Name:.....

Data Mining and Visualization

Assignment 1

Monte Carlo Method, Maximum Likelihood Estimation Method of Moments, Bootstrap, Parametric Bootstrap, Splines

Assignments constitute part of the examination and must be handed in time. You are asked to hand in the solutions during a week following the week on which the assignment has been discussed in classes. You must submit an electronic copy through CANVAS system.

Problem 1 – Method of Moments, Maximum Likelihood, MC Study, Bootstrap

In this problem we review some basic concepts of statistics. It is based on the analysis of Poisson distribution that is a frequently used model to define counting processes, i.e. processes that counts the number of occurrences of certain events. For example the daily number of visits to a certain webpage is given in the following data

```
107  90  71 102  73 100  73 107 116  83 109  99  76  76  97
116  80  80 104  91  73 118 110  73 107  71  82  80 118  75
```

One could be interested in knowing if this data can by possible from a Poisson distribution.

Consider a Poisson random variable X with the parameter λ.

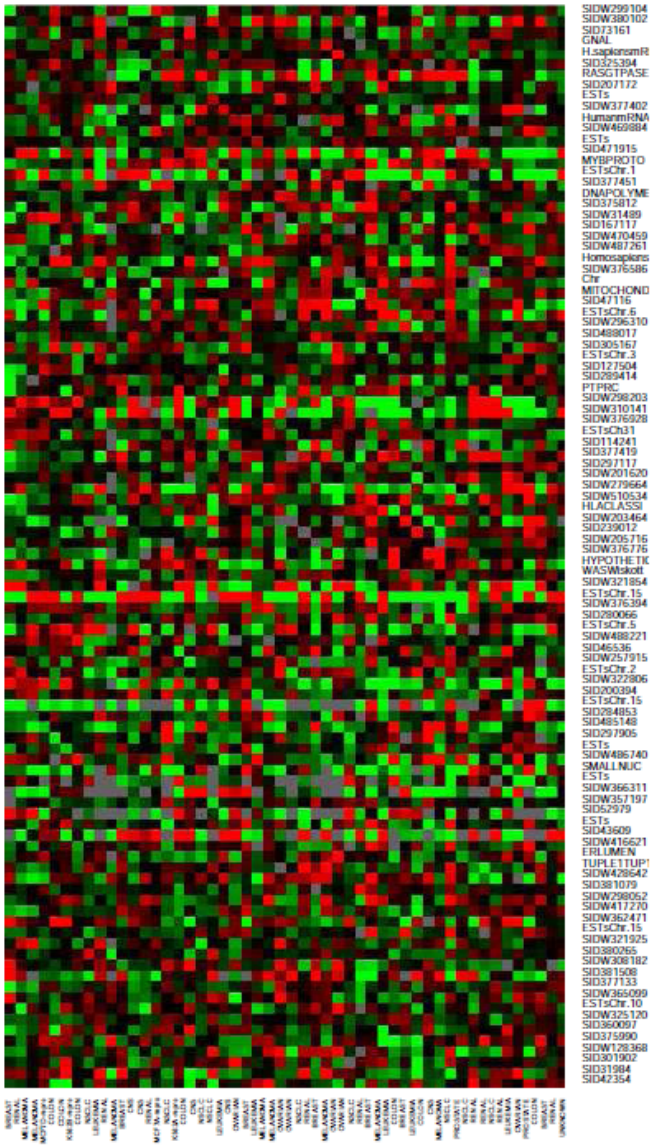
1. Provide the formula for the pdf (probability distribution function) of this distribution.
2. What is the expected value $E(X)$ and variance $Var(X)$?
3. In view of the above relations for the moments, propose two estimators of the parameter λ that would be based on a sample $X_1, \ldots, X_n$ of independent variables having this distribution.
4. Explain in well formulated steps how you could use R (or any other suitable package) to perform a Monte Carlo study to find out which of these two estimators is better. What would be a measure of goodness? Do you have any common sense expectation which of the two estimators will prove to be better?
5. Perform the Monte Carlo study as explained above and draw conclusions about which of the two estimators is better. (This part will be performed in Computer Lab 1, so you may well wait until this session before answering this and related questions).
6. What is the likelihood for the sample $X_1, \ldots, X_n$? What is the log-likelihood? What is the maximum likelihood estimator of λ?
7. Are the maximum likelihood estimators considered to be good estimators? Provide reasons for this. Do these properties 'explain' the results of the Monte Carlo study?
8. How formally, without a Monte Carlo study, can it be argued that one estimator is better than the other in this case?
9. Suppose that you have data $x_1, \ldots, x_n$ that you believe are taken from a Poisson distribution. Describe a bootstrap study that would investigate this hypothesis. What are pros and cons of using bootstrap to perform statistical analysis?

10. Use R to carry out the study suggested in the previous part for the data given for the problem. What are the conclusions? Do the data seem to come from a Poisson distribution?

Problem 2 – Bootstrap Study for Calibration of DNA Microarray

General introduction

"DNA stands for deoxyribonucleic acid, and is the basic material that makes up human chromosomes. DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA (messenger ribonucleic acid) present for that gene. Microarrays are considered a breakthrough technology in biology, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells. Here is how a DNA microarray works. The nucleotide sequences for a few thousand genes are printed on a glass slide. A target sample and a reference sample are labeled with red and green dyes, and each are hybridized with the DNA on the slide. Through fluoroscopy, the log (red/green) intensities of RNA hybridizing at each site is measured. The result is a few thousand numbers, typically ranging from say -6 to 6, measuring the expression level of each gene in the target relative to the reference sample. Positive values indicate higher expression in the target versus the reference, and vice versa for negative values.



A gene expression dataset collects together the expression values from a series of DNA microarray experiments, with each column representing an experiment. There are therefore several thousand rows representing individual genes, and tens of columns representing samples: in the particular example of Figure there are 6830 genes (rows) and 64 samples (columns), although for clarity only a random sample of 100 rows are shown. The figure displays the data set as a heat map, ranging from green (negative) to red (positive). The samples are 64 cancer tumors from different patients.

The challenge here is to understand how the genes and samples are organized. Typical questions include the following:

(a) which samples are most similar to each other, in terms of their expression profiles across genes?

(b) which genes are most similar to each other, in terms of their expression profiles across samples?

(c) do certain genes show very high (or low) expression for certain cancer samples?"

Assessing variability of the technology

However, before answering such medically important questions, it is often important to analyze the accuracy of microarray technology and provide a user some standard estimation of inaccuracy present in the measurements. In such assessment of these variabilities one is not interested in a discovery of any systematic relations between gene expressions but rather in the precision of measurements. Our discussion aims to propose a method to quantify the precision of measurements of gene expressions by designing an experiment and applying a bootstrap method on the obtained samples.

Thus it is of interest to assess the observational noise in results observed in DNA microarrays (uncontrollable inaccuracies). In principle, when data are collected each cell in the dataset matrix (columns are samples and rows refer to genes) may include its own variability independently of variability of genetic material. To eliminate the latter in what follows we assume

1. One organ from one subject is selected for testing. The organ is suppose to have the same genetic composition.
2. Different samples of tissue can be taken from this organ, prepared and applied to different microarrays.
3. One sample of tissue can be prepared in many replicates to be applied to different microarrays.

Further specification of the problem

- If one runs experiments by preparing many replicates from the same tissue sample, then in each cell there will be some variability expressed in different coloration in the outcomes. Since the samples are made of exactly the same tissue specimen, this variability is due to variability in the microarray technology and has nothing to do with variability due to preparation of the genetic material needed for application on a microarray. This is referred to as the microarray variability. From technological considerations, it is assumed that this variability is the same across all cells of a microarray (does not depend on what a specific gene is considered).
- On the other hand one can run the similar experiment but each time using a new sample of tissue (although the tissue is taken from the same organ of the same patient). This again will result in some variability of the observed coloration. This variability contains both the microarray variability and the (tissue)

specimen variability (between different preparations of the tissue). However, it is no longer assumed that different genes are having the same variability across different specimen of the tissue.

Our goal is now to design an experiment that collects data and allows for assessment of these variabilities by the bootstrap method.

Mathematical setup

A microarray yields a matrix of intensities of gene expressions for an individual sample form a tissue relative to the reference sample of genetic material. More precisely, let N be the number of investigated genes and n be a number of samples that are investigated. We assume that there were performed n experiments, each measuring the relative (to a reference sample) intensities of gene expressions for N genes. The outcome can be represented by an array of numbers (typically intensities are gene expression intensities in the range from -6 to 6). Formally, we write

$$\mathbf{X} = (x_{ij})_{i=1,j=1}^{N,n}$$

so that x_{ij} denotes the gene expression for the jth experiment and the ith gene.

One can formulate the problem in more precise terms by introducing proper parameters of experimental variability. Namely, one can assume that observations involve some true average value μ_i that may be different for each gene i but not between different microarrays or tissue samples from the same genetic material. There will be also some random noise that is the sum of the microarray noise ϵ_{ij} and the specimen noise $\tilde{\epsilon}_{ij}$ so that

$$x_{ij} = \mu_i + \epsilon_{ij} + \tilde{\epsilon}_{ij}. \tag{1}$$

Because the microarray noise is the same for each gene we can assume that the variance of ϵ_{ij} is the same between different genes (represented by i's). It is also natural always assume that the variance between samples is constant (the same conditions of the experiments: the same lab, equipment, personnel, etc). However the variance of $\tilde{\epsilon}_{ij}$ would change from gene to gene. Thus we can represent the variability of ϵ_{ij} by the common variance σ^2 and the variability of $\tilde{\epsilon}_{ij}$ by σ_i^2, $i = 1, \dots N$.

In a properly organized experiment, the noises should add independently so that variability from an experiment to an experiment (with different samples of the tissue) are represented by $\tilde{\sigma}_i^2 = \sigma_i^2 + \sigma^2$, for $i = 1, \dots, N$. With this set-up, our goal is to estimate both σ^2 and σ_i^2.

We assume that when a spray of the same tissue (the same genetic material) is put on different microarrays we obtain a version of the experiment given by

$$\tilde{x}_{ij} = \mu_i + \epsilon_{ij}, \tag{2}$$

while by using *different* samples of the tissue sprayed over different microarrays we obtain the full model in (1).

The task

Here are the conditions of experiment. We assume that we have 30 microarrays and 20 specimen of some neutral tissue (neutral so its genetic content is not relevant for our purpose, one could call it a 'calibrating' tissue). From each such a specimen one can take as many samples to be sprayed on a microarray as one wishes. Microarrays cannot be reused. Having these experimental limitations suggest a way to provide a microarray user with some errors assessment for the type of a microarray used in the experiment. More precisely, describe how would you estimate the variances σ^2 and σ_i^2.

1. Describe in detail the experiment that would allow estimation of these parameters. Specify a model for data obtained from the so-designed experiment.
2. Propose bootstrap techniques to assess accuracy of the noise parameter estimation.