

BI-PST/24-25 Homework

Oleksandr Slyvka
slyvkole@cvut.cz

Illia Lyhin
lyhinill@cvut.cz

Maksym Khavil
khavimak@cvut.cz

November 30, 2024

Abstract

With this document we present homework for BI-PST. Oleksandr Slyvka was chosen as a representative for our group. He was born 14.04.2006, so we get $K = 14$ and $L = 6$, then $M = 4$. Such value of M corresponds to case0202 Sleuth dataset, volume of hippocampus with respect to schizofrenia. We will analyse this dataset using numpy, pandas, matplotlib and scipy.stats Python modules.

1 Data and estimating moments

Firstly, we will load data and display it as a table. There are $n = 15$ samples in both categories, values seem to be greater than 1cm^3 and peak around 2cm^3 , each row represent a pair of twins, one of whom was affected by schizofrenia.

"Unaffected"	"Affected"
1.94000005722046	1.26999998092651
1.44000005722046	1.62999999523163
1.55999994277954	1.47000002861023
1.58000004291534	1.38999998569489
2.05999994277954	1.92999994754791
1.6599999666214	1.25999999046326
1.75	1.71000003814697
1.76999998092651	1.66999995708466
1.77999997138977	1.27999997138977
1.91999995708466	1.85000002384186
1.25	1.01999998092651
1.92999994754791	1.3400000333786
2.03999996185303	2.01999998092651
1.62000000476837	1.5900000333786
2.07999992370605	1.97000002861023

1.1 Estimations

Let X_{unaff} and X_{aff} denote random variables of hippocampues volume of those who are unaffected and affected by schizofrenia respectively. Then we will compute estimates of mean, median and variance of those random variables. Median estimation will be chosen as 7th value of sorted sequence of data points. Estimated mean and variance will be computed with following formulas:

$$\widehat{\mathbb{E}X} = m_1 = \bar{X} = \frac{1}{n} \sum_{k=1}^n X_i$$

$$\widehat{\text{var}X} = s^2 = \frac{1}{n-1} \sum_{k=1}^n (X_i - \bar{X})^2$$

Additionally we will compute uncorrected variance.

$$S^2 = \frac{1}{n} \sum_{k=1}^n (X_i - \bar{X})^2$$

All those computations are incapsulated in function estimate.

```
def estimate(X):
    print(f'{'Mean:' :<30}{X.mean():.3 f}')
```

```
    print(f'{'Median:' :<30}{X.median():.3 f}')
```

```
    print(f'{'Variance without correction:' :<30}{((X - X.mean()) ** 2).sum() / n :.3 f}')
```

```
    print(f'{'Variance with correction:' :<30}{((X - X.mean()) ** 2).sum() / (n - 1):.3 f}')
```

After plugging values in we obtain such results.

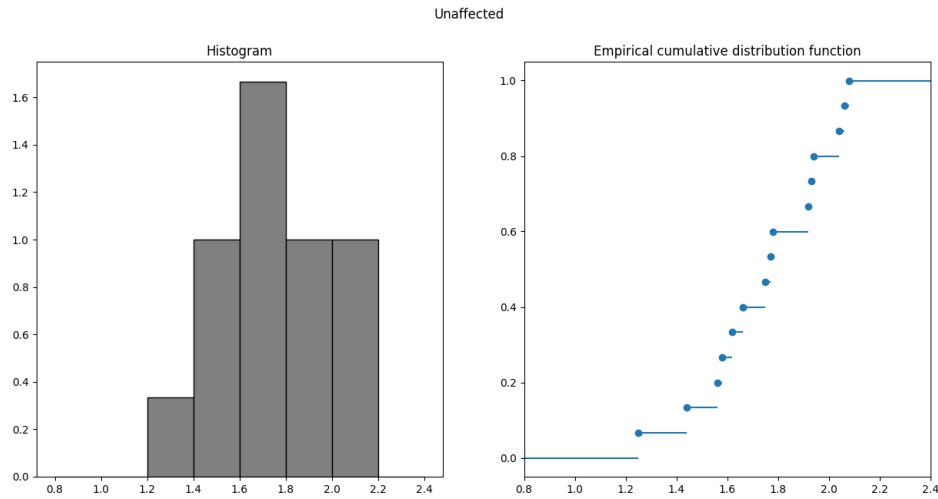
$\widehat{\mathbb{E}X_{\text{unaff}}} = 1.759$	$\widehat{\mathbb{E}X_{\text{aff}}} = 1.560$
$\widehat{F_{X_{\text{unaff}}}^{-1}}(0.5) = 1.770$	$\widehat{F_{X_{\text{aff}}}^{-1}}(0.5) = 1.590$
$\widehat{\text{var}X_{\text{unaff}}} = 0.059$	$\widehat{\text{var}X_{\text{aff}}} = 0.091$
$S_{\text{unaff}}^2 = 0.055$	$S_{\text{aff}}^2 = 0.085$

Let's notice that estimated expected value of unaffected distribution is greater than estimated expected value of affected distribution.

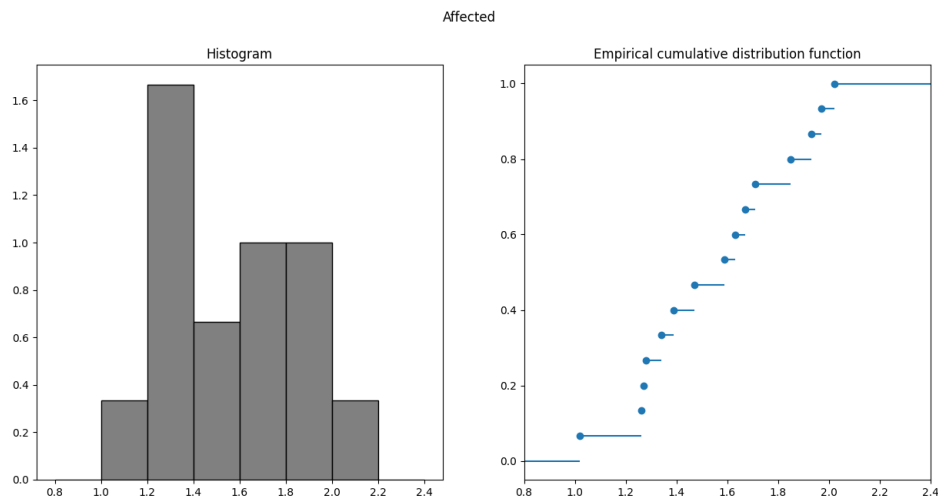
2 Histograms and empirical cumulative distribution functions

We will plot histograms, bins' width will be 0.2cm^3 . Near them we will plot empirical cumulative distribution functions, that will have a step of $1/n$ at each sample value.

Let's plot histogram and ecdf for distribution of unaffected twins.



Here is the same plot for distribution of schizophrenic twins.



We can see that later histogram is a little wider, that corresponds with its variance being greater than variance of unaffected distributions. Another observation is that it is altogether slightly shifted towards zero.

3 Choosing between normal, uniform and exponential distributions

We will try to model our data as one of the distributions listed above. To do that we need to find parameters of distributions, we will calculate them using method of moments.

3.1 Calculating parameters

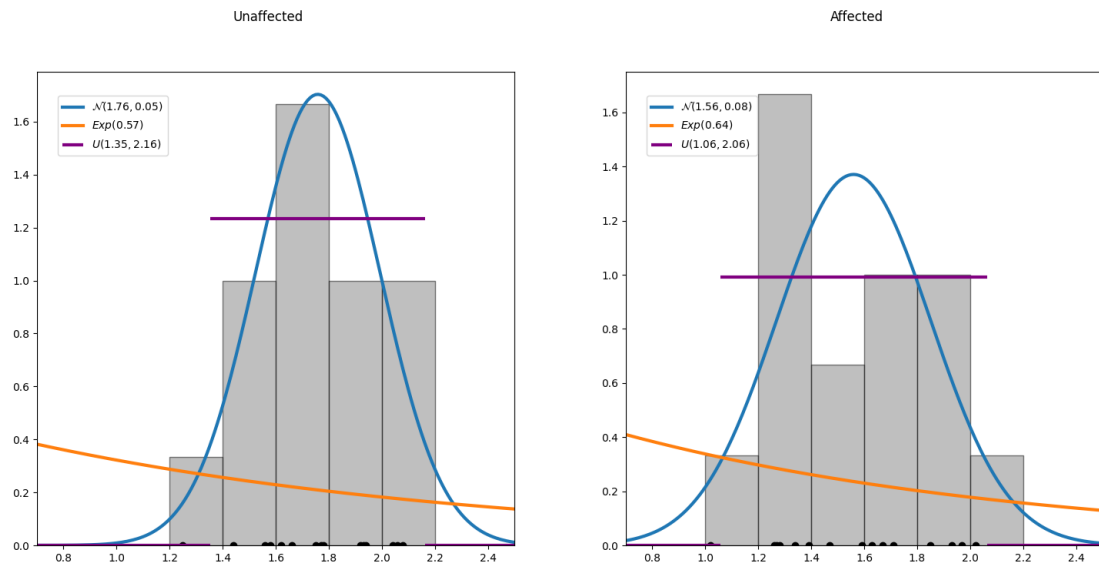
Normal distribution is parametrized by μ and σ^2 , uniform distribution is parametrized by its lower and upper bounds resp. a and b and exponential one is parametrized by the rate λ . All of them can be expressed as functions of at least two moments (exponential can be derived using single moment).

$$\begin{aligned}\hat{\mu} &= m_1 \\ \hat{\sigma}^2 &= m_2 - m_1^2 \\ \hat{a} &= m_1 - \sqrt{3(m_2 - m_1^2)} \\ \hat{b} &= m_1 + \sqrt{3(m_2 - m_1^2)} \\ \hat{\lambda} &= \frac{1}{m_1}\end{aligned}$$

Those computations are calculated in this code:

```
m1 = X.mean()
m2 = (X ** 2).mean()
mu_hat= m1
sigma2_hat= m2 - m1 ** 2
lambda_hat = 1 / m1
a_hat = m1 - np.sqrt(3 * (m2 - m1**2))
b_hat = m1 + np.sqrt(3 * (m2 - m1**2))
print(f"Normal: {mu_hat=}, {sigma2_hat=}")
print(f"Exp: {lambda_hat=}")
print(f"Uniform: {a_hat=}, {b_hat=}")
```

3.2 Comparing histogram and probability distribution functions



In both cases exponential distribution performs horribly, so we will reject that immediately. At both plots there are data points, that lay in zero-probability area for uniform distribution, so it will be rejected too. We are left only with normal distribution, so we naturally choose it, additionally it matches histogram peaks quite nicely.

4 Comparing histograms of 100 generated samples and data

lorem

5 Computing confidence intervals at 95% level

lorem

6 Testing if means equal K at 5% level

lorem

7 Testing if means of distributions of volumes affected and unaffected people equal at 5% level

lorem

References

- [1] BI-PST 24/25, lectures