

Case-based Social Statistics II

CSSS 322

Professor: Mark S. Handcock

Solutions to Homework 2

Due Thursday, April 18, 2002

Problems to be handed in:

- 1) Based on responses of 1467 subjects in the General Social Surveys in the mid-1980s, a 95% confidence interval for the mean number of close friends equals (6.8, 8.0). Which of the following interpretations of this interval is (are) correct?

a) We can be 95% confident that \bar{X} is between 6.8 and 8.0.

Solution(1): Not correct. \bar{X} is always in the interval.

✕

b) We can be 95% confident that μ is between 6.8 and 8.0.

Solution(1): Correct.

✕

c) Ninety-five percent of the values of $X =$ the number of close friends (for this sample) are between 6.8 and 8.0.

Solution(1): This is not true. The interval described in the interpretation is that containing 95% of the sample values.

✕

d) If random samples of size 1467 were repeatedly selected, then 95% of the time \bar{X} would be between 6.8 and 8.0.

Solution(1): This is not correct as \bar{X} will vary, but need not be in this precise interval 95% of the time. Remember that the interval itself is random, not just \bar{X} .

✕

e) If random samples of size 1467 were repeatedly selected, then in the long run 95% of the confidence intervals formed would contain the true value of μ .

Solution(1): This is correct.

✕

- 2) A survey is taken to estimate the mean annual family income for families living in public housing in Chicago. For a random sample of 30 families, the annual incomes (in hundreds of dollars) are as follows:

83	90	77	100	83	64	78	92	73	122
96	60	85	86	108	70	139	56	94	84
111	93	120	70	92	100	124	59	112	79

The data is entered into Datatools on the course website.

- a) Using Datatools, construct a stem-and-leaf plot of the incomes. What do you predict about the shape of the population distribution?

Solution(3):

Stem and leaf plot:

N = 30 Median = 88
 Quartiles = 77, 100

Decimal point is 1 place to the right of the colon

```

5 : 69
6 : 04
7 : 003789
8 : 33456
9 : 022346
10 : 008
11 : 12
12 : 024
13 : 9
  
```

The shape of the population distribution is probably mound-shaped.

- b) Construct and interpret estimates of μ and σ , the mean and standard deviation of the family incomes of all families living in public housing in Chicago.

Solution(3): The point estimate of μ is the sample mean \bar{X} . $\bar{X} = \frac{\sum_{i=1}^{30} x_i}{30} = 90$; x_i represents the data. The point estimate of σ is the sample standard deviation:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{30} (x_i - \bar{X})^2}{(30 - 1)}} = \sqrt{427.24} = 20.7.$$

⊗

- c) Construct and interpret a 95% confidence interval for μ .

Solution(3): To construct 95% confidence interval of μ we first have to find $\hat{\sigma}_{\bar{X}}$. $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{20.7}{\sqrt{30}} = 3.77$. The 95% confidence interval of μ is $90 \pm 1.96 (3.77) = 90 \pm 7.4 = (82.6, 97.4)$ We predict that μ falls between \$8260 and \$9740.

⊗

- d) Construct a 95% confidence interval for μ . Interpret the interval and compare it to the one in part (c).

Solution(3): The 99% confidence interval of μ is $90 \pm 2.58 (3.77) = 90 \pm 9.7 = (80.3, 99.7)$. We can be 99% confident that the mean family income of all families living in public housing in Chicago falls between \$8030 and \$9970. This interval is wider than the one in part c), since the confidence coefficient is larger.

⊗

3) A study is conducted of a geographic distribution of the residences of the employees at a large factory, in order to determine the suitability of initiating busing to that factory. One variable considered is the distance the employee lives from the factory. For a random sample of 100 employees, the mean distance is 6.3 miles and the standard deviation is 4.0 miles.

a) Find and interpret a 90% confidence interval for the mean residential distance from the factory for all employees.

Solution(3): To get the 90% confidence interval for the mean residential distance, we first have to get $\hat{\sigma}_{\bar{X}}$:

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{4}{\sqrt{100}} = 0.4.$$

The 90% confidence interval for the mean residential distance is

$$6.3 \pm 1.64(0.4) = 6.3 \pm 0.7 = (5.6, 7)$$

. We can be 90% confident that the mean residential distance from the factory of all employees is between 5.6 and 7 miles. ⊗

b) Find and interpret a 90% *prediction* interval for the residential distance from the factory for a randomly chosen employee.

Solution(3): To get the 90% prediction interval for a single employee's residential distance, we first have to get $\hat{\sigma}_X$:

$$\hat{\sigma}_X = s\sqrt{1 + \frac{1}{n}} = 4\sqrt{1 + \frac{1}{100}} = 4.02.$$

The 90% prediction interval for the residential distance is

$$6.3 \pm 1.64(4.02) = 6.3 \pm 0.7 = (-0.3, 12.9)$$

. We can be 90% confident that the residential distance from the factory of a single employees is between 0 and 12.9 miles (clearly it can not be negative). ⊗

c) About how large a sample would have been adequate if we merely needed to estimate the mean to within 1.0 miles, with 95% confidence?

Solution(3): Sample size required for estimating a mean μ to within 1, with 90% confidence is:

$$n = \sigma^2 \left(\frac{z}{B}\right)^2 = 4^2 \left(\frac{1.64}{1}\right)^2 = 43.$$

⊗

4) In the 1991 General Social Survey, respondents were asked whether people convicted of murder should receive the death penalty. 1078 responded yes and 336 responded no. Construct a 99% confidence interval for the proportion of American adults who would answer yes. Interpret this interval. Can you conclude that more than half of all American adults would answer yes? Explain briefly.

Solution(7): The proportion of respondents saying yes in the survey: $\hat{\pi} = \frac{1078}{(1078+336)} = \frac{1078}{1414} = 0.76$.
The 99% confidence interval is:

$$0.76 \pm 2.58 \sqrt{\frac{(0.76)(0.24)}{1414}} = 0.76 \pm 2.58(0.0113) = 0.76 \pm 0.03 = (0.73, 0.79).$$

We conclude that more than half of all American adults would answer yes, since all numbers in the confidence interval exceed 0.5. ∞

Extra Credit Problem:

5) The General Social Survey is a means of collecting information about the U.S. population. It is an annual survey that is based on randomly sampling people and interviewing them about the economic and social aspects of their lives. Suppose that you took a random sample of 100 families to find out about the weekly income of families in the U.S. Suppose you find that the mean income of families in the sample was \$874 and the sample standard deviation of the incomes was \$186.

a) Find the 95% confidence interval for the population mean.

Solution(3): The confidence interval formula is:

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

Here $\alpha = 0.05$, $s = \$186$ and $n = 100$, so the interval is:

$$\begin{aligned} & \left(874 - 1.96 \times \frac{186}{\sqrt{100}}, 874 + 1.96 \times \frac{186}{\sqrt{100}} \right) \\ & = (\$837.54, \$910.46) \end{aligned}$$

Note that as $n > 40$ we approximate the t -multiplier by that with infinite degrees of freedom, 1.96. ∞

b) Suppose a local urban planning organization is interested in the total income for all families in Western Washington. If there are 543,000 families in Western Washington, find a 95% confidence interval for the *total* weekly income for all these people. Briefly explain to the meaning of your answer.

Solution(6): From part a) we have a 95% confidence interval for the average family income. The total amount is 543,000 times this amount, so we can proportionally expand out the confidence interval for the average. That is, a 95% confidence for the total income is:

$$\begin{aligned} & 543,000 \times (\$837.54, \$910.46) \\ & = (543,000 \times \$837.54, 543,000 \times \$910.46) \\ & = (\$454,786,392, \$494,377,608) \end{aligned}$$

Your explanation of this process should be: If we sampled 100 families in Western Washington from the total of 543,000 families over and over again, each time calculating the confidence interval for the total balance due based on sample mean, then in the long-run the proportion of these intervals covering the total income will be 95%. ∞

- c) Suppose that the urban planning organization undertook a complete accounting of the entire population of families, and found a total of \$450,714,000. Your job is on the line. What do you say to him?

Solution(3): The true total income from all families, \$450,714,000, falls just outside our confidence interval, which is grating. Our analysis indicates that this will happen only $100\% - 95\% = 5\%$ of the time. Had we sampled more and more accounts, our confidence interval would have become more and more accurate. However, based on the 100 samples we had, the most accurate we could be is the interval estimate (\$454,786,392, \$494,377,608). ⊗