Inf2 – Foundations of Data Science 2023

Workshop solution: Semester 2 Week 4 Workshop

5th February 2024

1. **Distribution of the sample mean**

   (a) We haven't specified the distribution of tips. However, given that the sample size is $n = 100$, regardless of the distribution, we expect the distribution of the sample mean will be approximately normal, due to the Central Limit Theorem. Note that as the sample size increases, the sample mean distribution converges to normal.

   (b) The sample mean distribution is centred around the mean of the distribution itself, hence the correct answer is 9%.

   (c) The standard error of the sample mean is the standard error of the population divided by the square root of the sample size, hence 0.6%.

   (d) Here we rely on the sample mean distribution being approximately normal, and use the z-distribution to infer the requested probability. The sample mean, population mean and SEM are, respectively:

   $$\overline{x} = 8 \quad \mu = 9 \quad \sigma_{\overline{X}} = 6/\sqrt{100} = 0.6$$

   From this we compute the z-statistic:

   $$z = \frac{\overline{x} - \mu}{\sigma_{\overline{X}}} \tag{1}$$

   $$= (8–9)/0.6 = -1.667 \tag{2}$$

   We would like to compute the area under the standard normal distribution to right of this value:

   $$1 - \Phi(z) = 1 - \Phi(-1.667) = 1 - 0.0478 = 0.952 \tag{3}$$

2. **Confidence interval calculation 1** As $n = 110$ is over 40, we can assume that the sampling distribution of the statistic

   $$z = \frac{\overline{x} - \mu}{\hat{\sigma}_{\overline{X}}} \tag{4}$$

   is normal with mean 0 and variance 1 (the "z-distribution"); here $\sigma_{\overline{X}}$ is the standard error in the mean. A 99% confidence interval implies the area in the tails of the distribution is $\alpha = 0.01$. As we have been asked for a two-tailed confidence interval, we need to look up the z-critical value $z_{\alpha/2} = z_{0.005} = 2.58$. We have sample mean

and standard deviation $\overline{x} = 0.81$ and $s = 0.34$. Therefore, the standard error in the mean is $\sigma_{\overline{X}} = 0.34/\sqrt{110} = 0.0324$. We substitute the z critical value $z_{\alpha/2} = z_{0.005}$ and rearrange Equation (4) to obtain the upper and lower bounds of the confidence interval for μ:

$$(\overline{x} - \sigma_{\overline{X}} z_{\alpha/2} \ , \ \overline{x} + \sigma_{\overline{X}} z_{\alpha/2}) \tag{5}$$

$$= (0.81 - 0.34/\sqrt{110} \times 2.58 \ , \ 0.81 + 0.34/\sqrt{110} \times 2.58) \tag{6}$$

$$= (0.73, 0.89) \tag{7}$$

3. **Confidence interval calculation 2**

   (a) As $n = 20$, we cannot assume that the sampling distribution of the statistic

   $$t = \frac{\overline{x} - \mu}{\hat{\sigma}_{\overline{X}}} \tag{8}$$

   is normally distributed. The reason is that small number of samples causes considerable variability in the estimate of the standard error $\hat{\sigma}_{\overline{X}}$ between samples, making the distribution wider and flatter than a normal distribution. However, we can expect the statistic t to be distributed according to a t distribution with $n - 1$ degrees of freedom.

   A 95% confidence interval implies the area in the tails of the t-distribution is $\alpha = 0.05$. Because we have been asked for two-sided confidence interval, we need to look up the t-critical value with $n - 1 = 19$ degrees of freedom, $t_{\alpha/2,n-1} = t_{0.025,19} = 2.093$. We have the sample mean and standard deviation $\overline{x} = 25.05$ and $s = 2.690$. Therefore, the standard error in the mean is $\sigma_{\overline{X}} = 2.690/\sqrt{20} = 0.601$. Setting t in Equation (8) to $t_{\alpha/2,n-1} = t_{0.025,19}$ and rearranging, we obtain the confidence interval:

   $$(\overline{x} - \sigma_{\overline{X}} t_{\alpha/2,n-1} \ , \ \overline{x} + \sigma_{\overline{X}} t_{\alpha/2,n-1}) \tag{9}$$

   $$= (25.05 - 2.093 \times 0.601, 25.05 + 2.093 \times 0.601) \tag{10}$$

   $$= (23.79, 26.31) \tag{11}$$

   (b) The answer is (probably) yes. Although we don't know the CI over the ACT mean for the entire university population, we can probably assume the CI for the entire university population is pretty tight because it includes many more students. Even if there are only 2000 students in the university (i.e. 20×100), we would expect the CI to be $\sqrt{100} = 10$ times as tight, and then there would be no overlap in the CIs of the ACT mean between the calculus population and the university population. Hence, we can deduce that the ACT mean for calculus students is most likely higher than the ACT mean for the university population.