Technische Universität München
Lehrstuhl Informatik VIII
Prof. Dr.-Ing. Georg Carle
Dr. Heiko Niedermayer
Cornelius Diekmann, M.Sc.

Netzwerkanalyse Sommersemester 2014
Assignment 4

Task 1 Experimental Design

You experience a new network phenomenon. You think about the problem and identify three potential factors speed, breadth, and locality influencing it. When you increase speed from s_0 to s_{large}, the phenomenon becomes stronger. When you increase breadth from b_0 to b_{large}, it also becomes stronger, though less intense. Changing locality from l_0 to other values did not seem to change the phenomenon.

a) What will happen when the factors are set to s_{large}, b_{large}, and l_0?

Solution: Unknown as we have not seen the factor combination in our experiment. One cannot conlude safely that it will also increase or that both increases for the individual factors will even add up to a larger increase.

b) Can we conclude that locality has little or no influence on the phenomenon?

Solution: No, as this might change when the other factors speed and breadth take different values, in particular if they are of different magnitude as in the experiments.

Task 2 Scipy.stats, from Fitting to KS-Test

a) Generate three data sets with 100 random numbers (using scipy.stats.XXX.rvs(size=100, ...)):

- Normal Distribution with mean=4.0 and standard deviation=0.5
- Normal Distribution with mean=4.1 and standard deviation=0.5
- Triangle Distribution with mean=4.0 generating values from 0 to 8.0.

Fit each data set with the normal distribution (scipy.stats.XXX.fit(...)). The result of a fit is a tuple of (shape, loc, scale) or (loc, scale) in case of the normal where loc is the mean in case of the normal distribution and scale the standard deviation. What values do you see?

b) Fit the triangle data set to a lognormal distribution. Generate another data set with 100 random numbers from the fitted values with the lognormal distribution. Now, sort all four data sets (e.g. using dataset.sort()) and plot them with the data sets on the x-axis and the index as y-axis to generate a kind of CDF of the data sets.

c) Use a Kolmogorov-Smirnov test (scipy.stats.kstest(data, scipy.stats.norm.cdf(...))) to test your data sets against a normal distribution with mean 4.0 and standard deviation 0.5. The result tuple contains the p-value as second value (low value close to 0.0 indicating low chance that the distribution could have generated the data set). Which data set is accepted?

Solution: See code: aufg_4.py
The KS test will reject the data set of the triangle and lognormal distributions. It will accept both normal distributions, the one with slightly different mean with lower p value, yet still good enough. It would reject if for larger sample size like 1000 random samples in the data set.

Task 3 CHI-Square Test

The CHI-Square test needs as input two histograms and it then compares the two histograms. The data sets needs to be of the same size and the bins of both histogram have to be for identical ranges and none should be empty, best at least five elements in each bin.

a) Take three data sets from the previous task, split the range into bins and count the number of elements from each data set that fall into each bin. The result should be two array f_{obs} that you can put into the CHI square test.

b) Apply scipy.stats.chisquare on the two f_{obs} arrays. Interpret the p-value again for similar data sets and for data sets with different statistics.

c) For the CHI square test you need to determine the number of degrees of freedom of the data. Usually this is the number of bins minus 1 (all entries have to sum up to the size of the input data). Other constraints on the data can reduce this value. Now use the first result value *chisq* from the chisquare test to compute: 1.0-scipy.stats.chi2.cdf(chisq, numberOfBins - 1). What value do you get?

Solution: See code: aufg_4.py

3b) One observation is that selecting the bins is crucial and can cause problems. For alpha=0.05 (5%), we sometimes get the result of them being equal (p>alpha) or not equal (p<alpha). For larger size of the data set, the test will most certainly reject the equality for the two data sets.

3c) We get the p-value with the calculation.