Inf2 – Foundations of Data Science 2024

Workshop solution: Semester 2 Week 3 Workshop

22nd January 2025

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.

4. **Confidence intervals concepts**

   (a) The 90% confidence interval will be narrower than the 95% confidence interval, because a narrower interval means that the true mean will appear less frequently in the confidence interval, if the procedure for obtaining confidence intervals was repeated again and again. Note: you may find looking at this demo helps to deepen your understanding: https://rpsychologist.com/d3/ci.

(b) The statement is incorrect. According to the definition of a confidence interval there is a 95% chance that the confidence interval (a random variable) contains μ. In a frequentist interpretation, μ is not a random variable – we could conceivably compute it if we could measure the alcohol content in *all* the bottles. It therefore doesn't make sense to say that there is a "chance" that μ is within the confidence interval.

(c) The statement is incorrect. The confidence interval does not relate to the alcohol content in each bottle, but about the mean alcohol content in a sample of 50 bottles. We would expect the distribution of alcohol content to be much broader, as we saw e.g. in Q1. If we knew the pdf of the alcohol content per bottle, we could compute quantiles for 0.025 and 0.975 to get something like the question is asking for – but we don't know the original pdf of the alcohol content.

(d) We can't be sure that exactly 95 of the intervals will contain μ. However, we would expect 95 of the intervals to contain μ – or put another way, there is a 0.95 chance of each confidence interval containing μ. See Figure 1 in the lecture notes on confidence intervals, which provides the frequentist interpretation to the probability statement.

(e) 9 times as many samples. The width of the 95% confidence interval is proportional to the SEM. We therefore need to reduce the SEM by a factor of 3. Therefore, we need to increase n by a factor of 9, since SEM $\propto 1/\sqrt{n}$.

5. **Distribution of the sample mean**

(a) No change to (1a). For 1b, the distribution of the sample mean may appear non-normal, depending on what the distribution of the tips is. For 1c, the standard error of the sample mean is $6/\sqrt{10} = 1.897\%$. As we can see, the standard error increases as the sample size decreases. For 1d,

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{X}}} \tag{12}$$

$$= (8–9)/1.897 = -0.527 \tag{13}$$

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

$$1 - \Phi(z) = 1 - \Phi(-0.527) = 1 - 0.300 = 0.700 \tag{14}$$

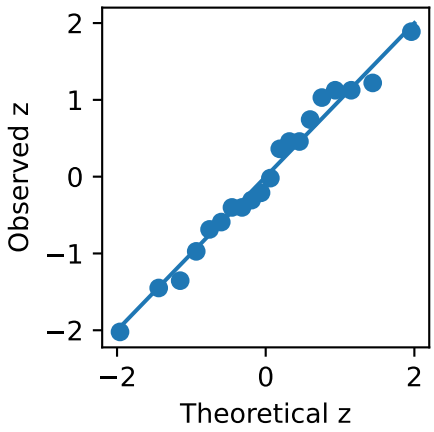
This is lower than when we had $n = 100$ samples, corresponding to the greater spread of the standard error in the mean.

(b) As the numbers in each sample are small, and the distribution of the tips is not normal, we might expect the sampling distribution of the mean tip not to be normal. Statistical simulations demonstrate that this the case – the sampling distribution actually has a bumpy appearance, because in each sample of 10 tips, 0, 1, 2, 3... of them may be zero. A statistical simulation also shows that probability of getting more than 8% is 0.69, so similar to the theoretical value.

6. **Thinking critically about data**

- If the sample is taken later in the semester and some students have dropped out, we might suppose that the students with lower ACT scores were more likely to drop out, in which case the mean ACT scores of students on the course may have gone up. We might therefore be wary about drawing conclusions about the students who enorlled on the course initially.
- We can't draw any conclusions about future performance. We might hypothesise that students with higher ACT scores would do better in the future, but without getting data on the future performance of the students, we cannot test this hypothesis.
- If we wanted to find out why students on the calculus course have higher ACT grades, we might want to examine the University and course admissions procedure: for example, perhaps higher school grades were required to get onto maths programmes? We should also examine how ACT scores are generated: for example, is it easier to get high scores in the maths area of the ACT than the English and Social Science? Students who find maths easier than English may then have higher ACT scores than those who find English easier than Maths.

7. **Checking for normality**
   To draw the Q-Q plot we plot the normalised data against the theoretical centiles. I.e. for $n = 20$, the theoretical locations of data points if there was a perfect normal distribution would be the at the following centiles: $2.5\%, 7.5\% \ldots 97.5\%$.



   The points cluster around the 45° line, indicating that the data is approximately normally distributed.
   Note that there are statistical tests of normality; see https://en.wikipedia.org/wiki/Normality_test. However, this is beyond the scope of the course.