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

January 13, 2010

Problem Set 2 - Solutions

This problem set will be not be graded and does not need to be turned in. However, it will provide a good review for material that will appear on the first midterm and good practice with Excel skills that you will continue to use later in the course. You are strongly encouraged to work through the entire problem set.

---

1. **Formulating a Hypothesis** For each scenario below, write down the most appropriate null and alternative hypothesis you would use for statistical inference on the population mean. Also write down the critical value you would use for your hypothesis testing if you want to use a five percent significance level.

   (a) It has been determined that mercury levels higher than two parts per billion in drinking water are unsafe. You are responsible for determining whether Davis drinking water is safe based on 100 water samples taking from different parts of Davis.

   $$H_o\text{: } \mu \geq 2$$

   $$H_a\text{: } \mu < 2$$

   $$c = -t_{.05;99}, \text{ reject null if } t^* < c$$

   We clearly want a one-sided test in this case. Whether we want to use an upper-tailed test or a lower-tailed test depends on whether we want to make it harder to declare the water safe or harder to declare the water unsafe. We probably want to error on the side of failing to reject that the water is unsafe so we will set up our null hypothesis as the water having mercury levels at or exceeding the cutoff value. The alternative is then that the water is safe, the mercury level is less than the cutoff level. Note that if we set it up the other way (null hypothesis is that the mercury level is less than or equal to the cutoff) we could have a situation where the sample had a dangerous level of mercury but we still fail to reject the null hypothesis that the water is safe.

   (b) A cookie company has determined that it is most profitable to have 15 chocolate chips in each cookie. If there are more than 15, the cookies become too expensive to produce. If there are less than 15, the company loses customers. They want to test whether they are operating at the efficient level by looking at a sample of 500 cookies.

   $$H_o\text{: } \mu = 15$$

$$H_a\text{: } \mu \neq 15$$

$$c = t_{.025;499}, \text{ reject null if } |t^*| > c$$

(c) A pollster wants to determine whether a majority of voters would vote for a constitutional convention for California. The pollster has the opinions of 1000 randomly sampled voters.

$$H_o\text{: } \mu \leq 50$$

$$H_a\text{: } \mu > 50$$

$$c = t_{.05;999}, \text{ reject null if } t^* > c$$

2. **Hypothesis Testing with Continuous Data** For this question, use the Yolo county census data available on Smartsite (cens00-yolo.csv). These data are a 5% sample of the Yolo county population aged 16 to 65 from the year 2000. HOURS gives the individual's average number of hours worked per week in the previous year. INCTOT gives the individual's total annual income for the previous year.

(a) Drop any observations for which hours worked last week are zero or total income is zero.

This can be done by first sorting the data by hours, selecting the chunk of observations for which hours are zero and deleting them, and then sorting by income and dropping those observations for which income is zero.

(b) Using the remaining observations, test the following hypothesis about the mean annual income μ using a significance level of 10%:

$$H_o : \mu \leq \$35000$$

$$H_a : \mu > \$35000$$

All calculations are shown in *ps2-solutions.xlsx*. To test the hypothesis, we first need to calculate our test statistic by computing the sample mean, sample standard deviation and number of observations and then plug those values into the formula for t^*. Next we need to either calculate the p value or the critical value. Both approaches are shown in the spreadsheet. As you can see in the spreadsheet, or p value is greater than .10 and our t^* is less than the critical value, so we fail to reject the null hypothesis at a 10% significance level.

(c) If you want to test whether the mean Yolo county annual income is equal to \$34,000, what would your p-value be?

See the calculations in *ps2-solutions.xlsx*.

(d) Based on your answer to part (c), what is the lowest significance level at which you would reject the null hypothesis that the mean annual income is equal to $34,000?

Our p value turned out to be .42. So we would only reject the null hypothesis if we used an α greater than .42.

(e) Calculate a 90% confidence interval for the population mean of annual income.

All calculations are shown in *ps2-solutions.xlsx*. The 90% confidence interval for the mean income is ($30720,$43564). So with a probability of 90% the true mean income for Yolo county is between $30720 and $43564.

(f) How would you expect your answers to the previous questions to change if you had a 10% sample of Yolo county residents rather than a 5% sample?

Switching to a 10% sample would double our sample size. This increase in sample size would narrow the distribution of the sample mean making us more likely to reject any of the null hypotheses if they were false. It would also narrow our confidence interval for the sample mean.

3. **Hypothesis Testing with Proportions Data** For this question, use the Gallup poll data on whether or not people favored statehood for Alaska and Hawaii (statehood-poll.csv). This was a poll taken in the 1950's and was meant to be a random sample of the adult population of the United States.

(a) Suppose that the government decided to grant statehood only if 70% of the population favored statehood. Based on the poll data, use an upper one-tailed test to determine whether Alaska should receive statehood. At what significance levels would Alaska be granted statehood?

See ps2-solutions.xlsx for the calculations. Note the use of the COUNTIF() command. You could also create a new variable that was equal to 1 for people who favored statehood and 0 for all other cases (no opinion or don't favor) and then take the average of this new variable to get the percentage of people favoring statehood. The p-value for an upper one-tailed test turns out to be .0026. So Alaska would be granted statehood at significance levels as small as .3%.

(b) Do the same for Hawaii. Are there any significance levels at which Hawaii would be granted statehood?

See ps2-solutions.xlsx for the calculations. The value of t^* turns out to actually be negative (less than 70% of the people in the sample favored statehood). With a negative t-stat, there is no reasonable significance level at which we'll reject the null hypothesis for an upper one-tailed test.

(c) Calculate a 95% confidence interval for the percentage of people who favor statehood for both Hawaii and Alaska. (Hint: You may need to construct a new variable to do this.)

See ps2-solutions.xlsx for the calculations. Note the use of the COUNTIFS() function. An alternative approach would be to create a new variable that equals one if a person favors statehood for both Alaska and Hawaii and equals zero in all other cases. This would be a more time consuming approach but you may find it simpler.

4. **Type I and Type II Errors**

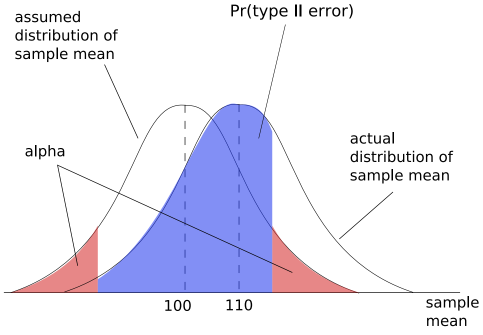
(a) Describe a situation in which a researcher would be very concerned about Type I errors. Would the researcher choose a large or small value for α in this situation?

There are countless answers for this question. In general, it would be a situation where rejecting the null hypothesis even though it is true is a very costly mistake to make. For specific examples, see the lecture notes and the practice midterm solutions. In these situations, a researcher would choose a small value for α.

(b) Describe a situation in which a researcher would be very concerned about Type II errors. Would the researcher choose a large or small value for α in this situation?

There are also countless correct answers for this question. In general, your answer should describe a situation in which failing to reject the null hypothesis when it is false is a very costly mistake. Specific examples are given in the lecture notes and the solutions to the practice midterm. In this situation, a researcher would opt for a larger value for α.

(c) Draw a graph showing the probability of a Type II error when μ_0 is 100 but the true population mean is 110.


assumed distribution of sample mean
Pr(type II error)
alpha
actual distribution of sample mean
100
110
sample mean

The blue area on the graph above gives the probability of Type II error. This probability depends both on our choice of the significance level α and how far the true population mean is from our guess of μ_0.

5. **Univariate Data Transformation**

(a) Using the GDP data from Problem Set 1, construct a graph that shows the growth rate of real GDP over time and a three-year moving average of the growth rate of GDP over time. Does the three-year moving average effectively smooth the graph? What happens if you switch to a seven-year moving average?

See ps2-solutions.xlsx for the calculations and the graph. Note that I have centered the moving average on the year of interest. In other words, the three-year moving average for the GDP growth rate in 1990 is equal to the average of the growth rates in 1989, 1990 and 1991. You can also construct moving averages in other ways. For example, you will sometimes see the average taken over the previous years (the averaged value for 1990 would be the average of the growth rates in 1990, 1989 and 1988). You can see from the graph that the moving average smooths out the data series and the longer the average is, the more it smooths the data. Because of this a longer moving average can make it easier to see long run trends but can cause you to lose some of the short run fluctuations in the data.

(b) Suppose that we have data on daily temperatures for a period of six months. The temperatures are given in degrees celcius. The data have a mean of 15 degrees celcius, a standard deviation of 4 degrees celcius and a range of 18 degrees celcius. Suppose that we convert the data into degrees fahrenheit. What will the new mean, standard deviation and range be? Are there any of the measures of central tendency and dispersion that we've discussed that would not change?

To understand how the statistics will be affected, we need to think of how the random variable giving temperature in degrees fahrenheit (we'll call this variable F) is related to the random variable giving temperature in degrees celcius (we'll call this variable C). The relationship between these two variables is simple:

$$f_i = \frac{9}{5}c_i + 32$$

We can use our formulas for the various descriptive statistics to see how they will change once we convert from celcius to fahrenheit:

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$$

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n}(\frac{9}{5}c_i + 32)$$

$$\bar{f} = \frac{1}{n}(\sum_{i=1}^{n}(\frac{9}{5}c_i) + \sum_{i=1}^{n} 32)$$

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n}(\frac{9}{5}c_i) + \frac{1}{n}\sum_{i=1}^{n}32$$

$$\bar{f} = \frac{9}{5}\bar{c} + 32$$

$$\sigma_f^2 = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})^2$$

$$\sigma_f^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{9}{5}c_i + 32 - \frac{9}{5}\bar{c} - 32)^2$$

$$\sigma_f^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{9}{5})^2(c_i - \bar{c})^2$$

$$\sigma_f^2 = (\frac{9}{5})^2\sigma_c^2$$

$$range_f = f_{max} - f_{min}$$

$$range_f = \frac{9}{5}c_{max} + 32 - \frac{9}{5}c_{min} - 32$$

$$range_f = \frac{9}{5}(c_{max} - c_{min})$$

$$range_f = \frac{9}{5}range_c$$

From the equations above, we can see that the mean will be 59 degrees fahrenheit, the standard deviation will be 7.2 degrees fahrenheit, and the range will be 32.4 degrees fahrenheit. Of the statistics we discussed in class, all of the central tendency measures would change (we are shifting the entire distribution by converting to fahrenheit). For the dispersion measures, all of them would change change as well. You might expect the coefficient of variation to stay the same because it is unitless. The problem here is that we are both rescaling the data (which by itself would not change the coefficient of variation) and shifting the distribution. The standard deviation is affected by the rescaling (the factor $\frac{9}{5}$) while the sample mean is affected by both this rescaling and the shift (the addition of 32). This means that the ratio of the standard deviation to the mean will change.