Advanced Statistical Analysis
Detailed Solutions - Exercise Series 3: Estimations

Exercise 1

Every week, a supermarket selects a sample of 100 customers to estimate the average spending amount. Based on numerous previous surveys, the supermarket assumes that the spending of each customer approximately follows a normal distribution with a standard deviation of 10 euros. This week, the sample mean found by the supermarket is 45 euros.

1. Estimate the average amount spent by the supermarket's customers using a confidence interval at the 95% confidence level.

   Solution:

   To construct a confidence interval, we use the following formula:

   $[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$

   where:

   - $\bar{x}$ is the sample mean (45 euros)
   - σ is the population standard deviation (10 euros)
   - n is the sample size (100)
   - $z_{\alpha/2}$ is the critical value for a standard normal distribution at $1 - \alpha/2$ level

   For a 95% confidence level, $\alpha = 0.05$ and $z_{\alpha/2} = z_{0.025} = 1.96$.

   Now, let's calculate the interval:

   $[45 - 1.96\frac{10}{\sqrt{100}}, 45 + 1.96\frac{10}{\sqrt{100}}]$

   $= [45 - 1.96, 45 + 1.96]$

   $= [43.04, 46.96]$

   Interpretation: We can be 95% confident that the true average amount spent by customers lies between 43.04 and 46.96 euros.

2. Same estimation for a confidence level of 99%.

   Solution:

   The method is the same, but we use a different $z_{\alpha/2}$ value.

   For a 99% confidence level, $\alpha = 0.01$ and $z_{\alpha/2} = z_{0.005} = 2.576$.

   $[45 - 2.576\frac{10}{\sqrt{100}}, 45 + 2.576\frac{10}{\sqrt{100}}]$

   $= [45 - 2.576, 45 + 2.576]$

   $= [42.424, 47.576]$

Interpretation: We can be 99% confident that the true average amount spent by customers lies between 42.42 and 47.58 euros.

Note: Notice that the 99% confidence interval is wider than the 95% one. This is logical because, to be more confident (99% instead of 95%), we must consider a broader range of possible values for the true mean.

On a farm, past studies have shown that the weight of a randomly chosen egg can be considered as the realization of a normal random variable X with a variance of 2. A random sample of 36 eggs is taken, and the measurements (in grams) are as follows:

50.34	51.41	51.51	52.07	52.22	52.38
52.62	53.13	53.28	53.3	53.32	53.39
53.79	53.89	54.63	54.78	54.93	54.99
55.04	55.12	55.24	55.28	55.56	55.82
55.91	55.95	57.05	57.18	57.31	57.67
57.99	58.1	59.3	60.58	63.15	-

Table 1: Weights of sampled eggs (in grams)

1. **Calculate the point estimator of the mean and variance of egg weights.**

   Solution:

   The point estimator of the mean is the sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{50.34 + 51.41 + \cdots + 63.15}{36} \approx 54.85 \text{ grams}$$

   The point estimator of the variance is the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx 7.45 \text{ grams}^2$$

   The sample standard deviation is:

$$s = \sqrt{7.45} \approx 2.73 \text{ grams}$$

   Interpretation: - The estimated average weight of the eggs is approximately 54.85 grams. - The estimated variability of the egg weights is 7.45 grams^2 (variance) or 2.73 grams (standard deviation).

2. **Estimate the average weight of an egg using a confidence interval at the 95% and 98% confidence levels.**

   Solution:

   We know the population variance $\sigma^2 = 2$, so we will use the normal distribution to construct the confidence interval.

   The general formula is:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

For a 95% confidence level:

$$z_{0.025} = 1.96$$

$$\left[54.85 - 1.96 \times \frac{\sqrt{2}}{\sqrt{36}}, 54.85 + 1.96 \times \frac{\sqrt{2}}{\sqrt{36}}\right]$$

$$= [54.85 - 0.47, 54.85 + 0.47]$$

$$= [54.38, 55.32]$$

For a 98% confidence level:

$$z_{0.01} = 2.326$$

$$\left[54.85 - 2.326 \times \frac{\sqrt{2}}{\sqrt{36}}, 54.85 + 2.326 \times \frac{\sqrt{2}}{\sqrt{36}}\right]$$

$$= [54.85 - 0.56, 54.85 + 0.56]$$

$$= [54.29, 55.41]$$

Interpretation: - We are 95% confident that the true average weight of an egg lies between 54.38 and 55.32 grams. - We are 98% confident that the true average weight of an egg lies between 54.29 and 55.41 grams.

3. **At the 95% confidence level, can the average weight of an egg be assumed to be 54 grams? 56 grams?**

   Solution:

   Recall that the 95% confidence interval is $[54.38, 55.32]$.

   For 54 grams: 54 is not included in the interval $[54.38, 55.32]$.

   Interpretation: At the 95% confidence level, we can reject the hypothesis that the average weight of an egg is 54 grams.

   For 56 grams: 56 is not included in the interval $[54.38, 55.32]$.

   Interpretation: At the 95% confidence level, we can reject the hypothesis that the average weight of an egg is 56 grams.

   Important Note: Rejecting these values does not prove that the mean is exactly 54.85 grams, only that 54 and 56 grams are not plausible based on this data and confidence level.

4. **Same questions at the 99% confidence level.**

   Solution:

   Let's first calculate the 99% confidence interval:

$$z_{0.005} = 2.576$$

$$\left[54.85 - 2.576 \times \frac{\sqrt{2}}{\sqrt{36}}, 54.85 + 2.576 \times \frac{\sqrt{2}}{\sqrt{36}}\right]$$

$$= [54.85 - 0.62, 54.85 + 0.62]$$

$$= [54.23, 55.47]$$

   For 54 grams: 54 is included in the interval $[54.23, 55.47]$.

   Interpretation: At the 99% confidence level, we cannot reject the hypothesis that the average weight of an egg is 54 grams.

For 56 grams: 56 is not included in the interval $[54.23, 55.47]$.

Interpretation: At the 99% confidence level, we can reject the hypothesis that the average weight of an egg is 56 grams.

Practical implications: - At 95% confidence, we would consider 54 and 56 grams as implausible. - At 99% confidence, we would consider 54 grams plausible, but not 56 grams, showing the effect of different confidence levels on the hypothesis testing.

Exercise 3

It is assumed that the weight of a newborn is a normal variable with a standard deviation of 0.5 kg. The average weight of 49 children born in January 2016 in a Marseille maternity ward was 3.6 kg.

1. Estimate the average weight of a newborn in this maternity ward using a confidence interval at the 95% confidence level.

   Solution:

   We have the following information: - $\bar{x} = 3.6$ kg (sample mean) - $\sigma = 0.5$ kg (population standard deviation) - $n = 49$ (sample size)

   For a 95% confidence level, $z_{0.025} = 1.96$.

   The confidence interval is given by:

   $[\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}}]$

   $= [3.6 - 1.96\frac{0.5}{\sqrt{49}}, 3.6 + 1.96\frac{0.5}{\sqrt{49}}] = [3.6 - 0.14, 3.6 + 0.14] = [3.46, 3.74]$

   Interpretation: We can be 95% confident that the true average weight of newborns in this maternity ward lies between 3.46 kg and 3.74 kg.

2. What would be the confidence level of an interval with length 0.1 kg centered at 3.6 for this average weight?

   Solution:

   The half-length of the interval is 0.05 kg (0.1 kg / 2). We are looking for $z_{\alpha/2}$ such that:

   $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.05$

   $z_{\alpha/2} = \frac{0.05\sqrt{49}}{0.5} = 0.7$

   From the standard normal table, we find that this value corresponds to a probability of about 0.758.

   Therefore, the confidence level is: $2 * 0.758 - 1 = 0.516$, or approximately 51.6%.

   Interpretation: An interval of length 0.1 kg centered at 3.6 kg corresponds to a confidence level of about 51.6%, which is relatively low for practical applications.

3. Same question for an interval of length 1 kg.

   Solution:

   The half-length of the interval is 0.5 kg (1 kg / 2). We are looking for $z_{\alpha/2}$ such that:

   $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.5$

   $z_{\alpha/2} = \frac{0.5\sqrt{49}}{0.5} = 7$

   This z value is very high. From the standard normal table, we find that the corresponding probability is practically 1.

Therefore, the confidence level is practically 100%.

Interpretation: An interval of length 1 kg centered at 3.6 kg corresponds to an extremely high confidence level, almost 100%. This means we are almost certain that the true average weight lies within this interval, but the interval is so wide that it is not very informative in practice.

Exercise 4

Naturalists want to estimate the number N of individuals of an animal species living on an island. They capture 800 individuals, mark them, and then release them. One month later, after allowing the individuals to mix back into the overall population, they recapture 1000 individuals, among which 250 marked individuals are found. Produce an interval estimate, at the 95% confidence level, of the number of individuals of this animal species present on the island.

Solution:

This situation is a classic example of the capture-recapture method, using the Lincoln-Petersen estimator.

Let: - $M = 800$ be the number of initially marked animals - $C = 1000$ be the number of animals captured during the second capture - $R = 250$ be the number of marked animals recaptured

The Lincoln-Petersen estimator gives:

$\hat{N} = \frac{MC}{R} = \frac{800 \times 1000}{250} = 3200$

For the confidence interval, we use Bailey's normal approximation:

$SE(\hat{N}) = \sqrt{\frac{\hat{N}^2(C-R)}{CR}}$

$SE(\hat{N}) = \sqrt{\frac{3200^2(1000-250)}{1000 \times 250}} \approx 173.21$

The 95% confidence interval is therefore:

$[\hat{N} - 1.96 \times SE(\hat{N}), \hat{N} + 1.96 \times SE(\hat{N})] = [3200 - 1.96 \times 173.21, 3200 + 1.96 \times 173.21] \approx [2860, 3540]$

Interpretation: We can be 95% confident that the true number of individuals of this species on the island lies between approximately 2860 and 3540.

Note: This method assumes the population is closed (no births, deaths, or migrations) and that all individuals have the same probability of being captured.

Exercise 5

A police patrol conducts speed checks on the side of a road where the speed limit is 70 km/h. For the first sixteen vehicles checked, the recorded speeds are as follows: 49; 71; 78; 58; 83; 74; 64; 86; 56; 65; 55; 64; 65; 72; 87; 56

1. Estimate the average speed of the checked vehicles.

   Solution:

   The average speed is given by:

   $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{49+71+...+56}{16} = \frac{1083}{16} = 67.6875$ km/h

   Interpretation: The estimated average speed of the checked vehicles is approximately 67.69 km/h.

2. Estimate the average speed of the checked vehicles using a confidence interval at the 95% confidence level.

   Solution:

   To construct the confidence interval, we must first calculate the sample standard deviation:

$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx 11.8741$ km/h

The 95% confidence interval is given by:

$[\bar{x} - t_{0.975,15}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.975,15}\frac{s}{\sqrt{n}}]$

Where $t_{0.975,15} = 2.131$ (value from the Student's t-distribution table for 15 degrees of freedom)

$[67.6875 - 2.131\frac{11.8741}{\sqrt{16}}, 67.6875 + 2.131\frac{11.8741}{\sqrt{16}}] \approx [61.7, 73.7]$ km/h

Interpretation: We can be 95% confident that the true average speed of vehicles on this road lies between 61.7 km/h and 73.7 km/h.

3. Estimate the proportion of vehicles not respecting the speed limit.

   Solution:

   Out of the 16 vehicles, 7 exceed the 70 km/h limit. The point estimate of the proportion is:

   $\hat{p} = \frac{7}{16} = 0.4375 = 43.75\%$

   Interpretation: It is estimated that 43.75% of vehicles do not respect the speed limit.

4. Estimate the same proportion using a confidence interval at the 95% confidence level.

   Solution:

   For a proportion, the 95% confidence interval is given by:

   $[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

   With $\hat{p} = 0.4375$ and $n = 16$:

   $[0.4375 - 1.96\sqrt{\frac{0.4375(1-0.4375)}{16}}, 0.4375 + 1.96\sqrt{\frac{0.4375(1-0.4375)}{16}}]$

   $\approx [0.1951, 0.6799]$

   Interpretation: We can be 95% confident that the true proportion of vehicles not respecting the speed limit lies between 19.51% and 67.99%. Note that this interval is very wide due to the small sample size.

Exercise 6

Two candidates are competing in an election. A survey conducted among 250 voters shows the first candidate leading with 55% of the votes against 45% for the opponent.

1. Estimate the proportion of voters in favor of the first candidate at the 95% confidence level. What can be concluded?

   Solution:

   The point estimate of the proportion is $\hat{p} = 0.55$.

   The 95% confidence interval is given by:

   $[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

   $[0.55 - 1.96\sqrt{\frac{0.55(1-0.55)}{250}}, 0.55 + 1.96\sqrt{\frac{0.55(1-0.55)}{250}}]$

   $\approx [0.4882, 0.6118]$

   Interpretation: We can be 95% confident that the true proportion of voters in favor of the first candidate lies between 48.82% and 61.18%.

Conclusion: Since the confidence interval includes 50%, we cannot say with certainty that the first candidate will win the election, despite their lead in the survey. The result remains uncertain.

2. Same question, assuming the survey was conducted among 1000 people.

   Solution:

   With $n = 1000$ and $\hat{p} = 0.55$:

   $[0.55 - 1.96\sqrt{\frac{0.55(1-0.55)}{1000}}, 0.55 + 1.96\sqrt{\frac{0.55(1-0.55)}{1000}}]$

   $\approx [0.5191, 0.5809]$

   Interpretation: We can be 95% confident that the true proportion of voters in favor of the first candidate lies between 51.91% and 58.09%.

   Conclusion: In this case, the confidence interval is entirely above 50%. We can therefore say with 95% confidence that the first candidate is likely to win the election.

3. Determine the number of people to survey in order to ensure the first candidate's probable victory at the 95% confidence level.

   Solution:

   To ensure a probable victory, the lower bound of the confidence interval must be greater than 50%. We are looking for n such that:

   $0.55 - 1.96\sqrt{\frac{0.55(1-0.55)}{n}} > 0.50$

   $1.96\sqrt{\frac{0.55(1-0.55)}{n}} < 0.05$

   $n > \frac{1.96^2 \times 0.55 \times 0.45}{0.05^2} \approx 380.3$

   Therefore, at least 381 people need to be surveyed to ensure a probable victory for the first candidate with a 95% confidence level.

   Interpretation: With a sample size of 381 or more, if the survey still shows 55% in favor of the first candidate, we could state with 95% confidence that this candidate is likely to win the election.

Exercise 7

An industrial company that manufactures mechanical parts has implemented a quality control process for production. During a production check of certain parts conducted on a sample of 150 parts, 17 parts were found to be defective.

1. Estimate the proportion of defective parts in the company's production.

   Solution:

   The point estimate of the proportion of defective parts is given by:

   $\hat{p} = \frac{\text{number of defective parts}}{\text{sample size}} = \frac{17}{150} \approx 0.1133$, or 11.33%

   Interpretation: It is estimated that 11.33% of the produced parts are defective.

2. Estimate this proportion using a confidence interval at the 95% confidence level.

   Solution:

   For a proportion, the 95% confidence interval is given by:

$[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

With $\hat{p} = 0.1133$ and $n = 150$:

$[0.1133 - 1.96\sqrt{\frac{0.1133(1-0.1133)}{150}}, 0.1133 + 1.96\sqrt{\frac{0.1133(1-0.1133)}{150}}]$

$\approx [0.0627, 0.1639]$

Interpretation: We can be 95% confident that the true proportion of defective parts in the production lies between 6.27% and 16.39%.

3. At the same 95% confidence level, how many parts need to be checked to limit the margin of error to 2% (interval length equal to 4%)? 1% (interval length equal to 2%)? 0.5% (interval length equal to 1%)? Comment.

   Solution:

   The general formula for the required sample size is:

   $n = \frac{z^2\hat{p}(1-\hat{p})}{e^2}$

   where e is the margin of error (half-length of the interval) and $z = 1.96$ for a 95% confidence level.

   For a margin of error of 2%: $n = \frac{1.96^2 \times 0.1133(1-0.1133)}{0.02^2} \approx 482$

   For a margin of error of 1%: $n = \frac{1.96^2 \times 0.1133(1-0.1133)}{0.01^2} \approx 1928$

   For a margin of error of 0.5%: $n = \frac{1.96^2 \times 0.1133(1-0.1133)}{0.005^2} \approx 7711$

   Comment: 1) It can be seen that to halve the margin of error, the sample size must be quadrupled. This illustrates the trade-off between precision and cost/time of the study. 2) To achieve a very small margin of error (0.5%), a very large sample is needed, which can be costly and time-consuming. 3) In practice, a balance must be found between the desired precision and the available resources for the study.

Exercise 8

A survey published by a polling institute reveals that 60%±2% of the French population is against Turkey joining the European Union in the near future. The survey was conducted on a sample of 914 adults. What confidence level did the polling institute use?

Solution:

The margin of error of 2% corresponds to half the length of the confidence interval. Let z be the desired confidence coefficient. We have:

$z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02$

where $\hat{p} = 0.60$ and $n = 914$.

Solving for z:

$z = \frac{0.02}{\sqrt{\frac{0.60(1-0.60)}{914}}} \approx 1.9559$

This value corresponds to a confidence level of approximately 95%.

To verify, we can calculate the corresponding probability: $P(-1.9559 < Z < 1.9559) \approx 0.9495$

Interpretation: The polling institute used a confidence level of approximately 95%. This is a standard confidence level in many statistical studies, offering a good balance between reliability and precision.

Exercise 9

A large company takes special care of its employees' health. To this end, it offered all employees a health examination. The cholesterol level of 100 employees was measured (it is assumed that these 100 people were randomly selected and all accepted the test). The results are as follows:

[Insert data table here]

1. Calculate the sample mean and sample standard deviation s of the cholesterol level.

   Solution:

   Without the actual data, we cannot calculate the exact values. However, here is the method:

   Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

   Sample standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

   Where $n = 100$ and x_i are the individual cholesterol levels.

2. Estimate the mean μ and standard deviation σ of the cholesterol level in the entire company.

   Solution:

   The estimator of the population mean is the sample mean: $\hat{\mu} = \bar{x}$

   The estimator of the population standard deviation is the sample standard deviation: $\hat{\sigma} = s$

   These estimators are unbiased and consistent, meaning they provide a good approximation of the population parameters, especially with a large sample like this.

3. Estimate the cholesterol level of the company's employees using a confidence interval at the 95% confidence level.

   Solution:

   The 95% confidence interval for the mean is given by:

   $[\bar{x} - t_{0.975,99}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.975,99}\frac{s}{\sqrt{n}}]$

   where $t_{0.975,99} \approx 1.984$ (value from the Student's t-distribution table for 99 degrees of freedom)

   Without the actual values of $\bar{x}$ and s, we cannot calculate the exact interval. However, this formula would give us an interval within which we can be 95% confident that the true mean cholesterol level of all employees lies.

4. Determine the minimum sample size required so that the amplitude of the confidence interval calculated in the previous question is less than 100.

   Solution:

   The amplitude of the interval must be less than 100:

   $2 \times t_{0.975,n-1}\frac{s}{\sqrt{n}} < 100$

   Approximating $t_{0.975,n-1}$ as 1.96 (value for the normal distribution, a good approximation for large samples), we get:

   $n > (\frac{2\times 1.96\times s}{100})^2$

   Without knowing s, we cannot calculate the exact value of n. The minimum sample size will depend on the variability of cholesterol levels in the population.

   For example, if $s = 50$ (a reasonable estimate for cholesterol level):

   $n > (\frac{2\times 1.96\times 50}{100})^2 \approx 384$

   In this case, a sample of at least 384 people would be required.

Exercise 10

The managers of a university program wish to conduct a survey on the average annual salary of graduates after their studies. They roughly estimate the standard deviation of the monthly salary to be 3,500€.

1. How many graduates should they survey if they want to obtain an estimate of the salary that does not deviate by more than 500€ from the true salary obtained at the 95% confidence level?

   Solution:

   For a 95% confidence level, $z_{0.975} = 1.96$. The desired margin of error is 500€. The monthly standard deviation is 3500€, so the annual standard deviation is $3500 \times \sqrt{12} \approx 12124$.

   We are looking for n such that:

   $$1.96 \frac{12124}{\sqrt{n}} = 500$$

   $$n = \left(\frac{1.96 \times 12124}{500} \right)^2 \approx 2261$$

   Therefore, at least 2261 graduates need to be surveyed.

   Interpretation: With a sample of this size, we can be 95% confident that the estimated average annual salary will not deviate by more than 500€ from the true value.

2. Same question for a deviation of no more than 100€.

   Solution:

   We use the same method with a margin of error of 100€:

   $$n = \left(\frac{1.96 \times 12124}{100} \right)^2 \approx 56525$$

   Therefore, at least 56525 graduates need to be surveyed.

   Interpretation: To achieve five times greater precision (100€ instead of 500€), the required sample size must be 25 times larger. This illustrates the high cost of increasing precision.

3. The survey costs 10€ per person surveyed. The managers have a budget of 30,000€. With what precision will they be able to estimate the average salary of their graduates?

   Solution:

   With a budget of 30,000€ and a cost of 10€ per person, they can survey:

   $$n = \frac{30000}{10} = 3000 \text{ people}$$

   The margin of error e for this sample size is given by:

   $$e = z_{0.975} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{12124}{\sqrt{3000}} \approx 433.85$$

   Interpretation: With their budget, the managers will be able to estimate the average salary with a precision of approximately ±434€ at the 95% confidence level. In other words, they can be 95% confident that their estimate will not deviate by more than 434€ from the true average salary of the graduates.