TASK LIST NO. 10: Advanced Statistical Methods

Task 1

The execution times of a certain task were measured. The results were ordered in the sequence they were received (in time). Then, for each result, it was determined whether it is above (a) or below (b) the median.

A sequence of symbols was obtained: $a, a, b, b, a, a, a, b, b, b, b, a, a, b, ...$

Verify the hypothesis that the sample is random (i.e., the results do not depend on time/order), using the **runs test**.

Task 2

We have two sorting algorithms (A and B). 5 independent time measurements were performed for each of them. The results do not follow a normal distribution (outliers are present).

- Algorithm A: $12, 18, 14, 15, 13$
- Algorithm B: $19, 21, 23, 20, 22$

Verify the hypothesis that algorithm A is faster than B using the rank-sum test (Mann-Whitney-Wilcoxon).

Task 3

Data on failures depending on the hardware manufacturer were collected in a table (contingency table):

Manufacturer Failure Type	Overheating	Disk Error	Memory Error
Manufacturer X	20	10	15
Manufacturer Y	30	50	25

Check at the significance level $\alpha = 0.05$ whether the type of failure depends on the manufacturer.

Task 4

We are testing the performance of 3 different frameworks (X, Y, Z). Since the data are strongly asymmetric, instead of classical analysis of variance (ANOVA), we use the non-parametric Kruskal-Wallis test.

For the ranking data from the table, verify the hypothesis that all frameworks have the same median performance.

Task 5

We have two datasets on network traffic (before and after firewall implementation). We want to check if the **entire distribution** (not just the mean) has changed.

Based on the empirical distribution functions of both samples, calculate the $D_{n,m}$ statistic and verify the hypothesis of identical distributions (Kolmogorov-Smirnov test).

Task 6

We investigate code compilation time (Y) depending on the number of files (X_1) and the number of lines of code in a file (X_2).

For the given data, determine the equation of the regression plane:

$$y = ax_1 + bx_2 + c$$

Task 7

The number of transistors in processors grows exponentially: $y = a \cdot e^{bx}$. Having historical data, reduce this problem to linear regression by taking logarithms ($\ln y = \ln a + bx$) and determine the growth parameters.

Task 8

Processor production generates a certain percentage of defects. Instead of taking a fixed sample of 100 units, we take units one by one. After each extraction, we decide: “good batch”, “bad batch”, or “continue sampling”.

Construct a sequential test (Wald test) to verify the hypothesis $p = 0.01$ against $p = 0.10$.

Task 9

For a simple sample $x_1, ..., x_n$ from an exponential distribution (failure-free operation time) with density $f(x) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right)$, where λ is the expected value, determine the estimator of the parameter λ using the maximum likelihood method (MLE).

Task 10

We have 3 servers. We want to check if they operate equally stably (if they have the same variance of response times) before comparing their average times. The sample variances are: $s_1^2 = 1.4,\ s_2^2 = 1.8,\ s_3^2 = 1.2$.

Verify the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ (e.g., using Bartlett’s test).