Statistical inference, ST703A, advanced level, 7.5 ECTS credits

Computer exercise 2

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.
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:)`.

Preparatory exercises

- Repeat Section 9.2.2, check the structure of pivot confidence interval for location and scale parameters, check Example 9.1.3 and Example 9.1.6.
- Check the Example 9.2.7 and Example 9.2.8, and Example 9.2.13.
- Read the Section 9.2.4, check Example 9.2.16 and Example 9.2.17.

1 Interval Estimation

In this part of the Exercise we consider the use of the bootstrap for finding approximate confidence intervals. Suppose that $\hat{\theta}$ is a point estimate of a (location) parameter θ, the true unknown value of which is θ_0, and suppose for the moment that the distribution of $\hat{\theta} - \theta_0$ is known. Then let the $\alpha/2$ and $1 - \alpha/2$ be quantiles of this distribution which we denote by $\underline{\delta}$ and $\bar{\delta}$ respectively, i.e

$$P(\hat{\theta} - \theta_0 \leq \underline{\delta}) = \frac{\alpha}{2}, \quad P(\hat{\theta} - \theta_0 \leq \bar{\delta}) = 1 - \frac{\alpha}{2}.$$

Then

$$P(\underline{\delta} \leq \hat{\theta} - \theta_0 \leq \bar{\delta}) = 1 - \alpha$$

or

$$P(\hat{\theta} - \bar{\delta} \leq \theta_0 \leq \hat{\theta} - \underline{\delta}) = 1 - \alpha.$$

Thus we have that $(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta})$ is $100(1-\alpha)\%$ confidence interval.

However the known distribution of $\hat{\theta} - \theta_0$ typically is not the case. Since θ_0 is unknown, the bootstrap technique suggest using $\hat{\theta}$ in its place: we can generate many samples (say b in total) from a distribution with value $\hat{\theta}$ and for each sample construct an estimate of θ, say θ_j^*, $j = 1, \ldots, b$. The distribution of $\hat{\theta} - \theta_0$ is then approximated by that of $\theta^* - \hat{\theta}$, the quantiles of which are used to form an approximate confidence interval.
Using the similar approach the bootstrap confidence interval for the scale (or shape) parameter θ can be constructed. We use

$$P(\frac{\hat{\theta}}{\theta_0} \leq \underline{\delta}) = \frac{\alpha}{2}, \quad P(\frac{\hat{\theta}}{\theta_0} \leq \bar{\delta}) = 1 - \frac{\alpha}{2}.$$

Then

$$P(\underline{\delta} \leq \frac{\hat{\theta}}{\theta_0} \leq \bar{\delta}) = 1 - \alpha,$$

or

$$P(\frac{\hat{\theta}}{\bar{\delta}} \leq \theta_0 \leq \frac{\hat{\theta}}{\underline{\delta}}) = 1 - \alpha,$$

where we assume that the distribution of $\frac{\hat{\theta}}{\theta_0}$ can be approximated by that of $\frac{\theta^*}{\hat{\theta}}$, with the corresponding quantiles $\underline{\delta}$ and $\bar{\delta}$.

1.1 Bootstrap CI for parameters of gamma distribution

Now we apply the technique presented above to find approximate confidence intervals for the parameters of the gamma distribution fitted to the precipitation data. The first step gives the ml estimates of α and λ (or β) using `gamfit`, (recall that the Matlab determines the estimate of $\beta = 1/\lambda$).

```
>> [paramhat,ci]=gamfit(rain(:,1))
>> datarain=gamrnd(paramhat(1),paramhat(2),1,227);
>> b=1000; [bootstat,bootsam] = bootstrp(b,@gamfit, datarain);
>> alphaboot=bootstat(:,1);
>> delta=alphaboot/paramhat(1);
>> delta=sort(delta);
>> qq_delta=delta(round([b*0.025,b*0.975]))
>> ci_boot_alpha=[paramhat(1)/qq_delta(2),paramhat(1)/qq_delta(1)]
```

Explain what the Matlab code does. Use the same technique to compute the bootstrap confidence interval for β. Compare the bootstrap intervals for α and β with those obtained by `gamfit`.
Notice that there are a number of different methods of using the bootstrap technique to find approximate confidence intervals. The preceding method was demonstrated because the fairly direct reasoning behind it. Another popular method, *the bootstrap percentile method*, uses the quantiles of the bootstrap distribution of $\hat{\theta}$ directly.

Answer:

1.2 Analysis of earthquakes data using credibility and confidence intervals

Recall Section 2.2 of the Computer Exercise 2 of the Probability course where we analysed the variability of an estimator using Bayesian approach. The main goal was to evaluate the probability of having a period of more than 1200 days between serious earthquakes using the data `dataquak`; see Rychlik et al (2006). We investigated the posterior distribution of

$$p = P(X > 1200) = e^{-1200\Lambda},$$

given the non-informative prior prior of Λ, $\pi(\lambda) = \lambda^{-1}$ It was shown that the posterior distribution of p can be approximated by the log-normal one, i.e

$$\log(p) \sim \mathcal{N}(\cdot, \cdot)$$

which is illustrated by

```
>> figure
>> x=[0.01:0.001:0.16];
>> pi_postpdf_p=lognpdf(x,-1200*(63/27120), 1200*sqrt(63/(27120^2)));
>> subplot(2,1,1)
>> plot(x,pi_postpdf_p)
>> title('Posterior density of p')
>> subplot(2,1,2)
>> pi_postcdf_p=logncdf(x,-1200*(63/27120), 1200*sqrt(63/(27120^2)));
>> plot(x,pi_postcdf_p)
>> title('Posterior distribution of p')
```

Use `zoom` and specify the credibility interval of p from the figure you just plotted.

$[p_{0.975}, p_{0.025}] =$

To calculate the bootstrap confidence interval for p we need the distribution of $\hat{p}/p_0$ which is unknown. We approximate it by the distribution of $\delta = p^*/\hat{p}$ where p^* is a bootstrap estimator of p_0 obtained by the following steps:

```
>> load data_quak
>> p_hat=exp(-1200./mean(data_quak))
>> b=1000;
>> [bootstat, bootsam]=bootstrp(b,@mean,data_quak);
>> data_boot=data_quak(bootsam);
>> p_star=exp(-1200./mean(data_boot));
>> delta= p_star./p_hat;
>> [f,x]=ecdf(delta);
>> ecdfhist(f,x,20)
>> colormap([0.5 0.5 0.5])
```

The last three steps give a normalized histogram representing the bootstrap distribution of δ. Now we specify the quantiles of this distribution and get the confidence interval:

```
>> delta=sort(delta);
>> q_delta=delta(round([b*0.025,b*0.975]))
>> ci_boot=[p_hat/q_delta(2),p_hat/q_delta(1)]
```

Compare the credibility interval with the bootstrap based confidence interval. Are they similar? Observe that the posterior distribution of p is skewed; you can relate the skewness of the distribution to the closeness of the upper endpoint of the credibility interval to zero.

Answer:

2 Hypothesis Testing

Random variation in the data often makes it difficult to determine whether samples taken under different conditions really are different. Hypothesis testing is a tool for analyzing whether sample-to-sample differences are significant and require further investigation or are consistent with random and expected data variation. Hypothesis testing for the difference in means of two samples from a normal distribution can be performed using `ttest2`. Read the instructions for this procedure using `help ttest2` and apply it to solve the Problem 8.41 (c) in CB. Try different assumptions about population variances.

Answer:

Observe that besides `ttest2` Matlab provides a number widely used parametric and nonparametric hypothesis testing procedures summarized in `Hypothesis tests`.

Further we consider some examples of calculating the power of a test, that is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. We focus on a two sided test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

that rejects the null hypothesis whether the sample mean is too high or too low. The test statistic is a t statistic, which is the difference between the sample mean and the mean being tested, divided by the standard error of the mean. Under the null hypothesis, the test statistics has Student t distribution with $n-1$ degrees of freedom. To investigate the power of the test we can look at power function as a function of μ using `sampsizepwr` function that allows for computing the sample size and test power. For a given $\mu_0 = 75$, sample size $n = 15$ and $\sigma = 6$ the power function can be illustrated by

```
>> n=15;x = linspace(65,85);
>> power=sampsizepwr('t',[75 6],x,[],n);
>> plot(x,power);
>> xlabel('True mean')
>> ylabel('Power')
```

Investigate how power increases as μ moves away from the null hypothesis value in either direction. Investigate also how the sample size, n effects the shape of the power function.

Answer:

Using `sampsizepwr` function we can also determine the mean closest to μ_0 that can be determined to be significantly different from μ_0 using a t test with a given sample size of $n = n_0$ and given power $\beta(\mu) = \beta_0$. Try for example

```
>> mu1 = sampsizepwr('t',[75 6],[],0.8,30)
```

and explain your results.

References

1. Rychlik I, & Rydén J. Probability and Risk Analysis, Springer Berlin Heidelberg, 2006.
2. Rice J. Mathematical statistics and data analysis. Duxbury Press, 1996.