HW №2. Hypothesis testing and ANOVA

Sergey Titov, Elena Kartysheva, Konstantin Grotov

July 31, 2023

1 Comparison of Distributions (6 points)

1.1 Theoretical Distributions (4 points)

From theory we know several facts:

1. Let $X_1, \ldots, X_n$ be a random sample from a $n(\mu, \sigma^2)$ distribution. Then $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student's t-distribution with $n-1$ degrees of freedom.
2. Let $X_1, \ldots, X_n$ be a random sample from a $n(\mu, \sigma^2)$ distribution. Then $(n-1)S^2/\sigma^2$ has a χ^2_{n-1} distribution.
3. Let $X_1, \ldots, X_n$ be a random sample from a $n(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_n$ be a random sample from a $n(\mu_Y, \sigma_Y^2)$ population. The random variable $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ has Snedecor's F distribution with $n-1$ and $m-1$ degrees of freedom.

Generate samples using these theoretical facts (using direct or non-direct method) and compare them with samples generated from the corresponding distributions. Compare means using parametric (t-test) and non-parametric (Mann–Whitney U) tests. **Assuming as Null hypothesis, that two samples have the same mean value get p-value and conclude, should we accept or reject the null hypothesis?** It's forbidden to use any built-in distribution functions (such as scipy.stats.f, scipy.stats.chi2, build-in inverse cdf's etc.), but it's allowed to use any special functions (such as scipy.special.gamma, scipy.special.gamma, scipy.special.hyp2f1 etc.). It's forbidden to use statistical tests from third-party libraries, but it's not forbidden to compare with them. Try for different sample sizes: 20, 100, 2000. Use $\alpha = 0.05$.

1.2 Given Distributions (2 points)

Use your test realisation and compare means for the given dataset. Explore the data, assume null hypothesis, select the correct method and make statistical inference.

2 ANOVA (4 points)

Implement two-way ANOVA test [1]. Check the ANOVA assumptions. Test the following hypothesis on the provided data: The "Y1" and "Y2" variables affect the "X" variable. Make a statistical inference and compare your p-values with values which give any statistical library (for example, "pingouin" in Python).

References

[1] Some information about Two-way ANOVA https://people.richland.edu/james/lecture/m170/ch13-2wy.html