Economics 30330: Statistics for Economics
Problem Set 7 - Suggested Solutions
University of Notre Dame
Instructor: Julio Garín
Spring 2012

Confidence Intervals (80 points)

1. In a sample of 22 undergraduates, a researcher finds the average student consumes 3 beers a week. Assume the population standard deviation is known and and is equal to 2.

   (a) Construct a 95 percent confidence interval around the estimate for the sample mean.
   We know σ, therefore we should use the z-table. The 95% confidence interval will be given then by,
   $$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$
   $$\left(3 - 1.96\frac{2}{\sqrt{22}}, 3 + 1.96\frac{2}{\sqrt{22}}\right)$$
   Therefore, $\mu \in (2.16, 3.84)$ with 95% of probability.

   (b) What does the 80 percent confidence interval look like?
   Now, with 80% of probability $\mu \in \left(\bar{x} \pm z_{0.2/2}\frac{\sigma}{\sqrt{n}}\right)$. Therefore, since the z-value is in this case 1.28, the confidence interval for the population mean, μ, becomes:
   $$(2.45, 3.55)$$

2. The Tornado Fuel Saver is a device that can be installed in the air-intake of your vehicle's engine that is advertised to improve gas mileage and horsepower. (You may have seen an infomercial for the Tornado Fuel Saver on late-night TV! This product is real!). The manufacturer advertises that the Tornado Fuel Saver can boost your car's miles per gallon by 28 percent. A car magazine recently conducted a test of the Tornado Fuel Saver and installed the device on 25 cars. The magazine found that miles per gallon improved by an average of 18 percent in this sample. The standard deviation of this estimate is an 8 percent improvement. From this sample of 25, construct a 95% confidence interval for the expected percent improvement in miles per gallon generated by installing the fuel saver.

   In this case we do not know the population variance, σ^2. However, we can use the fact that we know the sample standard deviation. Due to the binomial approximation to a normal that we covered in class, the distribution can be approximated by a z-distribution. So the confidence interval can be constructed as,
   $$\left(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$
   $$\left(\bar{x} - z_{0.025}\frac{s}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{s}{\sqrt{n}}\right)$$
   $$\left(18 - 1.96\frac{8}{\sqrt{25}}, 18 + 1.96\frac{8}{\sqrt{25}}\right)$$
   Therefore, the 95% confidence interval for the expected percent improvement in MPG generated by installing the fuel saver, μ, is $(14.86, 21.13)$.

3. A pharmaceutical company has developed a new drug that it believes will reduce cholesterol levels in some patients. In a preliminary study of 22 subjects, the company finds that average cholesterol levels fell by 18 points and with a standard deviation equal to 12. Construct a 95% confidence interval for the mean value of cholesterol reduction, assuming σ is unknown.

   As in the previous case we do not know the population variance, σ^2. While $np \leq$, we know that the distribution can be approximated by a t-distribution with $n-1$ degrees of freedom. So the confidence interval is now

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

$$\left(\bar{x} - t_{0.025,21}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025,21}\frac{s}{\sqrt{n}}\right)$$

$$\left(18 - 2.08\frac{12}{\sqrt{22}}, 18 + 2.08\frac{12}{\sqrt{22}}\right)$$

   Therefore, the 95% confidence interval for the mean value of cholesterol reduction is given by (12.68,23.32).

4. A survey of 101 Notre Dame undergraduates finds that 40 percent of the respondents reported having a "binge drinking" episode over the past month (binge drinking is defined as 5 or more drinks in a row on one occasion).

   (a) Construct a 99% confidence interval for this estimate.

   Since in this case we have a proportion, $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$, and from the sample size, we know the distribution of $\bar{p}$, can be approximated by a normal distribution. The confidence interval surrounding the population proportion can be constructed then as follows

$$\left(\bar{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \bar{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$\left(\bar{p} - z_{.005}\sqrt{\frac{p(1-p)}{n}}, \bar{p} + z_{.005}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$\left(0.40 - 2.57\sqrt{\frac{0.4(0.6)}{101}}, 0.40 + 2.57\sqrt{\frac{0.4(0.6)}{101}}\right)$$

   The confidence interval is then given by $(0.2745, 0.5253)$.

   (b) Nationally, 44% of college students report having a binge drinking episode. Suppose this number is known with certainty. Are ND students different from the national average?

   Notice that the 44% national average is within the interval. Therefore, although ND students consume less alcohol than the national average, they are statistically indistinguishable from that group.

5. A cell phone manufacturer advertises that the battery in its cell phone will allow for 90 minutes of talk time before running out. To test this claim, a famous consumer magazine ran a test on 16 fully charged cell phones and found that, on average, the phones ran out after only 84.5 minutes of talk time. The standard deviation of this estimate is 12 minutes. From this sample of 16, Construct a 90% confidence interval for the expected minutes of talk time.

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

$$\left(\bar{x} - t_{0.05,15}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.05,15}\frac{s}{\sqrt{n}}\right)$$
$$\left(84.5 - 1.753\frac{12}{\sqrt{16}}, 84.5 + 1.753\frac{12}{\sqrt{16}}\right)$$

Hence, the 90% confidence interval for the expected minutes of talk time, μ, is $(79.2, 89.8)$.

6. Below is some text from the October 3, 2011 *Slate* article "Poll: Most Don't Know What 'GOP' Stands For".

   > Do you know what GOP stands for? Its not a trick question, though apparently it is a tricky question for the majority of Americans.
   >
   > In a 60 Minutes/Vanity Fair poll released Sunday, just 45 percent correctly answered a multiple-choice question about the meaning of the widely used acronym for the Republican Party. (For the record, its "Grand Old Party.")
   >
   > More than a third guessed "Government of the People," while a few jokesters (one hopes) opted for "Grumpy Old People," "Gods Own Party," or "Gauntlet of Power." Fewer than one in 10 had the modesty to select "Dont know."
   >
   > Members of the GOP were marginally more likely to know what it means, with 51 percent of Republicans getting the right answer. Just 38 percent of Dems–er, Democrats–got it, though that may be partly because 9 percent of them opted for "Grumpy Old People." [...]
   >
   > The poll was based on a random sample of 1,165 adults across the country, with a margin of error of plus or minus 3 percent.

   (a) Using information in the text, recalculate the poll's margin of error.

   The margin of error will be equal to $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$. Since we do not know p, we should use $p = 0.5$. Therefore the margin of error, MOE, becomes:

$$\begin{aligned} MOE &= z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \\ &= z_{0.025}\sqrt{\frac{0.5(0.5)}{1165}} \\ &= 1.96(0.01464) \\ &= 0.02869 \end{aligned}$$

   (b) What is wrong with the estimated margin of error as described in the text?

   The margin of error is approximately equal to 3 percentage points, not 3 percent.

7. Suppose we have

$$\sum_{i=1}^{2}(Y_i - \hat{\beta}_1 X_i)^2$$

   which we will be calling, **sum of squared residuals (SSR)** (you will see why later on). Note that that sum is a function of the coefficient $\hat{\beta}_1$, or $f(\hat{\beta}_1)$. *Hint: this problem requires you to make use of standard calculus techniques.*

(a) Find the coefficient $\hat{\beta}_1$ that minimize the SSR. Put differently, you should solve the following problem:

$$\min_{\hat{\beta}_1} f(\hat{\beta}_1)$$

or, equivalently,

$$\min_{\hat{\beta}_1} \sum_{i=1}^{2} (Y_i - \hat{\beta}_1 X_i)^2$$

$$\begin{aligned}
\frac{df(\hat{\beta}_1)}{d\hat{\beta}_1} &= 0 \\
&= -2(Y_1 - \hat{\beta}_1 X_1)X_1 - 2(Y_2 - \hat{\beta}_1 X_2)X_2 \\
&= -Y_1 X_1 - Y_2 X_2 + \hat{\beta}_1(X_1^2 + X_2^2) \\
&= 0
\end{aligned}$$

Solving the previous equation for $\hat{\beta}_1$ we have[1]

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{2} Y_i X_i}{\sum_{i=1}^{2} X_i^2}$$

(b) Now, suppose that the function also has a constant term:

$$\sum_{i=1}^{2} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

In this part, you are required to find the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize that sum, which can also be written as function of those coefficients: $f(\hat{\beta}_0, \hat{\beta}_1)$. Hence, you should solve the following problem:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} f(\hat{\beta}_0, \hat{\beta}_1)$$

or

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{2} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

In this case, we have to take the partial derivative wrt the two arguments, ($\hat{\beta}_0$ and $\hat{\beta}_1$. Let's start with $\hat{\beta}_0$

$$\begin{aligned}
\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} &= 0 \\
&= -2(Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1) - 2(Y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2) \\
&= -(Y_1 + Y_2) - 2\hat{\beta}_0 - \hat{\beta}_1(X_1 + X_2) \\
&= 0
\end{aligned}$$

Therefore,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{1}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^{2} Y_i}{2} \quad \text{and} \quad \bar{X} = \frac{\sum_{i=1}^{2} X_i}{2}$$

[1] Since the second derivative is positive, we know that this is a minimum.

Now we should do the same, but this time wrt $\hat{\beta}_1$

$$\begin{aligned}\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} &= 0 \\ &= -2(Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1)X_1 - 2(Y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2)X_2 \\ &= -(Y_1X_1 + Y_2X_2) + \hat{\beta}_0(X_1 + X_2) + \hat{\beta}_1(X_1^2 + X_2^2) \\ &= 0\end{aligned}$$

Substituting $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$*, we have that*

$$\begin{aligned}0 &= -(Y_1X_1 + Y_2X_2) + (\bar{Y} - \hat{\beta}_1\bar{X})(X_1 + X_2) + \hat{\beta}_1(X_1^2 + X_2^2) \\ &= -\left[X_1(Y_1 - \bar{Y}) + X_2(Y_2 - \bar{Y})\right] - \hat{\beta}_1\bar{X}(X_1 + X_2) + \hat{\beta}_1(X_1^2 + X_2^2) \\ &= -\sum_{i=1}^{2} X_i(Y_i - \bar{Y}) - \hat{\beta}_1\left(\bar{X}X_1 + \bar{X}X_2 - X_1^2 - X_2^2\right) \\ &= -\sum_{i=1}^{2} X_i(Y_i - \bar{Y}) + \hat{\beta}_1\left[X_1\left(X_1 - \bar{X}\right) + X_2\left(X_2 - \bar{X}\right)\right] \\ &= -\sum_{i=1}^{2} X_i(Y_i - \bar{Y}) + \hat{\beta}_1\sum_{i=1}^{2} X_i\left(X_i - \bar{X}\right) \\ &= 0\end{aligned}$$

Which, solving for $\hat{\beta}_1$*, becomes*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{2} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{2} X_i\left(X_i - \bar{X}\right)}$$

(c) What if now we have

$$\min_{\hat{\beta}_0, \hat{\beta}_1} f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the function.

$$\begin{aligned}\frac{\partial f(\hat{\beta}_0, \hat{\beta}_0)}{\partial \hat{\beta}_1} &= 0 \\ &= -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= -\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n}\hat{\beta}_0 + \sum_{i=1}^{n}\hat{\beta}_1 X_i \\ &= -\sum_{i=1}^{n} Y_i + n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} X_i\end{aligned}$$

Which, after solving for $\hat{\beta}_0$*, becomes:*

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{2}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \quad \text{and} \quad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Now, let us find the value of $\hat{\beta}_0$ that minimize the SSR:

$$\begin{aligned}
\frac{\partial f(\hat{\beta}_0, \hat{\beta}_0)}{\partial \hat{\beta}_1} = 0 & \\
&= -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i \\
&= -\sum_{i=1}^{n}\left[Y_i - \left(\bar{Y} - \hat{\beta}_1\bar{X}\right)\hat{\beta}_1 X_i\right]X_i \qquad \text{(using (2))} \\
&= -\sum_{i=1}^{n}\left(Y_i X_i - \bar{Y}X_i\right) - \sum_{i=1}^{n}\hat{\beta}_1\left(\bar{X}X_i - X_i^2\right) \\
&= -\sum_{i=1}^{n} X_i\left(Y_i - \bar{Y}\right) + \hat{\beta}_1\sum_{i=1}^{n} X_i\left(X_i - \bar{X}\right)
\end{aligned}$$

Solving for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i\left(X_i - \bar{X}\right)} \tag{3}$$

In previous problems we showed that the numerator of (3) the can be written as

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

In other words, the numerator is the sample covariance between X and Y. Note also that $X_i\left(X_i - \bar{X}\right) = \left(X_i - \bar{X}\right)^2$. Hence, the denominator can be expressed as

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Therefore, the denominator is the sample variance of X.[2]

Hence, we can express $\hat{\beta}_1$ as,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

[2] This is true because dividing the numerator and denominator by $n-1$ does not change anything.

8. *You will need Excel, Stata, or other statistical programming software to complete this question. You may report your findings in the space provided below.*
A recent article reported that there are approximately 11 minutes of actual playing time in a typical National Football League (NFL) game (*The Wall Street Journal*, January 15, 2010). The article included information about the time devoted to commercials, and the amount of time the players spend standing around between plays. Data consistent with the findings published in *The Wall Street Journal* are in the file named "Standing" (available on Concourse). These data provide the amount of time players spent standing around between plays for a sample of 60 NFL games.

   (a) Use the Standing data set to develop a point estimate of the number of minutes during an NFL game that players are standing around between plays. Compare this to the actual playing time reported in the article. Are you surprised?
   The amount of playing time is approximately 11 minutes. So the time standing around is over 6 times as much. You may find this difference surprising; the authors did.

   (b) What is the sample standard deviation?
   $s = 4.4943$ *minutes*

   (c) Develop a 95% confidence interval for the number of minutes players spend standing around between plays.

$$\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

$$\left(\bar{x} - 2.001\frac{s}{\sqrt{n}}, \bar{x} + 2.001\frac{s}{\sqrt{n}}\right)$$

$$\left(66.93 - 2.001\frac{4.4943}{\sqrt{60}}, 66.93 + 2.001\frac{4.4943}{\sqrt{60}}\right)$$

   Therefore, the 95% confidence interval for the number of minutes players spend standing around between plays, μ, is (65.77, 68.09).

Table 1: Summary

Mean	66.93
Sample Standard Deviation	4.49
Confidence Interval LB	65.77
Confidence Interval UB	68.09

Table 2: Summary Alternative Calculation

Margin of error	1.16
Confidence Interval LB	65.77
Confidence Interval UB	68.09