Probability and Statistics. Lab assignment 4: Hypothesis testing

General comments

Team ID number and members

The id number of team, which is referred to in tasks, is calculated as the sum of last digits in students' IDs of all team members. In our case it's $2 + 6 = 8$.

Our team consists of:

- **Andrii Yaroshevych**
- **Pavlo Kryven**

Data generation

The data for problems 1-3 are generated as follows: set

$$a_k := \left\{ k \ln\left(k^2 n + \pi\right) \right\}, \quad k \geq 1,$$

where $\{x\} := x - [x]$ is the fractional part of a number x and n is your id number. Sample realizations $X_1, \ldots, X_{100}$ and $Y_1, \ldots, Y_{50}$ from the hypothetical normal distributions $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$ respectively are obtained as

$$\begin{aligned} x_k &= \Phi^{-1}\left(a_k\right), && k = 1, \ldots, 100, \\ y_l &= \Phi^{-1}\left(a_{l+100}\right), && l = 1, \ldots, 50, \end{aligned}$$

where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$ and Φ^{-1} is its inverse.

Instructions and problems

Instructions: In problems $1-3$, test H_0 vs H_1. To this end,

- point out what standard test you use and why;
- indicate the general form of the rejection region of the test H_0 vs H_1 of level 0.05;
- find out if H_0 should be rejected on the significance level 0.05;
- indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why.

Problem 1. $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$; $\quad \sigma_1^2 = \sigma_2^2 = 1$

Problem 2. $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 > \sigma_2^2$; $\quad \mu_1$ and μ_2 are unknown.

Problem 3. Using Kolmogorov-Smirnov test in $\mathbf{R}$, check if

a. $\{x_k\}_{k=1}^{100}$ are normally distributed (with parameters calculated from the sample);

b. $\{|x_k|\}_{k=1}^{100}$ are exponentially distributed with $\lambda = 1$;

c. $\{x_k\}_{k=1}^{100}$ and $\{y_l\}_{l=1}^{50}$ have the same distributions.

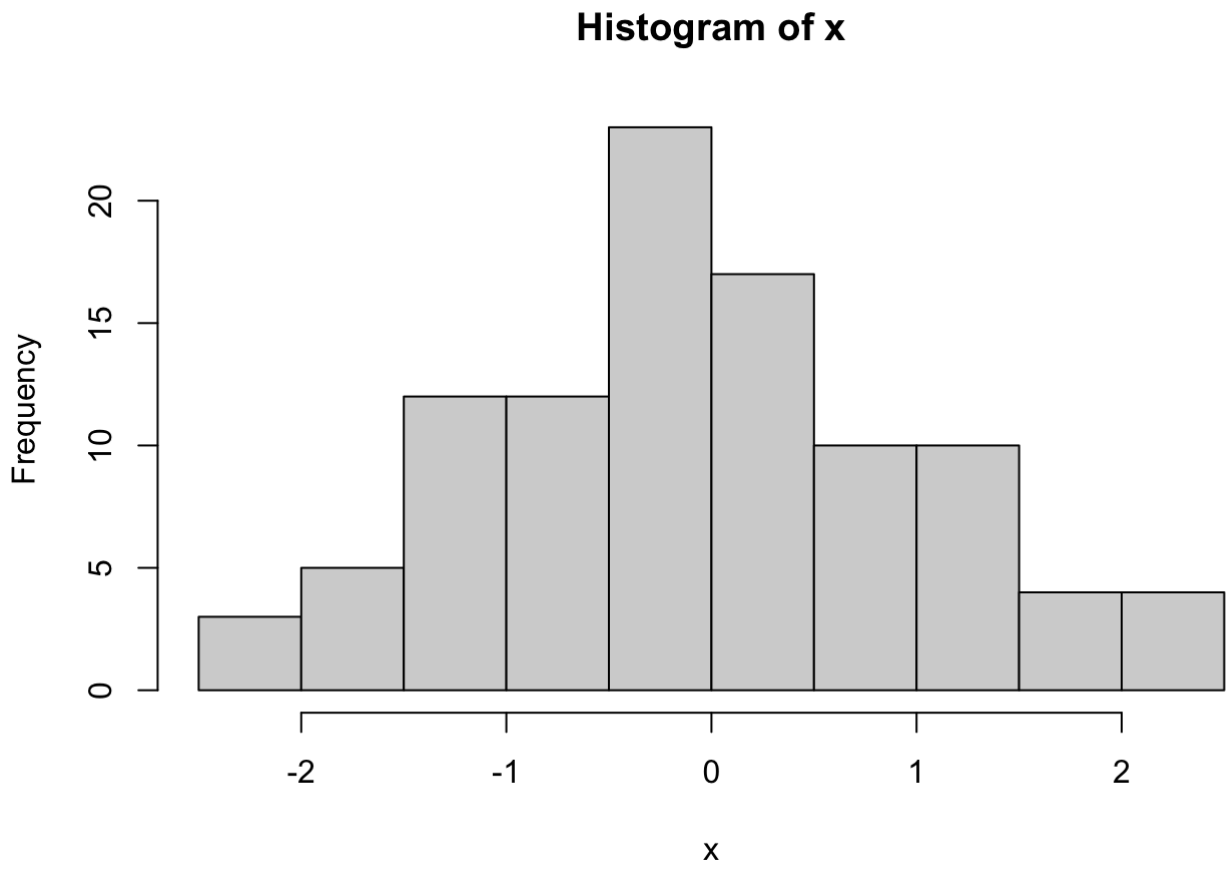
Explain the main idea behind the KS test and comment on the outcomes of the test.

Necessary libraries

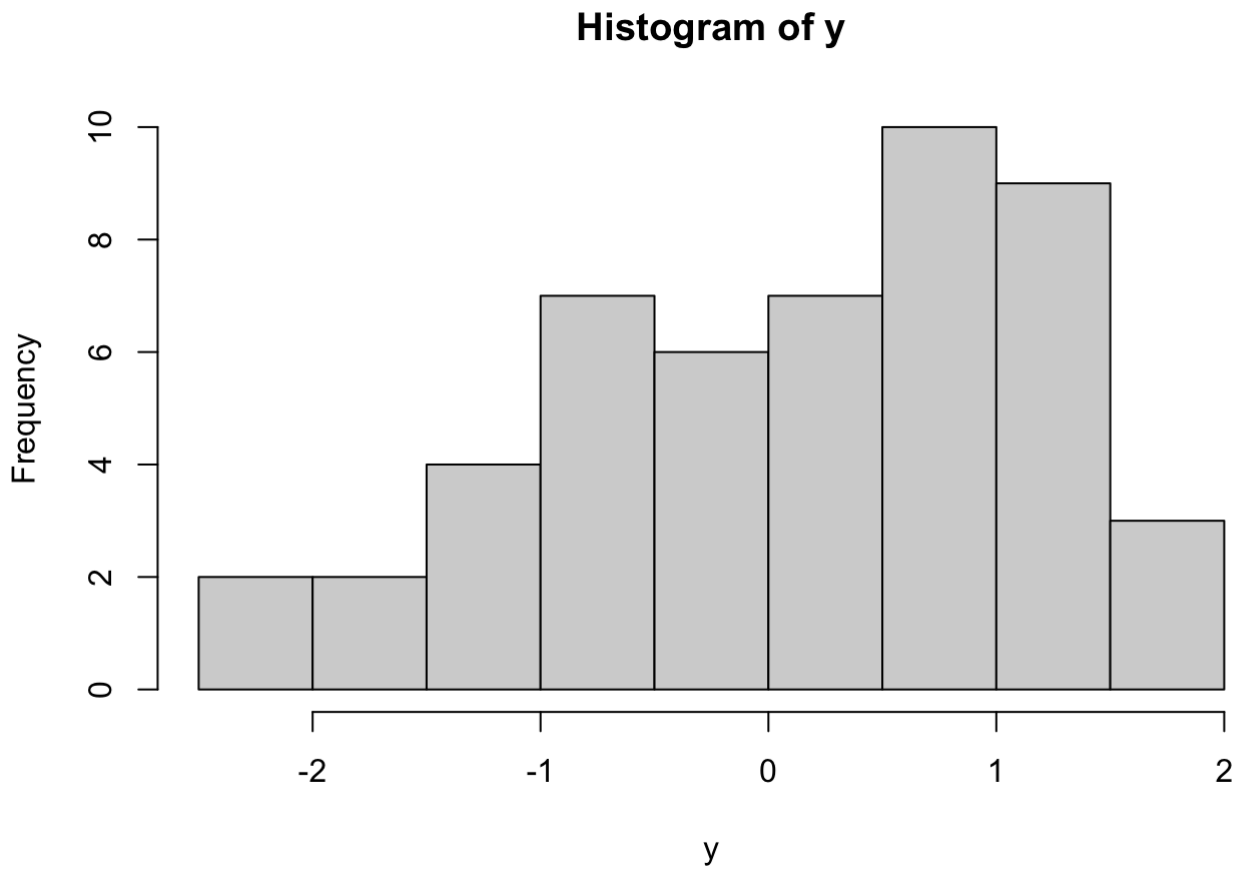
```
library(BSDA)
```

Generating data

```
id <- 8
n <- id

set.seed(8)

k <- 1:150
a.data <- (k*log(k^2*n + pi))%%1
x <- qnorm(a.data[1:100])
y <- qnorm(a.data[101:150])

hist(x)
```




```
hist(y)
```



Problem 1.

$H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$; $\sigma_1^2 = \sigma_2^2 = 1$

Done by ***Andrii Yaroshevych***

Point out what standard test you use and why

In this problem, we want to compare two samples from normal distributions with unknown means and known variances. We use the z-test for two independent samples:

```
z.test(x, y, alternative = "two.sided", mu = 0, sigma.x = 1, sigma.y = 1)
```

```
##
##  Two-sample z-Test
##
## data:  x and y
## z = -1.1217, p-value = 0.262
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5337676  0.1451839
## sample estimates:
##   mean of x   mean of y
## -0.07774598  0.11654586
```

Indicate the general form of the rejection region of the test H_0 vs H_1

The general form of the rejection region of the test H_0 vs H_1 is:

$$C_\alpha = \{\mathbf{x} \in \mathbf{R}^n \mid z(\mathbf{x}) \geq z_{1-\alpha}\}$$

where $z(\mathbf{x})$ is the z-statistic and $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

In our case, $\alpha = 0.05$ and $z_{1-\alpha} = 1.96$. The z-statistic of standard normal distribution is calculated as:

$$z = \frac{\overline{x} - \overline{y}}{\sigma}\sqrt{\frac{mn}{(m+n)}}$$

Now we can show the rejection region of the test H_0 vs H_1 for our problem:

$$C_{0.05} = \{\mathbf{x} \in \mathbf{R}^n,\ \mathbf{y} \in \mathbf{R}^n \mid\ |z(\mathbf{x}, \mathbf{y})| \geq 1.96\}$$

Find out if H_0 should be rejected on the significance level 0.05

H_0 should not be rejected on the significance level 0.05.

Indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why

The p-value of the test is 0.262. We would not reject H_0 for that value of p because $p > \alpha = 0.05$. We can reject H_0 at significance level $p(\mathbf{x})$ but not on smaller significance levels.

Problem 2.

> $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 > \sigma_2^2$; μ_1 and μ_2 are unknown.
>
> *Done by **Andrii Yaroshevych***

Point out what standard test you use and why

In this problem, we want to compare two samples from normal distributions with unknown means and hypothesis about the variances. So, we can use the f-test for two independent samples:

```
var.test(x, y, alternative = "greater")
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 1.0378, num df = 99, denom df = 49, p-value = 0.4516
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.6778692       Inf
## sample estimates:
## ratio of variances
##           1.037817
```

Indicate the general form of the rejection region of the test H_0 vs H_1

The general form of the rejection region of the test H_0 vs H_1 is:

$$C_\alpha = \{\mathbf{x} \in \mathbf{R}^n, \mathbf{y} \in \mathbf{R}^m \mid f(\mathbf{x}, \mathbf{y}) \leq f_\alpha\}$$

where f_α is the quantile of level α for the Fischer distribution with $n-1$ and $m-1$ degrees of freedom $\mathcal{F}_{n-1,m-1}$

In order to derive the test statistics, we will use the following estimators:

$S_{xx}/(n-1)$ is the estimate for σ_1^2

$S_{yy}/(m-1)$ is the estimate for σ_2^2

So, under the null hypothesis, the test statistics is:

$$F(\mathbf{X},\mathbf{Y}) = \frac{S_{XX}/(n-1)}{S_{YY}/(m-1)}$$

In our case, the rejection region of the test H_0 vs H_1 is:

$$C_{0.05} = \{\mathbf{x} \in \mathbf{R}^n, \mathbf{y} \in \mathbf{R}^m \mid f(\mathbf{x},\mathbf{y}) \leq f_{0.05}\}$$

Find out if H_0 should be rejected on the significance level 0.05

H_0 should not be rejected on the significance level of $\alpha = 0.05$.

Indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why

The p-value of the test is 0.4516. We would not reject H_0 for that value of p because $p > \alpha = 0.05$.

Problem 3.

> *Done by **Pavlo Kryven***

Point out what standard test you use and why

In this problem, we want to test the distribution of the sample $\mathbf{x}$ for normality.

$$H_0 \ : \ F_x = F_y \qquad \text{vs} \qquad H_1 \ : \ F_x \neq F_y$$

We use the Kolmogorov-Smirnov test for two independent samples.

Indicate the general form of the rejection region of the test H_0 vs H_1

The general form of the rejection region of the test H_0 vs H_1 is:

$$C_\alpha = \{\mathbf{x} \in \mathbf{R}^n \mid d \geq d^{(n)}_{1-\alpha}\}$$

Here we use the following test statistic:

$$D(X,Y) := \sup_{t \in \mathbb{R}} |\hat{F_x}(t) - F_0(t)|$$

under H_0, the distribution of d is independent of F_0 and is called the Kolmogorov distribution $\mathcal{D}_n$

So, for our problem, the rejection region is:

$$C_{0.05} = \{\mathbf{x} \in \mathbf{R}^n \mid d \geq d^{(n)}_{0.95}\}$$

a) $H_0 \ : \ \{x_k\}_{k=1}^{100} \sim \mathcal{N}(\mu, \sigma^2)$ vs $H_1 \ : \ \{x_k\}_{k=1}^{100} \nsim \mathcal{N}(\mu, \sigma^2)$

```
mean <- mean(x)
sd <- sd(x)

ks.test(x, "pnorm", mean = mean, sd = sd)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.057707, p-value = 0.8931
## alternative hypothesis: two-sided
```

Find out if H_0 should be rejected on the significance level 0.05

H_0 should not be rejected at the significance level of $\alpha = 0.05$.

Indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why

The p-value of our test is 0.8931. We would not reject H_0 for that value of p because $p > \alpha = 0.05$.

b) $H_0 : \{|x_k|\}_{k=1}^{100} \sim \mathcal{E}(1)$ vs $H_1 : \{|x_k|\}_{k=1}^{100} \nsim \mathcal{E}(1)$

```
lambda <- 1

ks.test(abs(x), "pexp", lambda)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  abs(x)
## D = 0.097408, p-value = 0.2989
## alternative hypothesis: two-sided
```

Find out if H_0 should be rejected on the significance level 0.05

H_0 should not be rejected at the significance level of $\alpha = 0.05$.

Indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why

The p-value of our test is 0.2989. We would not reject H_0 for that value of p because $p > \alpha = 0.05$.

c) $H_0 : \{x_k\}_{k=1}^{100} \sim \{y_l\}_{l=1}^{50}$ vs $H_1 : \{x_k\}_{k=1}^{100} \nsim \{y_l\}_{l=1}^{50}$

```
ks.test(x, y)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  x and y
## D = 0.19, p-value = 0.1732
## alternative hypothesis: two-sided
```

Find out if H_0 should be rejected on the significance level 0.05

H_0 should not be rejected at the significance level of $\alpha = 0.05$.

Indicate the p-value of the test and comment whether you would reject H_0 for that value of p and why

The p-value of our test is 0.1732. We would not reject H_0 for that value of p because $p > \alpha = 0.05$.

Conclusion

After conducting the experiments, it is clear that the results were as expected.

The Z-test and F-test are statistical tests used to compare the means of two samples or to compare the variances of two samples, respectively. The Z-test is used to determine if the difference between the means of two samples is statistically significant, while the F-test is used to determine if the ratio of the variances of two samples is statistically significant. Z-test requires knowing the population standard deviation. These tests involve comparing a calculated test statistic to a critical value taken from a standard distribution, such as the normal or t-distribution, to determine the significance of the difference or ratio. If the calculated test statistic exceeds the critical value, the null hypothesis that the means or variances are equal is rejected. The Z-test and F-test are useful tools for comparing samples and drawing conclusions about the underlying populations based on the sample data.

In terms of the first and second problem, it is unsurprising that the mean and variance of the two samples were similar, as they were generated in the same manner.

The Kolmogorov-Smirnov (K-S) test is a statistical test used to compare the distribution of a sample with a reference distribution or to compare two samples to determine if they come from the same distribution. The main idea behind the K-S test is to calculate the maximum difference, or distance, between the cumulative distribution functions (CDFs) of the sample and the reference distribution or between the two samples being compared.

The CDF is a function that describes the probability of a random variable being less than or equal to a certain value. The K-S test determines the significance of this maximum distance by comparing it to a critical value calculated from the sample size and the level of significance chosen for the test. If the maximum distance is greater than the critical value, the null hypothesis that the sample comes from the same distribution as the reference distribution or that the two samples being compared come from the same distribution is rejected.

It can be inferred that the distribution produced by module of the sample in task 3b, which inverts values around the origin, will exhibit a half-normal distribution due to the fact that the mean of a normal distribution is zero. This is confirmed by the results of the tests.

In addition, the final test did not reject the hypothesis that the two samples have the same distribution, which is also expected given that both samples were generated in the same way and follow a normal distribution.

Overall, the third problem was particularly interesting as it highlights the common practice of testing samples for the distributions they follow, which can provide valuable insights into the data.

The results of the K-S test were consistent with expectations, indicating that the distribution of the sample aligns with the reference distribution or that the two samples being compared come from the same underlying distribution. This suggests that the sample accurately reflects the characteristics of the population or that the two samples are representative of the same population. The successful application of the K-S test adds reliability to the conclusions drawn from the sample data.

In summary, the experiments conducted yielded results that were consistent.