Laboratory 5 - Inferential statistics

A statistical population is a set of individuals[1] whose attribute (weight, height, etc) has some random variation. We infer about these paramaters like follows

- we choose a simple random sample;
- we compute some statistics using this sample;
- using mathematical statistics and probability theory, we formulate an assertion about the parameter of interest.

A normally distributed variable with parameters μ and σ^2 has the following density function

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right].$$

If $X : N(\mu, \sigma^2)$, then

$$\boxed{M(X) = \mu} \text{ and } \boxed{D^2(X) = \sigma^2}$$

$N(0, 1)$ is called the *normal standard* distribution. The values of a standard normally distributed variable have the following spread:

%68 of them are found in an interval centered in the mean of length equal with two standard deviations;

%95 of them are found in an interval centered in the mean of length equal with four standard deviations;

%99.7 of them are found in an interval centered in the mean of length equal with six standard deviations;

1 Population mean estimation: sample mean

RStudio. Don't forget to set the working directory: **Session** → **Set Working Directory** → **Choose Directory**.

Let us consider a population having the mean μ and the variance σ^2; we measure an attribute for this population[2]: X. We get a simple random sample of size n from the population: $X_1, X_2, \ldots, X_n$. These values can be thought as independent, identical distributed (like X). The sample mean is

$$\overline{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

it is a statistic (a value), but can be viewed as a random variable also. Properties of the sample mean:

- $\overline{x}_n$ is an unbiased estimator of the population's expectation, μ.
- as a random variable:

$$\boxed{M(\overline{x}_n) = \mu,\ D^2(\overline{x}_n) = \frac{\sigma^2}{n}}$$

[1] In broad sense.
[2] X is a random variable with mean μ and variance σ^2.

- if the population is normally distributed, $N(\mu, \sigma^2)$, then the sample mean follows also a normal law, $N\left(\mu, \dfrac{\sigma^2}{n}\right)$;
- if the population is not normally distributed, but the size of the sample is large enough ($n \geqslant 30$), then the sample mean distribution is almost a normal variable: $N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

A function for the sample mean of a sample from a file:

```
selection_mean <- function(filename) {
   x = scan(filename);
   m = mean(x)
}
selection_mean("sample.txt")
```

RStudio. The file having the name *filename* must be in the working directory.

Exercise to work.

1.1 Write in a script the function from above. Then call this function on the file *history.txt*.

2 Confidence intervals for the mean of a population with known variance

We consider a population with known variance σ^2. We want an interval in which the population's mean μ (which is unknown) is found with high probability (0.90, 0.95 sau 0.99). Such an interval is the following

$$\left(\overline{x}_n - z^* \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x}_n + z^* \cdot \frac{\sigma}{\sqrt{n}}\right)$$

where z^*, the critical value, is determined like follows

$$z^* = -qnorm\left(\alpha/2, mean = 0, sd = 1\right) = qnorm\left(1 - \alpha/2, mean = 0, sd = 1\right)$$

and α is equal with $1-$ the confidence level. The sample mean, if it is not given, can be computed:

$$\overline{x}_n = mean(\text{sample\_data})$$

Solved exercise. The lifespan of a certain type of battery follows a normal law with variance of 9 hours. For a sample of 100 batteries we measure a sample mean of 20 hours. Find a confidence interval of 90% level of confidence for the true lifespan mean.

```
> alfa = 0.1
> sample_mean = 20
> n = 100
> sigma = sqrt(9)
> critical_z = qnorm(1 - alfa/2, 0, 1)
> a = sample_mean - critical_z*sigma/sqrt(n)
> b = sample_mean + critical_z*sigma/sqrt(n)
> interval = c(a, b)
> interval
```

The result is the interval $[19.50654, 20.49346]$.

Exercises to work

2.1 Write in a script a function (called **zconfidence_interval**) which has to compute the confidence interval (function's parameters will be: n, $\overline{x}_n$, α etc). This function will be used for solving the following exercises.

2.2 We want to find out a 90% confidence interval for the true mean of a population with known variance ($\sigma^2 = 100$). We use a simple random sample with 25 individuals whose sample mean is 67.53.

2.3 In a public institution there exists a coffee machine; the volume of a cup of coffee follows a normal law with standard deviation $\sigma = 0.5$ oz. For a sample of 50 cups of coffee we measure a sample mean of 5 oz. Determine a 95% confidence interval for the true average volume of a cup of coffee.

2.4 In a desperate try to General Electric, ACME company introduces a new type of electrical bulbs. ACME produces a lot of 100 bulbs whose sample mean is 1280 hour lifespan with a standard deviation of the entire population of bulbs of 140 hours. Find a 99% confidence interval for the true mean of the lifespan for this type of electrical bulbs.

2.5 We measure the weight of 35 athletes and we find an average of 60 kg. Suppose that the standard deviation of the entire population is 5 kg. Find 90%, 95% and 99% confidence intervals for the true mean of the population weight. Which is larger: the interval of 95% or that of 99% level of confidence? Why?

2.6 Modify the function from II.1 for the situation when the sample is given in a file (you have to open the file, compute the sample mean, and the sample size). Run this function on the history.txt file in order to find a 95% confidence interval ($\sigma = 5$).

3 Confidence intervals for the mean of a population with unknown variance

We consider a population with unknown variance. We use as an estimate for the standard deviation of the population, σ, the sample standard deviation, s. In this case the score, $t = \dfrac{\overline{x}_n - \mu}{s/\sqrt{n}}$, is Student distributed with $(n-1)$ degrees of freedom, $t(n-1)$: $t(n-1)$.

We look for an interval to which the true mean of the population, μ, belongs with a prescribed probability (0.9, 0.95 or 0.99). Such an interval is

$$\left(\overline{x}_n - t^* \cdot \frac{s}{\sqrt{n}},\ \overline{x}_n + t^* \cdot \frac{s}{\sqrt{n}}\right)$$

where t^*, the critical value, can be determined like follows

$$t^* = -qt\left(\alpha/2, n-1\right) = qt\left(1-\alpha/2, n-1\right)$$

α is $1-$ the level of confidence, and s is the standard deviation of the sample. When we know only the data from the sample, $\overline{x}_n$ and s will be computed like follows

$$\overline{x}_n = mean(\text{sample}),\ s = sd(\text{sample})$$

In what follows we will use an estimate for the *standard error of the mean* $se = \frac{s}{\sqrt{n}}$.

Solved exercise. A toy-company wants to know how appealing are its products to children. A random sample of children 60 is taken and they are asked to answer with a value from 0 to 5 about their interest in the products. We determine a sample mean of 3.3, with a standard deviation of $s = 0.4$. Find a confidence interval for the average grade, for the entire population (95% level of confidence).

```
> alfa = 0.05
> sample_mean = 3.3
> n = 60
> s = 0.4
> se = s/sqrt(n)
> critical_t = qt(1 - alfa/2, n - 1)
> a = sample_mean - critical_t*se
> b = sample_mean + critical_t*se
> interval = c(a, b)
> interval
```

The result is the interval $[3.19667, 3.40333]$.

Exercises to work

3.1 Write in a script a function (called **t_conf_interval**) which will compute the confidence interval like above (the parameters of the function will be: n, $\overline{x}_n$, α etc). This function will be used to solve the following exercises.

3.2 196 randomly chosen students were asked how much they pay for on-line shopping in a given week. The sample mean was 44.65$, with a sample variance of $s^2 = 2.25$. Find a confidence interval for this average amount of money in a given week for all students (99% level of confidence). (Assume the normality of the population)

3.3 A candy producing company considers that the level of sugar in its products can have a value between 1 and 20 and follows a normal law. We choose a 49 products random sample. The sample mean is 12 with a sample standard deviation of 1.75.

(a) Find the confidence intervals of 99% and 95% for the true mean sugar level.

(b) After some changes we choose again a 49 products random sample. This time the sample mean is 13.5 with a sample standard deviation of 1.25. Find the confidence interval of 95% for the true mean sugar level after the changes.

3.4 Change the above function for the case when the sample is given in a file (we must compute the sample mean, the sample standard deviation, and the sample size). Use this function for finding the confidence intervals of 95% and 99% for the sample from history.txt file.

3.5 From a simple random sample extracted from a normal population with unknown variance we measure the following data:

12 11 12 10 11 12 13 12 11 11 13 14 10

Determine the confidence intervals of 90%, 95%, and 99% for the true mean of the population.

Statistical hypotheses testing

We have a statistical population with an unknown distribution. A statistical test about some unknown features of the distribution[3] follows a general procedure

- formulate the null hypothesis, H_0, which completely establish the distribution of the population.
- the null hypothesis is attacked by an alternative hypothesis, H_a, which assumes a different state of the distribution.
- when we have significant enough evidences the **null hypothesis**, H_0, is rejected and the **alternative hypothesis**, H_a, is accepted.
- if the evidences are not statistically significant, then the **null hypothesis,** H_0**, cannot be rejected**, (**a statistical test doesn't end by accepting the null hypothesis**).

While performing a statistic test we cam make two types of errors

- **Type I errors:** we reject the null hypothesis although H_0 is true - this error is caused by an excessive confidence.
- **Type II errors:** we fail to reject null hypothesis, but H_0 is false - this error is caused by an excessive skepticism.

	H_0 is not rejected	H_0 is rejected
H_0 is true	correct	type I error
H_0 is false	type II error	correct

4 z proportions test

Consider a variable X which numbers the successes from n trials. X binomial distributed - $X : B(n, p)$. The proportion test infers about the probability p. We denote by $p' = \dfrac{X}{n}$ the frequency from the sample. Since $M(X) = np$ and $D^2(X) = np(1-p)$, we have

$$M(p') = p \text{ and } D^2(p') = \frac{p(1-p)}{n}.$$

For large enough n ($n \geqslant 20$ and $np \geqslant 5$) p' approximately follows a normal distribution. The statistic $z = \dfrac{p' - p}{\sqrt{p(1-p)/n}}$ has a standard normal distribution: $N(0, 1)$.

The proportion test is performed like follows:

1. formulate the null hypothesis (which says that p has a certain value):

$$\boxed{H_0 : \ p = p_0}$$

[3]Like the mean and the variance.

2. formulate the alternative hypothesis - that can be of three types:

$$\boxed{H_a:\ p < p_0} \quad \text{(left tailed) or}$$

$$\boxed{H_a:\ p > p_0} \quad \text{(right tailed) or}$$

$$\boxed{H_a:\ p \neq p_0} \quad \text{(two-tailed)}$$

3. give a significance level: α (usually is 1% or 5%);

4. compute the score :

$$\boxed{z = \frac{p' - p_0}{\sqrt{p_0(1-p_0)/n}}}$$

5. determine the critical value z^*:

$$\boxed{z^* = qnorm(\alpha, 0, 1)} \quad \text{for left tailed } H_a (z^* < 0),$$

$$\boxed{z^* = qnorm(1-\alpha, 0, 1)} \quad \text{for right tailed } H_a (z^* > 0),$$

$$\boxed{z^* = -qnorm(\alpha/2, 0, 1) = qnorm(1-\alpha/2, 0, 1)} \quad \text{for two-tailed } H_a (z^* > 0).$$

6. **the null hypothesis, H_0, is rejected if**

$$\boxed{z < z^*} \quad \text{for left tailed } H_a \text{ or}$$

$$\boxed{z > z^*} \quad \text{for right tailed } H_a \text{ or}$$

$$\boxed{|z| > |z^*|} \quad \text{for two-tailed } H_a,$$

otherwise we will say that **there is not sufficient evidence to reject H_0 and accept H_a**.

Solved exercise. A certain politician assumes that he will receive at most 60% of the votes from his electoral college. From a random sample of 100 electors 63 claim that they voted with this politician. Can we reject the politician's assertion? (1% level of significance)

```
> alfa = 0.01
> n = 100
> succeses = 63
> p_prim = succeses/n
> p0 = 0.6
> z_score = (p_prim - p0)/sqrt(p0(1 - p0)/n)
> critical_z = qnorm(1 - alfa, 0, 1)
> z_score
> critical_z
```

The result is $z = 0.61237 < z^* = 2.32634$, we cannot reject the politician's assertion.

Exercises to work

4.1 Write in a script a function (called **test_proportion**) which has to compute and return the critical value and the score for the proportions test (the parameters of this function will be α, n, x - the number of successes, p_0). This function will be used to solve the following exercises.

4.2 It is supposed that 10% of the population is left-handed. But some believe that the proportion of women left handers is slightly smaller. From a random sample of 150 women 13 were left handers. Can we say (with 1% level of significance) that the proportion of left-handed women is smaller than 10%?

4.3 In 2022 it was estimated that 11.2% of U. S. adults were smokers. U. S. Department of Health believe that the proportion of women smokers is smaller. From a random sample of 200 women 21 were smokers. Can we say (with 1% and 5% level of significance) that the proportion of women smokers is smaller than 11.2%?

4.4 The manufacturer of a patent medicine claims that it is at least 90% effective in relieving an allergy. From a sample of 120 people (independently selected at random) who had the allergy, the medicine provided relief for 110 people. With 1% and 5% level of significance, determine whether the manufacturer's claim is illegitimate.

4.5 Suppose that the proportion of defective components (for a certain product) is 10%. We want to test if this proportion was increased. A random sample of 150 contains 20 defective components. Can we say (with 5% level of significance) that the proportion of defective components is greater than 10%?