STOCKHOLM UNIVERSITY
Department of Statistics

Emma Pettersson, Michael Carlson
2019-03-28

Solutions to Plenary Exercises: Plenary Exercise 4

Basic Statistics for Economists, 15 ECTS, STE101

EXERCISE 1

Assume that the observations are independent. Note that the sample is small, $n < 30$.

The sample mean: and standard error of the sample mean are respectively

$$\bar{x} = \frac{5.01}{5} = 1.002 \qquad SE(\bar{x}) = \frac{\sigma_X}{\sqrt{n}}$$

The confidence interval can be calculated with the following formula:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$$

	95% CI:	99% CI:	99.9% CI:
α	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
$z_{\alpha/2}$ [from Table 2]	$z_{0.025} = 1.96$	$z_{0.005} = 2.5758$	$z_{0.0005} = 3.2905$
$z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$	$1.96 \cdot \frac{0.05}{\sqrt{5}} = 0.044$	$2.5758 \cdot \frac{0.05}{\sqrt{5}} = 0.058$	$3.2905 \cdot \frac{0.05}{\sqrt{5}} = 0.074$
$\bar{x} \pm z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}$	1.002 ± 0.044 $= (\mathbf{0.958, 1.046})$	1.002 ± 0.058 $= (\mathbf{0.944, 1.060})$	1.002 ± 0.074 $= (\mathbf{0.928, 1.076})$

The higher the confidence level, the wider the confidence interval. The more ‘confident’ you want to be that the true population parameter lies within the interval, the wider the interval will have to be. The only way to increase confidence without getting a wider interval is to increase the sample size. Or, the only way to reduce your uncertainty about the parameter (i.e. reduce the margin of error and get a smaller interval) without changing the confidence level is to increase the sample size.

EXERCISE 2

a. Assume that the total number of invoices is very large, almost as if $N \rightarrow \infty$ at least compared to the sample size $n = 100$. Also, assume that all the invoices are independently and identically distributed (iid).

Let Y be the number of invoices that are incorrectly reported from the sample of $n = 100$ invoices. The estimated probability of reporting an invoice incorrectly is $\hat{p} = Y/n$. Because the sample is large ($n > 30$), according to the CLT, $\hat{p}$ is approximately normally distributed.

$$\hat{p} = \frac{Y}{n} \sim N\left(P, \frac{P(1-P)}{n}\right)$$

Point estimate:

$$\hat{p} = \frac{11}{100} = 0.11$$

95% confidence interval for P:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.11 \pm 1.96\sqrt{\frac{0.11(0.89)}{100}} = 0.11 \pm 0.061 \Longrightarrow (\mathbf{0.049}, \mathbf{0.171})$$

b. Let $X_k$be the accounting error on invoice k and $\sum X_i$ be the total accounting error.

The whole sample needs to be included in the calculations, even the 89 invoices that have no accounting errors (see note below).

Because the sample is large, according to the CLT, $\bar{X}$ will be approximately normally distributed.

$$\bar{X} = \frac{\sum X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Point estimate:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{22\ 000}{100} = 220 \text{ kr}$$

95% confidence interval:

$$\bar{x} \pm z_{\alpha/2}\frac{s_x}{\sqrt{n}} \Longrightarrow 220 \pm 1.96\frac{700}{10} \Longrightarrow 220 \pm 137.2 \Longrightarrow (\mathbf{82.80}, \mathbf{357.20})$$

NOTE: It is a bit unclear in the description of the problem if the standard deviation 700 refers to the entire sample (including the non-erroneous invoice for which $X_i = 0$) or if it refers to only the eleven erroneous invoices. Above we have assumed that it refers to the entire sample. The average accounting error for the $n_E = 11$ erroneous invoices is obviously

$$\bar{x}_E = \frac{\sum x_i}{n_E} = \frac{22\ 000}{11} = 2000 \text{ kr}$$

Assuming that 700 refers to the entire sample the standard deviation can also be calculated from the given information. However, this is a bit more complicated but it turns out to be $s_E = 966.954 \approx 967$. If on the other hand 700 is the standard deviation only for the eleven erroneous invoice than the standard deviation for the entire sample is 667.121.

EXERCISE 3

A 95% confidence interval can be calculated with the following formula:

$$\bar{x} \pm z_{\alpha/2}\frac{s_x}{\sqrt{n}}$$

a. Industrial companies:

$$65.04 \pm 1.96\frac{35.72}{\sqrt{250}} \Longrightarrow 65.04 \pm 4.43 \Longrightarrow (\mathbf{60.61, 69.47})$$

b. Financial companies:

$$56.74 \pm 1.96\frac{34.87}{\sqrt{238}} \Longrightarrow 56.74 \pm 4.43 \Longrightarrow (\mathbf{52.31, 61.17})$$

c. When comparing the intervals, we do see that they overlap, which may imply that the time delays for the two sectors are the same.

However, a better alternative is to calculate a confidence interval for the **difference** between the average time delays of the two sectors $\mu_I - \mu_F$.

If $\bar{X}_I \sim N\left(\mu_I, \frac{\sigma_I^2}{n_I}\right)$ and $\bar{X}_F \sim N\left(\mu_F, \frac{\sigma_F^2}{n_F}\right)$ then for the linear combination of the two independent variables, the following distribution holds:

$$\bar{X}_I - \bar{X}_F \sim N\left(\mu_I - \mu_F, \frac{\sigma_I^2}{n_I} + \frac{\sigma_F^2}{n_F}\right)$$

We calculate a 95% confidence interval for the difference (since the samples sizes are fairly large we are satisfied in using the sample variances; we don't know the true variances!):

$$\bar{x}_I - \bar{x}_F \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_I^2}{\sqrt{n_I}} + \frac{s_F^2}{\sqrt{n_F}}} = 65.04 - 56.74 \pm 1.96\sqrt{\frac{35.72^2}{250} + \frac{34.87^2}{238}} = 8.3 \pm 6.26$$

$$\Longrightarrow (\mathbf{2.04, 14.56})$$

The interval does not contain the value 0, this indicates that the time delay difference between the two sectors is (absolute) greater than 0.

This alternative is better, because when we just compare two independent intervals we get a total 'uncertainty' of 4.43 + 4.43 = 8.86, which is greater than 6.26, the uncertainty we get from the second method.

$$\sqrt{\frac{s_I^2}{\sqrt{n_I}} + \frac{s_F^2}{\sqrt{n_F}}} \leq \sqrt{\frac{s_I^2}{\sqrt{n_I}}} + \sqrt{\frac{s_F^2}{\sqrt{n_F}}}$$

EXERCISE 4

Assume that the two samples are independent, and that within both samples, observations are *iid*.

X_k is a Bernoulli-distributed variable that has the value 0 with the probability $(1-P)$ and value 1 with probability P.

Let $Y = \sum X_k$be the number of individuals in a sample that answered “yes”, so that $Y \sim Bin(n, P)$.

Recall that we estimate P with $\hat{p} = \frac{Y}{n} = \frac{\sum X_k}{n}$. And $\hat{p}$ has the following expected value and variance: $E(\hat{p}) = E(\bar{X}) = \mu_X = P$ and $ar(\hat{p}) = Var(\bar{X}) = \frac{\sigma_X^2}{n} = \frac{P(1-P)}{n}$.

a. When the samples are large, we can estimate a 95% confidence interval for the population proportion using $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

	February	August
Point estimate $\hat{p}$	$\hat{p} = 0.42$	$\hat{p} = 0.30$
Critical value $z_{\alpha/2}$	$z_{\alpha/2} = z_{0.025} = 1.96$	$z_{\alpha/2} = z_{0.025} = 1.96$
Standard error $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\sqrt{\frac{(0.42)(0.58)}{320}} = 0.02759$	$\sqrt{\frac{(0.3)(0.7)}{200}} = 0.03240$
Confidence interval for P $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$0.42 \pm 1.96(0.02759)$ $= 0.42 \pm 0.054$ $\Rightarrow \mathbf{(0.366, 0.474)}$	$0.3 \pm 1.96(0.03240)$ $= 0.3 \pm 0.064$ $\Rightarrow \mathbf{(0.236, 0.364)}$

b. How large of a sample is needed for the August group interval to be of the same length as the February group interval? We want the margins of error $\left(z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ to be the same for the two groups. Since $z_{\alpha/2}$ is the same for both intervals we can omit this from the calculations and skip the square root:

$$\frac{\hat{p}_F(1-\hat{p}_F)}{n_F} = \frac{\hat{p}_A(1-\hat{p}_A)}{n_A}$$

$$\frac{0.42 \cdot 0.58}{320} = \frac{0.3 \cdot 0.7}{n_A}$$

$$n_A = \frac{0.3 \cdot 0.7 \cdot 320}{0.42 \cdot 0.58}$$

$$n_A = 275.8621 \approx \mathbf{276}$$