Laboratory 6 - Inferential statistics

I. Z-test - Inference for the mean of a population with known variance

We consider a statistical population whose variance (σ^2) is known. For a given random sample with sample mean $\overline{x}_n$, if the population is normally distributed or the sample is large enough ($n \geqslant 30$) the score $z = \dfrac{\overline{x}_n - \mu}{\sigma/\sqrt{n}}$ is standard normally distributed: $N(0,1)$.

The Z-test is performed like follows:

1. formulate the null hypothesis, which says that the mean of the population has a certain value:
$$\boxed{H_0 : \ \mu = \mu_0}$$
2. formulate the alternative hypothesis; we can have three different types of alternative hypothesis
$$\boxed{H_a : \ \mu < \mu_0} \quad \text{(left asymmetric) or}$$
$$\boxed{H_a : \ \mu > \mu_0} \quad \text{(right asymmetric) or}$$
$$\boxed{H_a : \ \mu \neq \mu_0} \quad \text{(symmetric hypothesis).}$$
3. choose a level of significance: α (usually 1% or 5%);
4. compute the z-score:
$$\boxed{z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}}$$
5. determine the critical value z^*:
$$\boxed{z^* = qnorm(\alpha)} \text{ for left asymmetric } H_a \ (z^* < 0),$$
$$\boxed{z^* = qnorm(1-\alpha)} \text{ for right asymmetric } H_a \ (z^* > 0),$$
$$\boxed{z^* = -qnorm(\alpha/2) = qnorm(1-\alpha/2)} \text{ for symmetric } H_a \ (z^* > 0).$$
6. **the null hypothesis H_0 is rejected if**
$$\boxed{z < z^*} \text{ for left asymmetric } H_a,$$
$$\boxed{z > z^*} \text{ for right asymmetric } H_a,$$
$$\boxed{|z| > |z^*|} \text{ for symmetric } H_a,$$
otherwise we say that **there is not sufficient evidence at the α level of significance** to reject the null hypothesis (**we fail to reject** H_0).

Solved Exercise. An electrical bulb producer want to test with 5% level of significance the assertion that the average lifespan of its bulbs is grater than 810 hours (we know that the population standard deviation is $\sigma = 50$ hours). A random sample of 200 bulbs is chosen and a sample mean of 816 hours is computed. Can we accept the producer's assertion?

```
> alfa = 0.05
> population_mean = 810
> sample_mean = 816
> n = 200
> sigma = 50
> critical_z = qnorm(1- alfa)
> z_score = (sample_mean - population_mean)/(sigma/sqrt(n))
> critical_z
> z_score
```

The score is $z = 1.69705 > z^* = 1.64485$ hence the null hypothesis can be rejected and we can accept that the true population mean is greater than 810 hours.

Exercises to work

I.1 Write a function (called **z_test**) which will compute and return the critical value and the score of a Z-test (the parameters of the function will be: the type of the alternative hypothesis, n, μ_0, $\overline{x}_n$, α, σ etc). This function will be used to solve the following exercises.

I.2 From past years we know that the students results for a math test follow a normal law with mean 75 and variance 17. The math department wants to know if the current year students have an unusual behavior: for a random sample of 36 students the sample mean is 85. Is this result due only to the chance, or there is an unusual behavior of this year students regarding the math test? (1%)

I.3 The boxes of a certain type of detergent indicate a weight of 21 oz. A consumers agency wants to test this weight with 1% level of significance. For 100 randomly chosen boxes it finds a sample mean weight of 20.5 oz. We know that the weight follows a normal law with a standard deviation of $\sigma = 2.5$ oz; the agency can pretend an increase in boxes weight?

I.4 A company that produces fluorescent lamps wants to know if it can be asserted that their product lifespan has an average of $1,000$ hours. The statistic department choose a random sample of 100 lamps and measures a sample mean of 970 hours. It is known that the standard deviation is 85 hours. For 5% level of significance can we conclude that the average lifespan is less than $1,000$? Same question for 1%?

I.5 It is required that the average lifespan of a certain type of electrical battery to be 220 hours. It is known (from the fabrication process) that the lifespan follows a normal law with a standard deviation of 9 hours. For a random sample of 36 batteries the sample mean is found to be 218 hours. Can we conclude that the real average lifespan is different from 220 hours?

II. T-test - Inference for the mean of a population with unknown variance

We consider a statistical population whose variance (σ^2) is unknown. For a given random sample with sample mean $\overline{x}_n$, if the population is normally distributed or the sample is large enough ($n \geqslant 30$) the score $t = \dfrac{\overline{x}_n - \mu}{s/\sqrt{n}}$ is Student distributed with $n-1$ degrees of freedom: $t(n-1)$.

The T-test is performed like follows:

1. formulate the null hypothesis, which says that the mean of the population has a certain value:

$$\boxed{H_0 : \ \mu = \mu_0}$$

2. formulate the alternative hypothesis; we can have three different types of alternative hypothesis

$$\boxed{H_a:\ \mu < \mu_0} \quad \text{(left asymmetric) or}$$

$$\boxed{H_a:\ \mu > \mu_0} \quad \text{(right asymmetric) or}$$

$$\boxed{H_a:\ \mu \neq \mu_0} \quad \text{(symmetric hypothesis).}$$

3. choose a level of significance: α (usually 1% or 5%);

4. compute the z-score:

$$\boxed{t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}}}$$

5. determine the critical value t^*:

$$\boxed{t^* = qt(\alpha, n-1)} \text{ for left asymmetric } H_a\ (t^* < 0),$$

$$\boxed{t^* = qt(1-\alpha, n-1)} \text{ for right asymmetric } H_a\ (t^* > 0),$$

$$\boxed{t^* = -qt(\alpha/2, n-1) = qt(1-\alpha/2, n-1)} \text{ for symmetric } H_a\ (t^* > 0).$$

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

$$\boxed{t < t^*} \text{ for left asymmetric } H_a,$$

$$\boxed{t > t^*} \text{ for right asymmetric } H_a,$$

$$\boxed{|t| > |t^*|} \text{ for symmetric } H_a,$$

otherwise we say that **there is not sufficient evidence at the α level of significance** to reject the null hypothesis (**we fail to reject** H_0).

Solved Exercise. For an experiment concerning the sugar level of insects, 5 individuals are fed with sugar. The glucose values (which follow a normal distribution) are:

55.95 68.24 52.73 21.5 23.78

Test (5% significance level) the hypothesis that the level of sugar is greater than 40.

```
> alfa = 0.05
> x = c(55.95, 68.24, 52.73, 21.5, 23.78)
> population_mean = 40
> sample_mean = mean(x)
> n = 5
> s = sd(x)
> se = s/sqrt(n)
> critical_t = qt(1 - alfa, n - 1)
> t_score = (sample_mean - population_mean)/se
> critical_t
> t_score
```

The score is $t = 0.47867 < t^* = 2.13184$, hence the null hypothesis cannot be rejected.

Exercises to work

II.1 Write a function (called **t_test**) which will compute and return the critical value and the score of a T-test (this function will receive as parameters the type of the alternative hypothesis, the sample mean, the sample standard deviation etc). This function will be used to solve the following exercises.

II.2 For a random sample of a normally distributed population we measure

$$36\ 32\ 28\ 33\ 41\ 28\ 31\ 26\ 29\ 34$$

With 1%, test the hypothesis that the true mean of the population is different from 34.

II.3 On a certain pack of cigarettes the nicotine concentration (which follows a normal law) is 11.4 mg. An NGO claims that the real level of nicotine is much greater. For 100 randomly chosen packs the sample mean is found 11.9 mg with a sample standard deviation $s = 0.25$ mg. Test (with 1% and 5%) if the claims are legitimate.

II.4 The average of an history test grade (which follows a normal distribution) is 80 points. The history department wants to know if the current year students have an usual behavior. The results for a random sample can be found in the file *history.txt*). Perform an appropriate test (1% and 5% level of significance).

II.5 We consider a random sample of size 64 with sample mean 52 and variance $s^2 = 89.5$ (the population is normally distributed). Test the hypothesis that the true population mean is 49 versus the hypothesis that the population mean is different from 49.

III. χ^2 (χ-square) test as a test of goodness-of-fit

As a test of goodness-of-fit the χ^2 test aims to determine how well a theoretical distribution fits an empirical one (that obtained from the sample). Suppose that in a certain sample the individuals can be grouped in k categories $C_1, C_2, \ldots, C_k$ with observed absolute frequencies $o_1, o_2, \ldots, o_k$. But on the other hand it is expected that the theoretical absolute frequencies are $e_1, e_2, \ldots, e_k$. Let $m = \sum_{i=1}^{k} o_i$.

The test is performed like follows:

1. we first formulate the null hypothesis which says that the observed frequencies and the theoretical ones are equal:

$$\boxed{H_0 :\ (o_1, o_2, \ldots, o_k) = (e_1, e_2, \ldots, e_k)}$$

2. we formulate the alternative hypothesis which says that the observed frequencies and the theoretical ones are different:

$$\boxed{H_a :\ (o_1, o_2, \ldots, o_k) \neq (e_1, e_2, \ldots, e_k)}$$

3. we choose a level of significance: $\alpha \in \{1\%, 5\%\}$;

4. we compute the χ^2-score (the statistic of the test):

$$\boxed{\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}}$$

5. we determine the corresponding critical value:

$$\boxed{\chi^{2*} = qchisq(1-\alpha, k-1)}\quad H_a \text{ is right asymmetric.}$$

6. **the null hypothesis, H_0, is rejected** and we accept H_a if

$$\chi^2 > \chi^{2*}$$

otherwise **there is not sufficient evidence at the α level of significance to reject the null hypothesis**.

Solved exercise. Table from below shows the observed frequencies in tossing a die 120 times. Test the hypothesis that the die is not fair using a 5% level of significance.

face	1	2	3	4	5	6
occurrences	25	17	15	23	24	16

The score is $\chi^2 = 5$, and the critical value is $\chi^{2*} = qchisq(0.95, 5) = 11.07$; since $\chi^2 \not> \chi^{2*}$ the null hypothesis cannot be rejected.

Exercises to work.

III.1 Write a function (called **chisq_goodness_of_fit_test**) which will compute and return the critical value and the score of a χ^2-test (the parameters of the function will be: o, e, and α). This function will be used to solve the following exercises.

III.2 Gregor Mendel[1] observed 556 peas chosen at random: 315 were round and yellow, 108 were round and green, 101 were wrinkled and yellow, and 32 were wrinkled and green. According to his theory these numbers should be in proportion $9 : 3 : 3 : 1$. With 1% level of significance can we doubt its theory?

III.3 In 360 tosses of a pair of dice, 37 sums equal with five, 74 sums equal with sevens, and 24 sums equal with eleven are observed. Using the 5% significance level, test the hypothesis that the dice are not fair.

III.4 Two coins are tossed 240 times and the results are given below. With 1% level of significance can we say that the two coins are not fair?

results	{ H, H }	{ H, T }	{ T, T }
frequencies	55	115	70

IV. χ^2 test for statistical independence

The χ^2 test for statistical independence infers on the independence of two categorical variables. The observed values of these two variables are given in a so called contingency table.

	Y				
X	o_{11}	o_{12}	$\dots$	o_{1r}	$o_{1,}$
	o_{21}	o_{22}	$\dots$	o_{2r}	$o_{2,}$
	$\vdots$	$\vdots$	$\dots$	$\vdots$	
	o_{p1}	o_{p2}	$\dots$	o_{pr}	$o_{p,}$
	$o_{,1}$	$o_{,2}$	$\dots$	$o_{,r}$	m

[1] Austrian matematician and biologist, founder of genetics.

where o_{ij} is the number of observations belonging to the category i of X and j of Y, $e_{ij} = \dfrac{o_{i,} \cdot o_{,j}}{m}$ are the expected frequencies, and $o_{i,} = \sum_{j=1}^{r} o_{ij}$ and $o_{,j} = \sum_{i=1}^{p} o_{ij}$, $m = \sum_{i=1}^{p} \sum_{j=1]^{r} o_{i}j}$.

The test is performed like follows:

1. we first formulate the null hypothesis which says that the variables are independent:

$$\boxed{H_0 : \text{ the variables are independent}}$$

2. we formulate the alternative hypothesis which says that the the variables are not independent:

$$\boxed{H_a : \text{ the variables are dependent}}$$

3. we choose a level of significance: $\alpha \in \{1\%, 5\%\}$;

4. we compute the χ^2-score (the statistic of the test):

$$\boxed{\sum_{i=1}^{p} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

5. we determine the corresponding critical value:

$$\boxed{\chi^{2*} = qchisq(1-\alpha, (p-1)(r-1))} \quad H_a \text{ is right asymmetric.}$$

6. **the null hypothesis, H_0, is rejected** and we accept H_a if

$$\chi^2 > \chi^{2*}$$

otherwise **there is not sufficient evidence at the α level of significance to reject the null hypothesis**.

Solved exercise. We want to knwo if there exists a relation between the gender and the preferred music. Table from below summarizes the results for a sample of 110 males and 115 females. Test this hypothesis using a 5% level of significance.

	dance	rap	jazz	rock	country	classic	
male	20	25	15	10	25	15	110
female	25	15	20	10	20	25	115
	45	40	35	20	45	40	225

The score is $\chi^2 = 6.717$, and the critical value is $\chi^{2*} = qchisq(0.95, 5) = 11.07$; since $\chi^2 \not> \chi^{2*}$ the null hypothesis cannot be rejected.

Exercises to work.

IV.1 Write a function (called **chisq_independence_test**) which will compute and return the critical value and the score of a χ^2-test (the parameters of this function will be: *contingency_table*, and α). This function will be used to solve the following exercises.

IV.2 It is supposed that the there exists a relation between the income and the schooling degree. Table from below summarizes the results of a survey. Test the above hypothesis using a 1% level of significance.

	middle school	high school	college	master	Ph. D.	
low income	40	25	25	10	25	125
middle income	35	35	30	30	35	165
upper income	20	30	35	40	45	170
	95	90	90	80	105	460

IV.3 It is believed that the smartphone brands are related to the geographic area. Table from below contains the results of a random sample of 746 smartphone owners. Test the above hypothesis using a 5% level of significance.

	North A.	Europe	Asia	South A.	
Apple	106	74	36	28	244
Samsung	64	70	36	60	230
Google	26	12	6	4	48
Xiaomi	4	26	28	22	80
Oppo	2	6	70	10	88
Motorola	6	4	10	36	56
	208	192	186	160	746

IV.4 Is gender dependent of education level? A survey on a random sample of 400 people gave the results from below. Test the above hypothesis with 1% level of significance.

	high school	college	master	Ph. D.	
male	65	52	45	42	204
female	41	45	52	58	196
	106	97	97	100	400

V. Inference about the ratio of two variances - F-test

We consider two normal populations and we choose two simple random independent samples with sample variances s_1^2 and s_2^2. The statistic $F = \frac{s_1^2}{s_2^2}$ is Fisher distributed $F(n_1 - 1, n_2 - 1)$.

The test is performed like follows:

1. we first formulate the null hypothesis which says that the true variances are equal:

$$\boxed{H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1}$$

2. we formulate the alternative hypothesis; we can have two different types of alternative hypothesis

$$\boxed{H_a : \frac{\sigma_1}{\sigma_2} > 1} \quad \text{(right asymmetric) for an one-tailed test}$$

$$\boxed{H_a : \frac{\sigma_1}{\sigma_2} \neq 1} \quad \text{(symmetric hypothesis) for a two-tailed test.}$$

3. we choose a level of significance: $\alpha \in \{1\%, 5\%\}$;

4. we compute the F-score (the statistic of the test):

$$\boxed{F = \frac{s_1^2}{s_2^2}}$$

5. we determine the corresponding critical value(s):

$$\boxed{F^* = qf(1-\alpha, n_1 - 1, n_2 - 1)} \text{ for right asymmetric } H_a,$$

$$\boxed{F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1), F_d^* = qf(1-\alpha/2, n_1 - 1, n_2 - 1)}$$

for symmetric H_a.

6. **the null hypothesis, H_0, is rejected** and we accept H_a if

$$F > F^* \text{ for right asymmetric } H_a,$$

$$F < F_s^* \text{ or } F > F_d^* \text{ for symmetric } H_a.$$

otherwise **there is not sufficient evidence at the α level of significance to reject the null hypothesis**.

Solved exercise. The results to a psycholgical test peformed on two samples (women and men) are the following:

$$\text{men: } n_1 = 120, s_1 = 5.05$$

$$\text{women: } n_2 = 135, s_2 = 5.44$$

Can we conclude that the variances differs (1%)?

```
> alfa = 0.01
> n1= 120
> n2 = 135
> s1 = 5.05
> s2 = 5.44
> critical_F_s= qf(alfa/2, n1 - 1, n2 - 1)
> critical_F_d= qf(1 - alfa/2, n1 - 1, n2 - 1)
> critical_F_s
> critical_F_d
> F_score
```

The score is $F = 0.86175$, and the critical values are $F_s^* = 0.62843$, $F_d^* = 1.58257$; since $F \in [F_s^*, F_d^*]$ the null hypothesis cannot be rejected.

Exercises to work.

V.1 Write a function (called **F_test**) which will compute and return the critical value and the score of a F-test (the parameters of the function will be: the type of the alternative hypothesis, α, n_1, n_2, s_1, s_2 etc). This function will be used to solve the following exercises.

V.2 Researchers study the amplitude of mice movement under nervous stimulation. For drugged mice the result are:

12.512 12.869 19.098 15.350 13.297 15.589

For the other group the data are

11.074 9.686 12.164 8.351 12.182 11.489

The influence of the drugs is significant - concerning the variances? (5% level of significance)

V.3 A professor believes that a certain lecture program can improve the reading abilities of its students. He choses two groups: one formed with 22 students which follow the program (A) and one with 22 students which don't follow the lecture program (B). The results are given in the file *program.txt*. Decide (1% and 5% level of significance) if the variances of the two populations are different.