of the population mean, the expected value of $\tilde{X}$ has to be equal to the population mean.



Use the figure above to answer questions 15 through 17. The graph on the left shows the distribution of random variable X and the graph on the right shows the distribution of random variable Y.

15. The median of X is _____ and the median of Y is _____.

    (a) Positive, positive.
    (b) Positive, negative.
    (c) Negative, negative.
    (d) Not enough information.

    (a) The median is the value at which the area under to curve to the left of the value is 50%. From the graphs, we can tell that this is a positive value in both cases.

16. Which of the following statements is true?

    (a) $skewness_x > skewness_y$.
    (b) $skewness_y > skewness_x$.
    (c) $skewness_x < 0$ and $skewness_y < 0$.
    (d) $skewness_x = 0$.

    (b) The distribution of x is left skewed and would have a negative value for skewness. The distribution of y is right skewed and will have a positive value for the skewness. So $skewness_y > skewness_x$.

17. The distribution of the sample mean of Y will be:
    (a) Right skewed.
    (b) Left skewed.
    (c) Symmetric.
    (d) Centered at 0.

    (c) The sample mean will be distributed normally and centered at the mean of Y (which is not necessarily equal to zero).

18. Which of the following would narrow the confidence interval for the population mean?
    (a) Decreasing the sample size.
    (b) Decreasing the significance level.
    (c) Both (a) and (b) would narrow the confidence interval.
    (d) Neither (a) nor (b) would narrow the confidence interval.

    (d) Decreasing the sample size or decreasing the size of α will both increase the width of the confidence interval for the population mean.

19. The distribution of the sample mean of X (using a sample size of 10) has a:
    (a) Larger variance than the distribution of X.
    (b) Smaller variance than the distriubution of X.
    (c) Variance equal to the variance of X.
    (d) None of the above.

    (b) The variance of the distribution of the sample mean is $\frac{\sigma^2}{n}$ which is smaller than the variance of the distribution of x (σ^2).

20. The probability of a Type I error when doing a two-tailed hypothesis test at a 5% significance level is:
    (a) Equal to the probability of a Type I error when doing a one-tailed hypothesis test at a 2.5% significance level.
    (b) Equal to the probability of a Type I error when doing a one-tailed hypothesis test at a 5% significance level.
    (c) Equal to the probability of a Type I error when doing a one-tailed hypothesis test at a 10% significance level.
    (d) Equal to the probability of a Type I error when doing a one-tailed hypothesis test at a 20% significance level.

    (b) The probability of a Type I error is equal to the value of α whether we are doing a one-tailed or two-tailed test.

SECTION II: SHORT ANSWER (40 points)

1. (14 points) A survey is given to 1000 college students asking them how many years they think it will take them to graduate from college. Responses ranged from three years to seven years. The distribution of students by their responses is given in the table below:

Years to graduate	Number of students
3	50
4	400
5	300
6	200
7	50

   (a) Calculate a 90% confidence interval for the proportion of students in the population who believe they will graduate in exactly four years.

   First we need to know the proportion of students in the sample that believe they will graduate in exactly four years:

$$\bar{x} = \frac{400}{50 + 400 + 300 + 200 + 50}$$

$$\bar{x} = \frac{400}{1000} = .4$$

   Now we can use our shortcut for proportions data to get the sample standard deviation:

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

$$s^2 = .4(1 - .4) = .24$$

$$s = .4899$$

   Now we have everything we need to calculate a confidence interval:

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

$$.4 \pm t_{\frac{.10}{2},999}\frac{.4899}{\sqrt{1000}}$$

   The value of $t_{.05,999}$ is given by TINV(.10,999) which is 1.65. So our confidence interval is:

$$.4 \pm 1.65 \cdot \frac{.4899}{\sqrt{1000}}$$

$$.4 \pm .0256$$

$$(.3744, .4256)$$

   (b) Suppose a researcher wants to use these data to create a 95% confidence interval for the average number of years it takes students to graduate from college. At what value would this confidence interval be centered?

The confidence interval will be centered at the sample mean of the number of years to graduate:

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{x} = \frac{1}{1000}(50 \cdot 3 + 400 \cdot 4 + 300 \cdot 5 + 200 \cdot 6 + 50 \cdot 7)$$

$$\bar{x} = 4.8$$

So the confidence interval will be centered at 4.8 years.

(c) Why might the researcher reach an incorrect conclusion? Fully explain your answer including whether the researcher would be likely to overestimate or underestimate the average number of years it takes students to graduate.

The variable we are observing is not the variable the researcher is trying to make inferences about. We observe the number of years students think it will take them to graduate, not the number of years it will actually take them to graduate. Students most likely don't account the possibility of bad events occurring that would delay graduation (failing classes, family emergencies, etc.). If this is the case, students expected years to graduate will tend to be lower than the actual years it takes to graduate. The sample mean from the data above would therefore provide an underestimate of the population mean of interest to the researcher.

2. (14 points) The true population mean of a random variable X is 50. Suppose that you draw a sample of 100 observations of X that has a sample standard deviation of 10 and you use this to test the following set of hypotheses using a 5% significance level:

$$H_o\text{: } \mu = 55$$

$$H_a\text{: } \mu \neq 55$$

   (a) Write down a formula for the test statistic you would use for the test. The only variable in your expression should be the sample mean, $\bar{x}$ (you should plug in the appropriate numerical values for any other variables or constants).

   The formula for the test statistic is:

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

   Plugging in our values gives us:

$$t^* = \frac{\bar{x} - 55}{\frac{10}{\sqrt{100}}}$$

$$t^* = \bar{x} - 55$$

   (b) For what range of values of $\bar{x}$ would you end up committing a Type II error?

   We will commit a Type II error if we fail to reject the null hypothesis even though the null hypothesis is false. So we need to figure out the range of values for $\bar{x}$ for which we would not reject the null hypothesis. To do this, we first need to figure out the critical values we would use. For a two-sided test with a 5% significance level, the critical values would be given by:

$$c_{upper} = t_{\frac{\alpha}{2},n-1} = t_{.025,99}$$

$$c_{lower} = -c_{upper} = -t_{.025,99}$$

   The value for $t_{.025,99}$ is given by TINV(.05,99) and is 1.98. So we will fail to reject the null hypothesis if t^* is between -1.98 and 1.98. This combined with our equation for t^* above allows us to figure out the range of $\bar{x}$ for which we will commit a Type II error:

$$c_{lower} < t^* < c_{upper}$$

$$-1.98 < \bar{x} - 55 < 1.98$$

$$53.02 < \bar{x} < 56.98$$

   So we will make a Type II error whenever the sample mean is between 53.02 and 56.98.

3. (12 points) For each scenario below, write down the null and alternative hypotheses you would use, the formula for the test statistic you would use, the critical value you would use, and the basis on which you would reject the null hypothesis. For the test statistic and the critical value, include specific numbers whenever possible.

   (a) You have a sample of 100 books. You want to test whether or not the average number of pages in a book is greater than 200 pages using a 5% significance level.

   If you choose to use an upper one-tailed test:

   $$H_o\text{: } \mu \leq 200$$

   $$H_a\text{: } \mu > 200$$

   $$t^* = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 200}{\frac{s}{10}}$$

   $$c = t_{\alpha,n-1} = t_{.05,99} = TINV(.1, 99) = 1.66$$

   reject the null if $t^* > c$

   If you choose to use a lower one-tailed test:

   $$H_o\text{: } \mu \geq 200$$

   $$H_a\text{: } \mu < 200$$

   $$t^* = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 200}{\frac{s}{10}}$$

   $$c = -t_{\alpha,n-1} = -t_{.05,99} = -TINV(.1, 99) = -1.66$$

   reject the null if $t^* < c$

   (b) You take the temperatures of 100 people to test whether the average body temperature is equal to 98.6 degrees. You want to use a 5% significance level.

   $$H_o\text{: } \mu = 98.6$$

   $$H_a\text{: } \mu \neq 98.6$$

   $$t^* = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 98.6}{\frac{s}{10}}$$

   $$c = t_{\frac{\alpha}{2},n-1} = t_{.025,99} = TINV(.05, 99) = 1.98$$

   reject the null if $|t^*| > c$

   (c) You have a sample of 100 people's wages. You want to use a lower one-tailed test whether the average wage is below $20 an hour using a 10% significance level.

   $$H_o\text{: } \mu \geq 20$$

   $$H_a\text{: } \mu < 20$$

$$t^* = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 20}{\frac{s}{10}}$$

$$c = -t_{\alpha,n-1} = -t_{.10,99} = -TINV(.2, 99) = -1.29$$

reject the null if $t^* < c$