Statistical inference, ST703A, advanced level, 7.5 ECTS credits

Computer exercise 1

Assessment

This assignment is a compulsory part of the course. At the end of the session each group's/individual results will be reviewed as **pass** or **fail**. Observe that in order to be able to finish the assignment in the specified time, you are supposed to read the whole assignment and to perform the preparatory exercises **before** you attend the computer session.

1 Estimation of Parameters

This exercise is concerned with parameter estimation, the method of moments and the traditional method of maximum likelihood are demonstrated using Matlab. We will also focus on two different approaches to investigate the variability of estimates, the bootstrap technique and Bayesian analysis. These two methods will then be used for finding interval estimators. Simple examples of the hypotheses testing and some properties of power function are demonstrated.
Special instruction for ST703A. To make files and data available on your computer enter the following commands:

- Go to `System start, Statistical programs`, and choose Matlab.
- Go to `Current directory`, and click on the top-right corner, □. You get a dialog window, `Select a directory`.
- Choose `Den här datorn` and then `Inluppgifter på Studentserver statistik (M:)`
- Go to `casberlab` directory.

If you are doing the exercise at another place, all necessary files are downloadable from `Studentserver statistik (M:)`.

1.1 Preparatory exercises

- Read section 9.2.4 in CB and go through Example 9.2.16.
- Read section 10.1.4 in CB and go through Examples 10.1.19-10.1.22.

2 The Method of Moments and Maximum Likelihood Estimation

We will take the following basic approach to the study of parameter estimation. The observed data will be regarded as realizations of random variables $X_1, \ldots, X_n$, whose joint distribution depends on an unknown parameter θ. Note that θ can be a vector, such as (α, β) in *gamma* distribution. An estimate of θ will be a function of X_is and will hence be a random variable with the probability distribution called its *sampling distribution*. We will use approximations to the

sampling distribution to assess the variability of the estimate.

In this part we will use numerical examples in Matlab to investigate the properties of point estimates. We will also be concerned with sampling distributions of estimates and with assessing variability using bootstrap.

2.1 Poisson Distribution

As a concrete example, let us consider a study done at the National Institute of Science and Technology (Steel et al 1980, see Rice (1996)). Asbestos fibers on filters were counted as part of a project to develop measurements standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and punches of 3-mm diameter were taken from the filter and mounted on a transmission electron microscope. An operator counted the number of fibers in each of 23 grid squares, yielding the counts that are stored in the file

`asbestos.mat`

The Poisson distribution, $Po(\lambda)$ would be a plausible model for describing the variability from grid square to grid square in this situation and could be used to characterize the inherent variability in future measurements. Determine the method of moments estimate of λ. Specify the log-likelihood function for λ and determine the ML estimate. Are these two estimates the same?

Answer:

To make a visual inspection of the log-likelihood function based on the asbestos data we can plot it using the following commands:

```
>> lambda=[20:0.05:30];
>> l=log(lambda)*sum(asbestos)-23*lambda-sum(log(factorial(asbestos)));
>> plot(lambda,l)
```

Explain your results.

If the experiment were to be repeated, the counts, and therefore the estimate would not be the same. In order to investigate how stable the obtained estimates are we turn to the standard technique based on deriving the sampling distribution of the estimate or an approximation to that distribution. In our example the statistical model stipulates that the individual counts X_i are iid from $Po(\lambda_0)$. Letting $S = \sum_{i=1}^{n} X_i$, the parameter estimate $\hat{\lambda} = S/n$ is a random variable, and to determine its sampling distribution we observe that by the properties of Poisson distribution, $S \sim Po(n\lambda_0)$. Thus the pmf of $\hat{\lambda}$ is

$$P(\hat{\lambda} = \nu) = P(S = n\nu) = \frac{(n\lambda_0)^{n\nu} e^{-n\lambda_0}}{(n\nu)!}$$

for such ν that $n\nu$ is a nonegative integer. Suggest an approximation to the distribution of S and $\hat{\lambda}$ if $n\lambda_0$ is large and determine corresponding parameters. Plot the density of the suggested approximate distribution.

Answer:

Use the *estimated* standard error $s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}$ instead of the standard error of $\hat{\lambda}$, $\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda_0}{n}}$ when deriving the approximation and compute s from the asbestos data. Give some ideas about theoretical justification of using $\hat{\lambda}$ instead of λ_0. Use properties of the approximate distribution to make assessments about the variability of the estimates.

Answer:

In some cases the sampling distribution is of an explicit form depending upon the unknown parameters. In these cases we could substitute our estimates for unknown parameters in order to approximate the sampling distributions. In other cases the form of sampling distribution is not so obvious, but we could use bootstrap resampling technique to simulate it. Furthermore by using the bootstrap we avoid doing perhaps difficult analytic calculations. We have seen that in the case of $Po(\lambda)$ the closed form approximation of the sampling distribution of $\hat{\lambda}$ can be obtained. The bootstrap approximation can also be suggested using

```
>> [bootstat,bootsam] = bootstrp(1000,@mean, asbestos);
>> [f,x]=ecdf(bootstat);
>> ecdfhist(f,x,20)
```

Explain commands and results.

2.2 Gamma distribution

Here we will use the following form of the density of gamma distribution

$$f(x|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \quad 0 \le x < \infty, \quad \alpha,\lambda > 0$$

and assume that both parameters are unknown. Observe that in the notations of CB $\lambda = 1/\beta$, see p.99. Recall that during the lecture 3 were we used the method of moments to estimate α and λ that can shortly be presented as follows: the first two population moments of the gamma distribution are

$$E(X) = \frac{\alpha}{\lambda}, \quad E(X^2) = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

Then, using the sample moments

$$m_1 = \frac{x_1 + \ldots + x_n}{n}, \quad m_2 = \frac{x_1^2 + \ldots + x_n^2}{n}$$

we get

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2} \quad \text{and} \quad \hat{\lambda} = \frac{m_1}{m_2 - m_1^2}$$

It appears that it would be difficult to derive the exact forms of the sampling distributions of $\hat{\alpha}$ and $\hat{\lambda}$ since they are each rather complicated functions of the sample $X_1, \ldots, X_n$. To evaluate the variability of $\hat{\alpha}$ $\hat{\lambda}$ we use the bootstrap approach.
As an example, we consider the fit of the amount of precipitation during 227 storms in Illinois from 1960 to 1964 to a gamma distribution (LeCam and Neyman 1967, see Rice). The idea with the data analysis was to characterize the natural variability in precipitation from storm to storm. The data gathered are summarised in the file `rain.mat`. The first column of the data represents the average amount of rainfall (in inches) from each storm, the second column represents the corresponding year of measurements where 0 stands for 1960, 1 for 1961, and so on. Display the data as a normalized histogram and explain why gamma distribution is a good candidate for a model.

Answer:

Calculate then the method of moments estimators for this data and use these estimators as parameters to fit the density of gamma distribution to the data. Apply then `gammafit` to the data to specify the ML estimators of α and λ, and plot the fitted density with $\hat{\alpha}_{\mathrm{ML}}$ and $\hat{\lambda}_{\mathrm{ML}}$ on the same figure using `hold on`, (details of `gammafit` will be discussed later). Explain the discrepancy between the fitted densities? Keep in mind that the gamma distribution is only a possible model for the data and should not be taken as being literally true.

Answer:

We now turn to the analysis of the sampling distributions of $\hat{\lambda}$ and $\hat{\alpha}$ obtained by the method of moments. We generate many samples of the size $n = 227$ from the rain data and then illustrate the estimates variability by histograms.

```
>> [boot_mean,bootsam] = bootstrp(1000,@mean,rain(:,1));
>> [boot_var,bootsam] = bootstrp(1000,@var,rain(:,1));
>> m_1=bootstat_mean;
>> m_2=boot_var+m_1.*m_1;
>> alphahat=m_1.*m_1./(m_2-m_1.*m_1)
>> betahat=(m_2-m_1.*m_1)./m_1;
>> figure
>> subplot(1,2,1)
>> hist(alphahat)
>> subplot(1,2,2)
>> hist(betahat)
>> colormap([0.5 0.5 0.5])
```

The obtained histograms indicate the variability that is inherent in estimating the parameters from a sample of the size $n = 227$. Can you suggest some approximations to the obtained distributions from the shape of histograms?

The variability shown by the histograms can be summarized by calculating the standard deviations of the 1000 estimates, thereby providing estimated standard errors of $\hat{\lambda}$ and $\hat{\alpha}$. To be more precise, if we get 1000 estimates of α denoted by α_i^*, $i = 1, \ldots, 1000$, the standard error of $\hat{\alpha}$ is estimated as

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\alpha_i^* - \bar{\alpha})^2}$$

where $\bar{\alpha}$ is the mean of the 1000 values. Calculate the standard errors for $\hat{\lambda}$ and $\hat{\alpha}$ using the bootstrap samples above and explain your results.

Answer:

Now we turn to the method of maximum likelihood and investigate the variability of ml estimates of α and λ using the same precipitation data summarised in the file `rain.mat`. When both α and λ are unknown, the log-likelihood function of an i.i.d sample $X_1, \ldots, X_n$ is

$$l(\alpha, \lambda) = n\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i - n \log \Gamma(\alpha).$$

By calculating the partial derivatives $\partial l/\partial \alpha$ and $\partial l/\partial \lambda$ and setting them equal to zero we find $\hat{\lambda} = \frac{\hat{\alpha}}{\bar{x}}$ and by substituting this into the equation for $\partial l/\partial \alpha$ we obtain a nonlinear equation for the ML estimate of α:

$$n \log(\hat{\alpha}) - n \log \bar{x} + \sum_{i=1}^{n} \log x_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

This equation cannot be solved in the closed form and therefore obtaining the exact sample distributions of the estimates would appear to be intractable. Recall that we discussed the same problem with the ML estimates of gamma distribution parameters, see Exercise 7.10 in CB. However iterative methods for finding estimates are used in `gamfit(data)` function which returns the maximum likelihood estimates (MLEs) for the parameters of the gamma distribution given the data in vector `data`. Do the following steps

```
>> [paramhat,ci]=gamfit(rain(:,1))
>> datarain = gamrnd(paramhat(1),paramhat(2),1,227);
>> [bootstat,bootsam] = bootstrp(1000,@gamfit, datarain);
>> bootstat
>> figure
>> subplot(1,2,1)
>> hist(bootstat(:,1))
>> subplot(1,2,2)
>> hist(bootstat(:,2))
>> colormap([0.5 0.5 0.5])
```

The first step gives us the ML estimates for α and $\beta = 1/\lambda$ obtained from the rain data. To evaluate the variability of these estimates we use the bootstrap, and since the true values of α

and β are unknown we let our ML estimates play their roles. Explain the commands above. Can we regard the obtained histograms as approximations to the sampling distributions of $\hat{\alpha}$ and $\hat{\beta}$?

Answer:

Now we can compare the sampling distributions of α and $\beta = 1/\lambda$ obtained by method of moments and maximum likelihood. What can you conclude about the dispersion of these distributions? Relate your conclusion to the properties of estimators. Calculate the standard errors for $\hat{\alpha}$ and $\hat{\beta}$ and compare them to those for estimates obtained by the methods of moments. Observe that two different bootstrap approaches were used when specifying the method of moments and the ml estimates. Explain what is the difference and how does it effect the variability of estimates. *Optional*: Try to use the same bootstrap approaches for both estimation techniques and compare the variability of resulting estimates.

Answer: