Exercises

E1. Consider the behavior of sample mean $(1/n)\sum_{i=1}^{n} x_i$ for n random variables that follow independently the standard Cauchy-distribution. Note that m random standard Cauchy samples, each of size n, can be easily generated in Matlab using the command cauchy_sample=randn(n,m)./randn(n,m). Investigate how the variance of the sample mean changes as n increases (consider the values 10, 50, 100, 500, 1000). The value $m = 1000$ should be sufficient to characterize the distribution of the sample mean. Draw histograms of the sample mean values to see what the underlying distribution looks like.

E2. Compare the findings in the exercise E1 to the situation where the data $x_1, ..., x_n$ follow the standard Normal distribution instead. Standard Normal samples are generated by the Matlab command randn(n, m).

E3. Here we consider the behavior of maximum likelihood (ML) estimates when the assumed model is not correct. Simulate 1000 values from a Poisson process with a heterogeneous rate function according to the following scenario. Firstly, we note that a Poisson process with a constant rate can be simulated by defining the event times successively as $t_{n+1} = t_n - \lambda \log(\text{rand})$, where 'rand' is a uniformly distributed random number in the interval [0,1] (in Matlab given by the command rand) and λ is the reciprocal of rate of the process. This means that the interarrival times between events will be distributed according to an Exponential distribution with mean λ.

The heterogeneous process will be based on two parameters $\lambda_1 = 5$ and $\lambda_2 = 50$ and a stochastic switching mechanism between them. Assume that at $t = 0$, the parameter $\lambda_1 = 5$ will be used to generate t_1. Then, at each $t_i, i \geq 1$, a biased coin is flipped to determine whether the process behavior should from that time point onwards be governed by the other parameter (either λ_1 or λ_2). Let the success probability of this transition (or switch) be $p = 0.1$.

Given the 1000 simulated event times, calculate the ML-estimate of λ (expected value of an Exponential distribution) assuming that all observations come from a single identical distribution. Using this estimate $\hat{\lambda}$, calculate the probability that the waiting time to an event will be $\leq$1. NB This probability can be easily obtained using the cumulative distribution function for the exponential distribution. Compare this probability to the empirical probability of the event, i.e. the relative number of simulated values ≤ 1.

In Matlab, the decision whether to do the parameter switch can be done by simulating one more uniformly distributed random number u in the interval [0,1], and checking whether $u \leq p$ (if yes, then the transition to the other λ-value is done). NB When simulating the event times, keep record of the parameter value under which they are generated (e.g. create a vector of length 1000, where the elements are either 1 or 2, denoting the used parameter value). The knowledge stored in these values will be used in the next exercise.

E4. Continue with data obtained in the exercise E3. The task is to predict the probability of the the waiting time being ≤ 1, under the knowledge of the process state (which parameter is governing the distribution). Make the ML-estimation separately for the two sets of data corresponding to λ_1 and λ_2, respectively. Calculate then separately for $\hat{\lambda}_1$ or $\hat{\lambda}_2$ the estimates of the probability that the waiting time to an event will be ≤ 1. Compare these to the empirical estimates from the corresponding data sets and to the probability obtained in the previous exercise.

E5. Entropy is a central concept in information theory and physics. It is also widely used as a diversity measure in biology. Formally, we may define entropy as

$$h(\mathbf{p}) = -\sum_{j=1}^{k} p_j \log p_j, p_j \geq 0, \sum_{j=1}^{k} p_j = 1, \tag{1}$$

that is, entropy equals the average log-probability of observing a realization from some particular category (among k possible categories), when a population is characterized by the discrete distribution $\mathbf{p}$ =$(p_1, ..., p_k)$. Assume we obtain a sample of $(n_1, ..., n_k), n_j \geq 0, j = 1, ..., k$, observations from an underlying distribution $(p_1, ..., p_k)/$, such that the sample is Multinomial$(p_1, ..., p_k)$-distributed. Notice that n_j equals the number of observations that come from the category j, $j = 1, ..., k$. Show that the maximum likelihood estimate $\hat{h}(\mathbf{p})$ of the entropy $h(\mathbf{p})$ equals $-\sum_{j=1}^{k} \hat{p}_j \log \hat{p}_j$, where $\hat{p}_j = n_j/n$, and $n = \sum_{j=1}^{k} n_j$. Investigate the bias (i.e. the difference between the estimate and the true value) of $\hat{h}(\mathbf{p})$ when $n = 5, 10, 20$ and $k = 5, 10, 15$, and $\mathbf{p}$ is the uniform distribution in each case. Remember that a discrete uniform sample is easily generated in Matlab by ceil(rand(n,1)*k).