University of Wrocław: Data Science

Theoretical Foundations of Large Data Sets, List 1

1. Let $X_1, \ldots, X_n$ be the simple random sample from the distribution with the density $f(x, \alpha) = (\alpha + 1)x^\alpha$, for $x \in (0, 1)$, $\alpha > -1$.

   a) Find the maximum likelihood estimator $\hat{\alpha}_{MLE}$ of the parameter α.

   b) Find the Fisher Information and the asymptotic distribution of this estimator. Use these results to estimate the mean squared error of $\hat{\alpha}_{MLE}$.

   c) Calculate the moment estimator $\hat{\alpha}_{mom}$ of the parameter α.

   d) Fix $\alpha = 5$ and generate one random sample of the size $n = 20$. Calculate both estimators and the respective values of $\alpha - \hat{\alpha}$ and $(\alpha - \hat{\alpha})^2$. Which estimator is more accurate?

   e) Generate 1000 samples of the size $n = 20$ and

      i) draw histograms, box-plots and q-q plots for both estimators;

      ii) estimate the bias, the variance and the mean-squared error of both estimators and construct approximate 95% confidence intervals for these parameters. In case of MLE compare the values of these parameters to the values provided by the asymptotic distribution of $\hat{\alpha}_{MLE}$.

      Which estimator is more accurate?

   f) Repeat point e) for $n = 200$. Compare the results with those for $n = 20$.

2. Let $X_1, \ldots, X_n$ be the simple random sample from the distribution with the density $f(x, \lambda) = \lambda e^{-\lambda x}$, for $x > 0$, $\lambda > 0$.

   Find the uniformly most powerful test at the level $\alpha = 0.05$ for testing the hypothesis $H_0 : \lambda = 5$ against $H_1 : \lambda = 3$.

   a) Provide the formula for the critical value for this test.

   b) Provide the formula for the power of this test.

   c) Provide the formula for the p-value for a given random sample. For $n = 20$ generate one random sample from H_0 and one random sample from H_1 and find the respective p-values. What conclusions can be drawn based on the p-values?

   d) What is the distribution of the p-value when data come from H_0?

   e) Generate 1000 samples of the size $n = 20$ from H_0 and calculate respective p-values.

      i) Compare the distribution of these p-values to the distribution derived in d): draw a histogram and a respective q-q plot.

ii) Use these simulations to construct the 95% confidence interval for the type I error of the test.

f) Generate 1000 samples of the size $n = 20$ from H_1 and calculate respective p-values.

i) Compare the distribution of p-values under H_0 and under H_1.

ii) Use these simulations to construct the 95% confidence interval for the power of the test. Compare with the theoretically calculated power.

g) Repeat points e)-f) for $n = 200$. Critically compare these results with the results for $n = 20$.