DS 122 Homework 3

Xiang Fu
xfu@bu.edu
Boston University Faculty of Computing & Data Sciences

Contents

1 Question 1

A recent survey by the ABC Health Institute suggests that tech professionals, given the high demands of their job get, on average, 6 hours of sleep per night with a standard deviation of 1.5 hours. Assume that the distribution of sleep hours is approximately normal.

1.1 Part A

Find the probability that a randomly selected tech professional will get more than 7 hours of sleep.

1.2 Answer

Given:

- μ = 6 hours (mean hours of sleep)
- σ = 1.5 hours (standard deviation)
- Distribution of sleep hours is approximately normal

We want to find: $P(X \geq 7)$, where X represents the random variable denoting the hours of sleep.

1.2.1 Standardize the variable

The z-score for a value X from a normal distribution with mean μ and standard deviation σ is given by:

$$z = \frac{x - \mu}{\sigma}$$

For x = 7 hours:

$$z = \frac{7 - 6}{1.5} \approx 0.67$$

1.2.2 Finding the probability using the z-score

Given:

$$P(Z > 0.67) = 1 - P(Z < 0.67) = 0.2525$$

1.2.3 Conclusion

Thus, the probability that a randomly selected tech professional will get more than 7 hours of sleep is approximately 25.25%.

1.3 Part B

Find the probability that the mean sleep duration of a random sample of 49 tech professionals will exceed 7 hours (Hint: you can use the standard deviation from the original survey to calculate the standard error, you do not need to calculate a sample standard deviation).

1.4 Answer

Given:

- μ = 6 hours (mean hours of sleep for the population)
- σ = 1.5 hours (standard deviation for the population)
- n = 49 (sample size)

We want to find:

$$P\left(\overline{X} > 7\right)$$

where $\overline{X}$ represents the sample mean.

1.4.1 Calculate the standard error (SE)

$$\sigma_{\text{sample mean}} = \frac{\sigma}{\sqrt{n}}$$

Plug in the values:

$$\sigma_{\text{sample mean}} = \frac{1.5}{\sqrt{49}} \approx 0.2143$$

1.4.2 Standardize the sample mean

Using ths formula:

$$z = \frac{\overline{x} - \mu}{\sigma_{\text{sample mean}}}$$

For $\overline{x} = 7$ hours:

$$z = \frac{7 - 6}{0.2143} \approx 4.67$$

1.4.3 Finding the probability using the z-score

$$P(Z > 4.67) = 1 - P(Z < 4.67) \approx 1.53 * 10^{-6}$$

Thus, the probability that the mean sleep duration of a random sample of 49 tech professionals will exceed 7 hours is approximately 0.000153%.

1.5 Part C

You are worried that some tech professionals may be sleep deprived, so you want to determine how little sleep the worst rested individuals are getting. Determine the number of hours of sleep that the 10th percentile of tech professionals are getting (hint: use the z-score formula as usual, but invert it to solve for the x that corresponds to the z-score for the 10th percentile)

1.6 Answer

Given the z-score formula:

$$z = \frac{x - \mu}{\sigma}$$

We can rearrange it to solve for x:

$$x = \mu + z * \sigma$$

1.6.1 Find the z-score for the 10th percentile

The z-score corresponding to the 10th percentile, $P(Z < z) = 0.10$, can be found using the inverse of the cumulative distribution function (the quantile function)

For the 10th percentile, $z_{0,10}$ is around -1.28.

Using this value:

$$x \approx 6 + (-1.28 * 1.5) = 4.08$$

This means that 10% of tech professionals are getting 4.08 hours of sleep or less.

2 Question 2

A renowned car manufacturing company in Detroit is considering setting up new showrooms across Europe. They have gathered data on the expenses related to opening showrooms in 15 European cities to gauge the average expense. Based on their sample, the average expense comes out to be 15,000 dollars, and the standard deviation is 3,500 dollars.

2.1 Part A

Assuming the expenses follow an approximately normal distribution, find the 95% confidence interval for the true mean expense of opening a showroom in Europe (hint: 15 is a small sample size)

2.2 Answer

Use the t-distribution since the sample size is small (less than 30).

Given:

- Sample mean, $\overline{x} = 15000$
- Sample standard deviation, $s = 3500$
- Sample size, $n = 15$

The formula for the confidence interval using the t-distribution is:

$$\overline{x} \pm t_{\frac{a}{2},\,\mathrm{df}} * \left(\frac{s}{\sqrt{n}}\right)$$

For a 95% confidence interval and 14 degrees of freedom, we will find $t_{0.025,14}$.

Using the t-table, the t-value for a 95% confidence interval with 14 df is approximately 2.145.

Next, we need to calculate the margin of error, which is given by:

$$\text{MOE} = t_{\frac{a}{2},\ \text{df}} * \left(\frac{s}{\sqrt{n}}\right)$$

Given:

$$s = 3500, n = 15$$

$$\text{MOE} = 2.145 * \left(\frac{3500}{\sqrt{15}}\right) \approx 1938.24$$

Next, using the formula:

$$\text{CI} = \overline{x} \pm \text{MOE}$$

where $\overline{x} = 15000$

The 95% confidence interval is:

$$\text{CI} = (15000 - 1938.24, 15000 + 1938.24)$$

$$\text{CI} = (13061.76, 16938.24)$$

This means that we are 95% confidence that the true average expense of opening a showroom in Europe lies between \$13061.76 and \$16938.24.

2.3 Part B

How does the size of the sample affect the width of the confidence interval?

2.4 Answer

1. As the sample size (n) increases, the width of the confidence interva decreases, and vice versa. This relationship is due to the standard error (SE) component in the margin of error (MOE) formula:

$$\text{SE} = \frac{s}{\sqrt{n}}$$

Where s is the sample standard deviation. As n increases, $\sqrt{n}$ also increases, leading to a smaller SE. A smaller SE results in a small MOE, which in turn narrows the confidence interval.

2. A larger sample size provdies more information about the population, which leads to more precise estimates of the population parameter. In other words, with a larger sample, we can be more confident that our sample mean is close to the true population mean, leading to a narrower confidence interval.

3. Also, as the sample size increases, the sampling distribution of the sample mean becomes more normally distributed. This makes the use of the z-distributon more appropriate for larger samples ($n \geq 30$). For samller samples, especially when the population distribution is not known or not normal, the t-distribution is used, which has wider tails and therefore results in a wider confidence interval compared to the z-dsitribution.

2.5 Part C

If the company wants to be more certain that their interval contains the true mean, and so instead aims for a 99% confidence interval, what would be the interval?

2.6 Answer

Given:

- Sample mean, $\overline{x} = 15000$
- Sample standard deviation, $s = 3500$
- Sample size, $n = 15$
- Degrees of freedom, df = 14

For a 99% confidence interval, $\alpha = 1 - 0.99 = 0.01$. Thus, we want to find $t_{0.005,14}$.

Using the t-table, the t-value for a 95% confidence interval with 14 df is approximately 2.977.

$$\text{MOE} = 2.977 * \left(\frac{3500}{\sqrt{15}}\right) \approx 2690.30333$$

$$\text{CI} = (15000 - 2690.30333, 15000 + 2690.30333)$$

$$\text{CI} = (12309.69667, 17690.30333)$$

This means that we are 99% confidence that the true average expense of opening a showroom in Europe lies between \$12309.69667 and \$17690.30333.

3 Question 3

An online streaming platform claims that, on average, gamers on their platform stream games for 300 hours a year. A group of enthusiastic gamers feels that the actual average streaming time is higher than the platform's claim. To investigate this, a random sample of 40 gamers was taken, revealing an average streaming time of 320 hours per year with a sample standard deviation of 60 hours. Please note that we know nothing about the distribution of the population.

You can use the following t-value calculator : https://goodcalculators.com/student-t-value-calculator/

3.1 Part A

Perform a one-tailed hypothesis test to test the streaming platform's claim. Use a level of significance of .05 and your random sample to make a conclusion about the null hypothesis.

3.2 Answer

$$H_0 : \mu = 300 \text{ (The average streaming time on the platform is 300 hours)}$$

$$H_1 : \mu > 300 \text{ (The average streaming time on the platform is more than 300 hours)}$$

Since the population standard deviation is not known and the sample size is relatively large ($n = 40$), we can use the t-test statistic.

Using t-table, a level of signifiance of 0.05 for a one-tailed test with df = 40 - 1 = 39, we found out that the critical t-value is about 1.685.

Using the formula:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where:

- $\overline{x} = 320$
- $\mu = 300$
- $s = 50$
- $n = 40$

Plugging in the values:

$$t = \frac{320 - 300}{\frac{60}{\sqrt{40}}} \approx 2.58$$

Given tat 2.58 > 1.685, we would reject the null hypothesis H_0.

Thus, there is sufficient evidence, at the 0.05 level of signifiance, to conclude that the average streaming time on the platform is more than 300 hours.

3.3 Part B

If your test conclusion is in error, what type of error is it? Type I or II (i.e. false positive or false negative)

3.4 Answer

That would be a Type I Error (false positive).

In our case, the null hypothesis was that the average streaming time is 300 hours. Since we've rejected this null hypothesis in favor of the alternative hypothesis (that the average streaming time is greater than

300 hours), if we arw wrong in our conclusion, it means we have seen evidence in our sample that does bot relfect the actual situation in the whole population.

3.5 Part C

Assume you did not reject the null hypothesis, what are you concluding about the difference between the actual streaming hours and the platform's claim?

3.6 Answer

In this situation, our conclusion would be:

We do not have sufficient evidence to suggest that that actual average streaming time is different from the platform's claim of 300 hours a year.

In other words, based on the sample data, we cannot say that the average streaming time is greater than 300 hours. This does not necessarily prove that the platform's claim is correct, but rather that we don't have strong enough evidence from our sample to dispute it.

4 Question 4 Computational

See the Jupyter Notebook for more info.