Probability Theory, ST701A, 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.

Properties of a random sample: Convergence concepts, resampling technique and Bayesian methodology

In the first part of this computer exercise, we will encounter some fundamental properties of the limiting behavior of the sum of independent random variables as the number of summands become large. We focus on the Law of Large Numbers and Central Limit Theorem. The results demonstrated in the Exercise are both intrinsically interesting and useful in e.g. statistical inference, since many commonly computed statistical quantities, such as averages, can be represented as sums. In the second part of the Exercise we introduce two different techniques for assessing variability of estimates, the bootstrap resampling and Bayesian analysis.

Special instruction for ST701A. 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 down-loadable from `Studentserver statistik (M:)`.

Preparatory exercises

As a preparation the Exercise you need to read the instructions for the computer exercise, Sections 2.1, 3.2-3.3 4.5-4.6 in the course book by Casella, Berger (CB) and present answers to the following questions

1. Explain the following types of limiting behavior of certain sample quantities and relationship between them: Convergence in probability, almost sure convergence and convergence in distribution.

2. State Week and Strong Low of Large Numbers.

3. Describe the Central Limit Theorem and give an example of how it can be used in probabilistic reasoning?

4. Let X be a random variable showing the numbers 1 through 6 when you roll a single six-sided (fair) die, i.e. $f_X(k) = 1/6$ for $k = 1, \ldots, 6$. Which distribution the sum of the n independent dice rolls will have approximately when n is large?

5. A six-sided die is rolled independently 100 times. Suggest an approximation and find the probability that the face showing a six turns up between 15 and 20 times. Find the approximate probability that the sum of the face values of the 100 trials is less than 300.

6. Suppose that a coin is tossed independently 100 times and lands heads up 60 times. Should we be surprised and perhaps doubt that the coin is fair? Suggest an exact and approximate solution.

7. How do you interpret a normal probability plot?

8. Check the Example 5.6.6 from the Section 5.6 in CB.

9. Explain Bayes Rule and Total Probability Theorem.

10. Recall the idea of sampling with replacement and check the Example 1.2.20 in CB.

1 Convergence concepts

1.1 Law of Large Numbers

The Week Law of Large Numbers (WLLN) states that, under certain conditions, the sample mean $\bar{X}_n$ approaches the population mean, μ as $n \to \infty$. Observe that the property summarized by the WLLN, that a sequence of the samesample quantity approaches a constant as $n \to \infty$, is known as *consistency*. We will focus on this property more closely in the Inference part of the course.
A simple empirical demonstration of the LLN is to simulate rolling dice and then observe that the successively calculated sample mean converges towards a population mean as the number of trials increases. We use a simple dice generator and simulate first 100 rolls:

```
>> x=floor(6*rand(1,150)+1)
```

where the function `floor` rounds the elements of x downward to the nearest integers. Observe that each element of x is an observation from the distribution of X describing rolling die. In order to calculate a successive sample mean we do

```
>> x_bar=cumsum(x)./(1:150)
```

where the function `cumsum` returns a vector whose ith element is the cumulative sum of the first i elements of the generated vector x and ./ denotes element-by-element division. In this way the elements of the resulting vector, $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_{100})$, are successively calculated sample means. Try to plot them by

```
>> plot(x_bar,'.')
```

and explain the resulting figure. Do you observe what was expected?

Answer:

Try now increase the number of rollings in the generator above and make a number of plots (use `subplot`) for different n. What do you observe?

The most important cases to which the LLN does not apply involve Cauchy distribution with the density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad -\infty < x < \infty$$

The density is symmetric about zero, so it would seem that $E[X] = 0$. However,

$$\int_{-\infty}^{\infty} \frac{|x|}{1+x^2} dx = \infty,$$

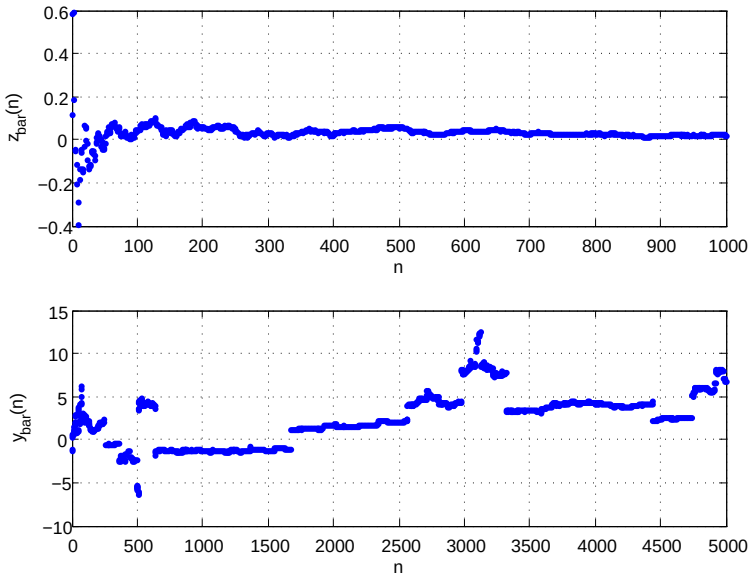
recall the Exercise 3.21 from CB, and therefore the expectation does not exist. The reason that it fails to exists that the density decreases so slowly that very large values of $X \sim f(x)$ can occur with substantial probability. Recall also that if Z_1 and Z_2 are independent standard normal random variables then $X = Z_2/Z_1$ has a Cauchy distribution; see Example 4.3.6 in CB. Like the standard normal density, the Cauchy density is bell-shaped and symmetric about zero but the tails of the Cauchy tend to zero very slowly compared to the tails of the normal. This can be interpreted as being due to a substantial probability that Z_1 in the quotient Z_2/Z_1 is near zero.

To give an empirical illustration of this property of Cauchy distribution, we first notice that the the relationship between the standard normal and Cauchy distributions gives a simple way of generating Cauchy random variable. We first generate two sequences z_1 and z_2 of observations of the independent standard normal random variables and then get their quotient which follows a Cauchy distribution. To depict the asymptotic behavior average of n independent random variables as a function of n for normal and Cauchy random variables we do the following commands

```
>> figure
>> subplot(2,1,1)
>> z_1=randn(5000,1); z_2=randn(5000,1);
>> z_bar=cumsum(z_1(1:1000))./(1:1000)'; plot(z_bar,'.')
>> subplot(2,1,2)
>> y=z_1./z_2; y_bar=cumsum(y).\(1:5000)'; plot(y_bar,'.')
```

Observe that you would probably need to perform these a number of times to achieve a graph like in Figure 1 where $\bar{Z}_n$ appears to be tending to the limit, whereas $\bar{Y}_n$ does not. Explain why. Recall the property of t distribution and think how it is related to the Cauchy distribution. Use this relationship and suggest another technique for generating Cauchy random variables. *Hint:* Try to use

```
>> t=trnd(1,5000,1); t_bar=cumsum(t)'./(1:5000); plot(t_bar,'.')
```



Figur 1: The average of n independent random variables as a function of n for normal random variables (upper panel) and for Cauchy random variables (lower panel).

Compare behaviour of the sample mean of the Cauchy distribution to the sample median m, where $P(X \geq m) = P(X \leq m)$ which is known to be a more stable estimator of location parameter.(see Exercise 3.39) You can calculate the successive sample median by the first 1000 observations and explain your results.

```
>> t=trnd(1,5000,1);for i=1:1000; t_median(i)=median(t(1:i)); end
>> plot(t_median,'.')
```

1.2 Central Limit Theorem

We start by investigation of a discrete distribution, for example a uniform distribution that we can obtain by the above mentioned dice generator. We represent the distribution by the vector

```
>> f=[0 1 1 1 1 1 1]/6
```

where the first component 0 represents the probability the the outcome is 0 is used here for the technical convenience. The graph representing the distribution is given by

```
>> bar(0:length(f)-1,f)
```

Now in order to find out the probability mass function of the sum of two dice we use the *convolution* function `conv` that gives the distribution of sum of two independent random variables.

```
>> f_2=conv(f,f)
>> bar(0:length(f_2)-1,f_2)
```

To establish the distribution of the sum of 4,8 and more dice, we do

```
>> f_4=conv(f_2,f_2);
>> subplot(2,1,1)
>> bar(0:length(f_4)-1,f_4)
>> f_8=conv(f_4,f_4);
>> subplot(2,1,2)
>> bar(0:length(f_8)-1,f_8)
```

where f_4 and f_8 are the probability function of a sum of four and eight dice, respectively. Investigate the shape of the distribution, what do you observe? When the resulting graph starts to approach the symmetric shape?

Answer:

Try to perform the same steps with a very skew distribution, investigate how many summands will be needed to observe a symmetric shape of the resulting distribution.
When you roll a symmetric single six-sided die, of outcomes have mean 3.5 and variance 35/12,

```
>> mu=sum((0:6).*f)
>> sigma=sqrt(sum(((0:6)-mu).^2.*f))
```

and so the corresponding mean and variance of rolling of n dice is n times greater. One can apply the Central Limit Theorem to approximate the resulting distribution by the normal one which is given by $\mathcal{N}(n\mu, n\sigma^2)$. For example, we can compare the distribution of the sum of eight independent dice rolling with $\mathcal{N}(n\mu, n\sigma^2)$ where $n = 8$.

```
>> bar(0:length(f_8)-1,f_8)
>> hold on
>>
```

It is known that for i.i.d $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$, the distribution of the sample mean, $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/n)$, i.e also normal. However if we are interested in using a more robust estimator of location such as the median, it becomes more difficult problem to derive its distribution. For a sample of size $n = 15$ from $\mathcal{N}(0, 1)$, simulate the distribution of a median $\tilde{X}$. Compare then the sample distribution of $\tilde{X}_n$ with the asymptotic distribution of the median $\sqrt{n}(\tilde{X} - \mu)$ using the following result for a symmetric distribution with location parameter (distribution median) θ

$$\sqrt{n}(\tilde{X}_n - \theta) \to \mathcal{N}(0, 1/(4f_X^2(\theta))), \quad \text{as} \quad n \to \infty$$

in distribution, where f_X is the pdf of X. Is $n = 15$ is large enough for the asymptotic to be valid?

Answer:

2 Analysis of earthquake data

In this section we use the earthquake data to explore some properties of random samples and sampling distributions. The patterns of occurrence of earthquakes in terms of time, space and magnitude are very erratic, and attempts are sometimes made to construct probabilistic models of these events. The models may be used in a purely descriptive manner or, more ambitiously, for purposes of predicting future occurrences and consequent damage.
The data we are going to work with are stored in the file

`data_quak.math`

and represent the observed times separating a sequence of serious earthquakes (>7.5 on Richter scale) worldwide. These data have been gathered during the period of time from the December 1902 to March 1977, giving in total 63 earthquakes during 27120 days (see Rice, 1999 and the references there in, Udias and Rice, 1975 and Wafo).
The natural first step in exploring the data is the graphical representation.

```
>> load data_quak
>> figure
>> [f,x] = ecdf(data_quak);
>> ecdfhist(f,x,10);
>> colormap([0.5 0.5 0.5])
```

Let X denote the length of time unit until the next earthquake, or the interoccurence time. Then a good candidate to model this time is an exponential distribution, $Exp(\mu)$ where μ is a return period of earthquakes, i.e. we can assume that

$$F_X(x|\mu) = 1 - e^{-x/\mu}.$$

To show this we can try to fit the exponential density to the data. Explain the commands below

```
>> xx = 0:1:max(data_quak); yy = exppdf(xx,mean(data_quak));
>> hold on; plot(xx,yy,'g-');
```

Observe that an exponential model for interoccurence times would be memory less; that is, knowing that an earthquake had not occurred in the last t time units would tell us nothing about the probability of occurrence during the next s time units. Can you suggest some other distribution that dose not have this property to model the interoccurence times? Try to fit the suggested distribution to the data and explain your results.

Answer:

Now we use the data above to evaluate the probability of a period of more than 1200 days (about 3.5 years) between two serious earthquakes. Assuming that the data are coming from the exponential distribution, $Exp(\mu)$ we then need to specify

$$p = P(X > 1200|\mu) = 1 - F_X(1200|\mu) = e^{-1200/\mu}$$

Since the distribution parameter, μ, is unknown we use our data to estimate it and recall that μ is a mean parameter, so that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where $n = 62$ and x_i are the observed interoccurence times. Then the estimated probability $\hat{p}$ is given by

$$\hat{p} = e^{-1200/\hat{\mu}}.$$

Another way to estimate p is to use the empirical distribution of the earthquake data. We can plot it by

```
>> figure
>> [f,x] = 1-ecdf(data_quak); stairs(x,f);
>> grid on
```

and then use `zoom` in the figure window to get the value p_{emp} at $x = 1200$. Compare p_{emp} with $\hat{p}$ and explain the discrepancy.
Note that both the results above gives a point estimation of p, i.e. only one estimate $\hat{p}$ or p_{emp} of p. But how accurate this result is? To measure the estimator variability we can try to evaluate its variance. We focus on $\hat{p}$ and observe that by considering X_is as n i.i.d. random variables from $F_X(x|\mu)$, $\hat{\mathcal{M}}$ and $\hat{P}$ are also random. However finding an analytic expression for the variance of $\hat{P}$ is not so easy (Try!). Below we will consider two different methods of evaluating the variability of $\hat{P}$. The first one is based on the resampling methodology whereas the second one uses the Bayesian approach.

2.1 Variability of $\hat{P}$: the Bootstrap technique

In this part we use the Bootstrap to approximate the sampling distribution of the estimator $\hat{P}$ and thereby get an idea about the variability of $\hat{P}$. It is important to keep in mind that $\hat{P} = \hat{P}(X_1, \ldots, X_n)$, i.e. $\hat{P}$ is a function of random variables, and hence has a a probability distribution, its sampling distribution, which is determined by n and the underlying distribution $F_X(x)$. We would like to know this sampling distribution, but we are faced with two problems: (1) generally we don't know $F_X(x)$ (recall that we *assume* that the earthquake data are coming form $Exp(\mu)$, however, another distribution, e.g. $gamma(\alpha, \beta)$ can also be a data generating candidate), and (2) even if we knew $F_X(x)$, it is relatively complicated to find the closed form analytic expression for the distribution of $\hat{P}$. To overcome these problems, we use the resampling

approach.

To explain the underlying rationale of the Bootstrap resampling technique we first address the second problem. Suppose for a moment that $F_X(x|\mu)$ is $Exp(\mu)$ and $\hat{\mu}$ is an estimation of μ obtained from the original data $x_1, \ldots, x_n$. To avoid complicated analytic calculations of the distribution of $\hat{P}$ we turn to simulation: we generate many, many samples, say b in number, of size n from $F_X(x|\hat{\mu})$; from each sample we calculate the value of $\hat{p}$. The empirical distribution of the resulting values $p_1^*, \ldots, p_b^*$ is an approximation to the distribution function of $\hat{P}$, which is good if b is large. If we wish to know the variance $\hat{P}$, we can find a good approximation to it by calculating variance of the collection of values $p_1^*, \ldots, p_b^*$ as

$$s_{\hat{p}}^2 = \frac{1}{b}\sum_{i=1}^{b}(p_i^* - \bar{p}^*)^2,$$

where $\bar{p}^*$ is the mean of $p_1^*, \ldots, p_b^*$ Note that we can make these approximations arbitrary accurate by taking b to be arbitrary large. Note also that the samples used above are not resamplings of the data, but actual random samples drawn from *plug-in distribution* $F_X(x|\hat{\mu})$. This approach is a version of the *parametric bootstrap* since we use a parametric model at the sampling stage. See more details and examples in e.g. Section 5.6 of CB.

For the first problem, i.e. for the more general case when the underlying distribution F is unknown, the bootstrap solution is to view the empirical cdf F_n as an approximation of F and sample from F_n. That is F_n would be used in place of F in the previous paragraph. To run sampling from F_n we will reason as follows: F_n is a discrete probability distribution that gives probability mass of $1/n$ to each observed value $x_1, \ldots, x_n$. A sample of size n from F_n is thus a sample of size n drawn *with replacement* from the collection $x_1, \ldots, x_n$. That is we learn about the sample characteristics by taking *resamples*, i.e we *retake* samples from the original data set $x_1, \ldots, x_n$. We thus draw b samples of size n with replacement from the observed data, producing $p_1^*, \ldots, p_b^*$. The variance of $\hat{P}$ can be approximated as above. This type of bootstrapping is called *non-parametric bootstrap* as we have assumed no functional form for F.

Now we use the non-parametric bootstrap to approximate the distribution of $\hat{P}$. Using the `bootstrap` function you can resample the vector `dataquak` as many times as you like and consider the variation in the resulting estimation of p^*. To get the output samples `bootsam` without applying a function, set `bootfun` argument to empty ([]) (doesn't work properly some time, that's why we use `mean` in the command below). To create 1000 bootstrap samples from `dataquak` we start by

```
>> [bootstat,bootsam] = bootstrp(1000,@mean, data_quak);
>> bootsam(:,1:5)
```

where the last command displays the indices of the data selected for the first 5 bootstrap samples. Observe that `bootsam` contains 1000 index vectors of the length of 62, and one index can occur more than once (resampling with replacement). By the commands

```
>> data_boot=data_quak(bootsam);
>> size(data_boot)
```

the earthquakes data are resampled to create 1000 different data sets, and we compute the estimators $p_1^*, \ldots, p_b^*$ by

```
>> p_star=exp(-1200./mean(data_sam));
```

Check the size of resulting vector `pstar` and convince yourself that you get $b = 1000$ estimators p_i^*, $i = 1, \ldots, b$. Further, to visualise the variability of $\hat{P}$ we depict its the approximate distribution as follows:

```
>> figure
>> [f,x] = ecdf(p_star); ecdfhist(f,x,20);
>> ecdfhist(f,x,20);
>> colormap([0.5 0.5 0.5])
```

Explain the graph. Try to use the simulated data to approximate the variance of $\hat{P}$.

Answer:

2.2 Estimating the variability of p using Bayesian approach

The Bayesian approach to statistics is fundamentally different from the classical inference that we have used above. In the classical approach the parameter μ is thought to be unknown, but fixed, quantity. In the Bayesian inference μ is considered to be a quantity whose variation can be described by a probability distribution (called *prior distribution*). A sample is then taken from a population indexed by $\mathcal{M}$ and the prior distribution is updated with with the sample information. This updated prior is called the *posterior distribution*. The updating is done with the use of Bayes' rule, see Chapter 1 in CB, and hence is the name Bayeisan approach. In this part of the Exercise we will only point out some basic ideas and show how to analyse the same earthquake data set with a Bayesian method. More detailed discussion of these concepts is left for the Inference part of the course.

We assume that the return period, is a random variable and focus instead on the intensity of earthquakes $\Lambda = \frac{1}{\mathcal{M}}$ instead of the return time $\mathcal{M}$. Then the probability p that we wish to evaluate is

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

As before our goal is to evaluate the distribution of p. In the Bayesian approach, p is a function of a random variable Λ, and therefore also is random. (*Break of notation convention*: Lowercase p is used here for denoting a random variable in order to avoid confusion with the probability function P).
In order to specify the posterior distribution of p we fits find the posterior distribution of Λ. Assuming that the interoccurence time is exponentially distributed one can show that earthquakes, Y, occur in time as a Poisson process with a given parameter $\Lambda = \lambda$, that is

$$P(Y = y|\Lambda = \lambda) = e^{-\lambda s}\frac{(\lambda s)^y}{y!},$$

which gives the probability of observing y earthquakes during time $[0, s]$. Generally, the times between events in the Poisson process are i.i.d.exponential random variables; see e.g. CB Example 3.3.1 (Gamma-Poisson relationship) or the book by Rice, Chapter 2, p.50.

Now, given the observed ineteroccurence times $\mathbf{x} = (x_1, \ldots, x_{62})$ (i.e 63 earthquakes) and using Bayes' rule we can establish the updating formula

$$\pi(\lambda|Y = 63) = const \cdot P(Y = 63|\Lambda = \lambda)\pi(\lambda)$$

where *const* does not depend on λ and $\pi(\lambda)$ is a prior distribution of Λ. Observe that the posterior distribution $\pi(\lambda|Y = y)$ is a conditional distribution, conditional upon observing 63 earthquakes during the time period.

The updating formula requires a prior distribution $\pi(\lambda)$ which can be considered as the experimenter's belief/knowledge about Λ, and is formulated before the data are seen. If we know very little about Λ, it would be sensible to choose

$$\pi(\lambda) = \lambda^{-1}$$

for $\lambda > 0$, which which would suggest that a value for an intensity becomes less likely"in inverse proportion to its value. This type of distribution is called for *non-informative* prior. Observe that the posterior probabilities will still sum (or integrate) to 1 even if the prior values do not, and so the priors only need to be specified in the correct proportion.
Now, to specify the posterior density of Λ we rearrange $P(Y = 63|\Lambda = \lambda)$. We recall that $s = 27120$, $y = 63$ and $Y \sim Po(\lambda)$ which gives

$$P(Y = 63|\Lambda = \lambda) = e^{-\lambda \cdot 27120}\frac{(\lambda \cdot 27120)^{63}}{63!} = const \cdot e^{-\lambda \cdot 27120}\lambda^{63}.$$

Hence, the posterior density of Λ is

$$\pi(\lambda|Y = 63) = const \cdot e^{-\lambda \cdot 27120}\lambda^{63} \cdot \lambda^{-1} \propto \lambda^{63-1}e^{-\lambda \cdot 27120},$$

which is a kernel of a gamma density with parameters $\alpha = 63$ and $\beta = 1/27120$. To see the graph we do

```
>> figure
>> x=[0:1/10000:0.005];
>> plot(x,gampdf(x,63,1/27120))
```

Observe that the symmetric shape of the density suggests that the posterior distribution of Λ can be approximated by a normal distribution. We recall that for $X \sim gamma(\alpha, \beta)$ $E[X] = \alpha\beta$ and $Var[X] = \alpha\beta^2$, see CB, p.624 and plot the approximate density in the same figure by

```
>>hold on
>> pi_post_lambda=normpdf(x,63/27120,sqrt(63/(27120^2)));
>> plot(x,pi_post_lambda,'r')
```

What do you observe?

Answer:

Recall that our initial goal was to specify the posterior distribution of $p = e^{-\Lambda \cdot 1200}$. By taking the logarithm and assuming that Λ is approximately normally distributed one can conclude that p is approximately log-normal, i.e.

$$log(p) \sim \mathcal{N}(-1200\alpha\beta, 1200^2\alpha\beta^2).$$

To plot the posterior density of p we do

```
>> figure
>> x=[0.01:1/1000:0.15];
>> pi_post_p=lognpdf(x,-1200*(63/27120), 1200*sqrt(63/(27120^2)));
>> plot(x, pi_post_p)
```

Compare this density with that one which you obtain in the previous subsection using the bootstrap resampling. Explain your results.

Answer: