1st Semester, 2021-2022

3rd Year, Math & CompSci

Mathematical Statistics

Seminar Exercises: Week 10

Recap. *Throughout this class, $X_1, X_2, \ldots, X_n, \ldots$ will be i.i.d. random variables that follow the distribution of a given characteristic X.*

How to find a $100(1-\alpha)\%$ confidence interval(CI):

The problem: *Let θ be a parameter of the distribution of X. We want to find two statistics θ_L and θ_U, depending on $X_1, \ldots, X_n$, so that:*

$$P(\theta_L \leq \theta \leq \theta_U) = 1 - \alpha$$

i.e. we want to estimate θ with a ***confidence level*** *of $1-\alpha$.*

Step 1: Find the right ***pivotal quantity*** *(in short,* ***pivot****) Z for the job (see below);*

Step 2: Use Octave/Matlab to find the quantiles $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ corresponding to the law of Z; if the law is symmetrical, e.g. $\mathcal{N}$, T, then

$$z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$$

Step 3: Write $z_{\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}$;

Step 4: From the inequality above replace θ in terms of Z to get the confidence interval.

The confidence intervals for every situation:

- *One population, $X \sim \mathcal{N}(\mu,\sigma)$ or $n > 30$, known variance σ^2:*

$$\mu \in \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],\ z_\beta = norminv(\beta)$$

- *One population, $X \sim \mathcal{N}(\mu,\sigma)$ or $n > 30$, unknown variance:*

$$\mu \in \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right],\ z_\beta = tinv(\beta, n-1)$$

- *One population, $X \sim \mathcal{N}(\mu,\sigma)$ or $n > 30$:*

$$\sigma \in \left[\sqrt{\frac{(n-1)s^2}{z_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)s^2}{z_{\frac{\alpha}{2}}}}\right],\ z_\beta = chi2inv(\beta, n-1)$$

- *Two populations, $X_{(1)} \sim \mathcal{N}(\mu_1,\sigma_1)$ and $X_{(2)} \sim \mathcal{N}(\mu_2,\sigma_2)$ or $n_1 + n_2 > 40$, σ_1 and σ_2 are known.*

$$\mu_1 - \mu_2 \in \left[\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{X_1} - \overline{X_2} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

$z_\beta = norminv(\beta)$

- *Two populations, $X_{(1)} \sim \mathcal{N}(\mu_1,\sigma_1)$ and $X_{(2)} \sim \mathcal{N}(\mu_2,\sigma_2)$ or $n_1 + n_2 > 40$, $\sigma_1 = \sigma_2$ unknown.*

$$\mu_1 - \mu_2 \in \left[\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{X_1} - \overline{X_2} - z_{\frac{\alpha}{2}} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

$z_\beta = tinv(\beta, n_1 + n_2 - 2)$

- *Two populations, $X_{(1)} \sim \mathcal{N}(\mu_1,\sigma_1)$ and $X_{(2)} \sim \mathcal{N}(\mu_2,\sigma_2)$ or $n_1 + n_2 > 40$, σ_1 and σ_2 unknown.*

$$\mu_1 - \mu_2 \in \left[\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{X_1} - \overline{X_2} - z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

$z_\beta = tinv(\beta, n)$

- *Two populations, $X_{(1)} \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_{(2)} \sim \mathcal{N}(\mu_2, \sigma_2)$.*

$$\frac{\sigma_1}{\sigma_2} \in \left[\sqrt{\frac{1}{z_{1-\frac{\alpha}{2}}}} \cdot \frac{s_1}{s_2}, \sqrt{\frac{1}{z_{\frac{\alpha}{2}}}} \cdot \frac{s_1}{s_2}\right]$$

$$z_\beta = \mathit{finv}(\beta, n_1 - 1, n_2 - 1)$$

- *One proportion, $n > 30$:*

$$p \in \left[\overline{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \overline{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right], \ z_\beta = \mathit{norminv}(\beta)$$

- *Two proportions, $n_1 + n_2 > 40$:*

$$p_1 - p_2 \in \left[\overline{p}_1 - \overline{p}_2 - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}, \right.$$
$$\left. \overline{p}_1 - \overline{p}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}\right]$$
$$z_\beta = \mathit{norminv}(\beta)$$

<u>*How to find the sample size that ensures a certain length for the confidence interval:*</u>

If the pivot Z follows a normal distribution and $\overline{\theta}$ is an unbiased estimator for θ, then the confidence interval is:

$$\theta \in \left[\overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}, \overline{\theta} - \sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right]$$

Thus, the length of the confidence interval is:

$$\sigma_{\overline{\theta}} \cdot \left(z_{1-\frac{\alpha}{2}} - z_{\frac{\alpha}{2}}\right) = 2\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} = -2\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}$$

We bound this length and get an expression involving the sample size n.

*The **margin of error** is half of the length of the confidence interval.*

Exercise 1. Consider the following sample data for the weight (in kg) of the people in a certain city:

$$67.6 \quad 84.7 \quad 88.1 \quad 68.0 \quad 64.2 \quad 75.9 \quad 69.2 \quad 71.3 \quad 82.4 \quad 78.6$$

Assume that the weight is a characteristic that follows the normal distribution. Find 95% confidence intervals for:

(a) the mean value of the weight, given that the standard deviation of the weight is 10 (kg);

(b) the mean value of the weight, given that the standard deviation of the weight is unknown;

(c) the standard deviation of the weight.

Exercise 2. In a pre-election poll, we are interested in the proportion p of people who plan to vote for candidate A against candidate B.

(a) Find a 95% confidence interval for p, given that 64 persons out of a random sample of 100 persons support A;

(b) Estimate the minimum number of persons polled to obtain a confidence interval for p with a marginal error less than 2.5% and a confidence level at least 95%.

Exercise 3. A vending machine contains the following numbers of bills:

1 RON	5 RON	10 RON
25	15	10

Assume that each client pays the vending machine with only one bill. Find 95% confidence intervals for:

(a) the mean value of the amount of money paid by a client;

(b) the standard deviation of the amount of money paid by a client;

(c) the proportion of clients that pay the vending machine with a 5 RON bill.

Exercise 4. In an orange juice factory, cans are filled by a machine according to the normal distribution.

(a) Consider a random sample of 100 cans that contain a total amount of 24.8 liters of orange juice. Find a 95% confidence interval for the mean value of the amount of orange juice in a can, given that the standard deviation for the filling machine is 5 ml;

(b) Estimate the minimum number of cans in the sample to obtain a 95% confidence interval for the mean value m of the amount of orange juice in a can with a marginal error less than 1 ml, given that the standard deviation for the filling machine is 5 ml;

(c) Find a 95% confidence interval for the proportion of cans that contain between 249 ml and 251 ml, given the following sample:

251.2	250.2	249.6	247.2	250.4	250.2	251.4	251.0	250.3
250.4	251.3	250.5	249.1	248.4	251.3	250.5	250.2	251.7
250.0	250.6	250.0	250.1	247.7	249.7	249.2	249.8	249.3
249.5	249.5	249.7						

Exercise 5. Consider the following random sample data for the height (in cm) of the 10-year-old children of a city:

141 138 136 142 145 140 139 137 132 140

Next, consider another independent random sample data for the height (in cm) of the 15-year-old children of the same city:

172 167 168 165 163 171 168 166 161 169

Assume the height is a characteristic that follows the normal distribution. Find 95% confidence intervals for:

(a) the difference between the mean values of the heights, given that the standard deviation of the height for 10-year-old children is 3 cm and for 15-year-old children is 4 cm;

(b) the difference between the mean values of the heights, given that the standard deviations of the heights for 10-year-old and for 15-year-old children are equal but unknown;

(c) the difference between the mean values of the heights, given that the standard deviations of the heights for 10-year-old and for 15-year-old children are unknown;

(d) the ratio between the standard deviations of the heights.

Exercise 6. Consider the following random sample data for the height (in cm) of the 10-year-old children of a city:

141 138 136 142 145 140 139 137 132 140

The previously chosen children are measured again after 5 years and we have the following corresponding data:

171 166 164 159 165 160 162 158 155 160

Assume the height is a characteristic that follows the normal distribution. Find 95% confidence intervals for the difference between the mean values of the heights.

Exercise 7. A new type of battery for a certain brand of laptop is tested in order to replace the old one. 40 laptops were tested with the old type of battery and a sample mean of 3.5 hours and a sample standard deviation of 0.1 hours were recorded. 30 laptops were tested with the new type of battery and a sample mean of 4 hours and a sample standard deviation of 0.2 hours were recorded. Find a 95% confidence interval for the difference between the mean values of the two battery lifetimes, if the corresponding standard deviations are unknown and:

(i) equal

(ii) unequal

Exercise 8. A new medication for isolated systolic hypertension was tested on a sample of 10 patients. The following are the systolic blood pressures (in mmHg) before and after administration of the drug:

145	147	152	145	156	141	151	148	144	151
125	136	121	121	128	124	131	125	127	132

Assume that the systolic blood pressure is a characteristic that follows the normal distribution. Find a 95% confidence interval for the difference between the mean values of the systolic blood pressures before and after administration of the drug.