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 $x \in (0,1)$, $\alpha > 0$.

   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. How would you estimate the mean squared error of $\hat{\alpha}_{MLE}$?

   c) Calculate $E(X_1)$ and use it to derive the moment estimator $\hat{\alpha}_M$ of α.

   d) Fix $\alpha = 5$ and generate one random sample with $n = 20$.

   e) Calculate both estimators and the respective values of $\hat{\alpha} - \alpha$ and $(\hat{\alpha} - \alpha)^2$. Which estimator is more accurate ?

   f) Repeat point d) 1000 times (i.e. generate 1000 samples) 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 $\hat{\alpha}_{MLE}$.

   g) Repeat points d)-f) 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 exponential distribution with the density $f(x;\theta) = \lambda e^{-\lambda x}$, 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 $H_1 : \lambda = 3$ and

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

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

   c) For $n = 20$ generate one random sample from H_0 and find the respective p-value.

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

   e) Repeat point c) 1000 times (i.e. generate 1000 samples) and calculate respective p-values.

      i) Compare the distribution of these p-values to the distribution derived in point 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 your test.

   f) Generate 1000 samples of size $n = 20$ from the alternative distribution and calculate respectve p-values.

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

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

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