Datorlab 1

Fördelningar och slumpvariabler.

(Understanding Distributions)

In this computer exercise we will encounter some fundamental concepts, firstly, from probability theory: the probability density function, expectation, and variance of a random variable; and, secondly, from statistics: the histogram, the empirical distribution, and probability papers. The Weibull distribution and the Gumbel distribution, both often used in reliability/safety analysis and in studies of environmental hazards, will serve as examples. At first we will rely on simulations, but eventually we will investigate real-world data: measurements of wave heights from the Atlantic Ocean.
You need a copy of Matlab that includes the Statistical Toolbox. Most versions of Matlab (including our in the lab) have this toolbox included. On some occasion you will need access to data files including samples that you will be asked to analyze. All necessary files are downloadable from the course home page
http://www.math.chalmers.se/Stat/Grundutb/CTH/ess011/1415/files/labfiles.zip.
Please download the labfiles.zip file and uncompress it at the directory you plan to use for the computer exercises.

1 Preparatory exercises

1. Make sure you understand what probability and density functions are and how they are related to the distribution function.

2. Given a sample $\{x_1, .., x_n\}$ from a random variable (r.v.) X, how do you construct the empirical distribution function? What is the empirical distribution function?

3. Explain what is meant by the α-quantile of a distribution.

Question 1: Write down the definitions of expectation and variance of a continuous random variable X, i.e. $E(X)$ and $V(X)$. Derive the expectation and variance of X if X is exponentially distributed.

2 Relative frequencies and distributions

In this section, we will use numerical examples in Matlab to approach the concepts "probability" and "distribution". The aim is that you should obtain an intuitive feeling for probabilistic reasoning, rather than to be immediately confronted with theory.

Exploring data

For illustrational purposes, we will use artificial data, which are simulated from a statistical distribution. This is opposite to real-world data, where no labels, explaining the statistical distributions, are found. However, although we, statistically speaking, know the origin of the data, this approach is useful from a pedagogical point of view. Even in research on statistical-computation algorithms, simulated data are often used for analysis and testing.

To obtain a random data-set of 50 values, type

```
>> data=randn(1,50);
```

Question 2: What is the distribution of your random sample (use `help randn`)? Write down the density function.

A good rule, whenever a new set of data is encountered: try to plot it in some kind of diagram! Use the plot command: `plot(data,'.-')`. Another way of presenting the data is to plot the sorted data: `plot(sort(data),'.-')`. From the data set got above, choose a relatively high number, say, $x = 1.1$. It may be interesting to calculate the percentage of data which have values *less than or equal to* this number. When the number of observations in the sample increases, we may interpret the ratio as the *probability* to obtain a value less than x. The ratio is calculated as follows:

```
>> x=1.1; ratio=sum(data<=x)/length(data)
```

(See that you understand the commands!)

Question 3: Try three other values of `x` and write down the answer. How do you expect the percentage to change with the change of x?

The opposite procedure, that is, find the value x corresponding to a given probability, is often more important. This is referred to as finding the *quantiles*. We will return to this later.

We can of course let the computer choose a large number of values `x` to examine, and then try get an overview. This is implemented in a home-written function `empcdf`. The function delivers two vectors: `x` contains the values chosen, while the ratios are collected in `ratio`. (If you want to see the code, try `type empcdf`.) The result is visualised in a new figure:

```
>> cdfplot(data);
>> figure(2);
>> x=sort(data); ratio=(1:length(data))/length(data);
>> plot(x,ratio,'.')
>> grid on
```

The figure should look similar to Figure 1 in this paper. It shows in some sense how the values in data are distributed, and the resulting function is called the *empirical distribution function*[1]. For a value on the abscissa, say, 1.1, we find the percentage of values in the sample with values less than this number.

Another way to plot the empirical distribution function is to make use of the command `stairs`:

```
>> n=length(data);
>> figure(3);
>> stairs(sort(data),(1:n)/n)
>> grid on
```

[1]Distribution functions are often called *cumulative distribution functions*; that is why our home-written routine is called `empcdf`, `emp`irical `c`umulative `d`istribution `f`unction.

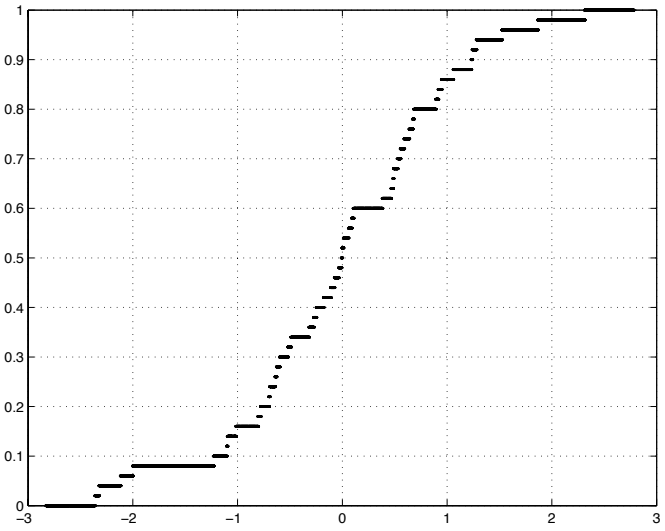

Figure 1: Empirical distribution function, an example.

Larger samples. The distribution function of a random variable

Let us now study another sample of 50 observations, obeying the same random law as the previous data. We simulate data, and plot them in the same figure as before:

```
>> data=randn(1,50);
>> figure(2);
>> cdfplot(data);
```

Continue with a larger set of data, say, 2000 observations. Analyse them as before:

```
>> hold on
>> data=randn(1,2000);
>> cdfplot(data);
```

With a large number of observations, the result approaches the *distribution function*, that is, for a random variable X, the function

$$F_X(x) = \mathsf{P}(X \leq x). \tag{1}$$

In our case, X was chosen from a Gaussian distribution (normal distribution); we had that $X \in \mathrm{N}(0;1)$. It is instructive to plot the theoretical distribution function, implemented in `normcdf`, in the same figure as before:

```
>> plot(x,normcdf(x),'r')
>> hold off
```

For every distribution function F_X, we have that $F_X(x) \to 1$ when $x \to \infty$ and that $F_X(x) \to 0$ when $x \to -\infty$.

Question 4: Estimate the median. Is it easy to estimate the *mean* of the distribution from the plot?

Quantiles

The concept of *quantile* (or *fractile*), mentioned before, is important. The quantile can be defined in different ways – when studying tables &c., you should always make sure of which definition is used. We *here* define the quantile as a number x_α which satisfies

$$\mathsf{P}(X \leq x_\alpha) = 1 - \alpha \tag{2}$$

where α is some small number (common choices: 0.05, 0.01, 0.001). The quantile is not always unique; for some values of α there might be infinitely many x_α satisfying (2); for other values of α there might be no quantile x_α at all.

Question 5: From your Matlab-plot (`figure(2)`), using (1) and the definition of quantile, can you estimate the quantile $x_{0,05}$ when $\alpha = 0,05$? Write the estimate down. Compare with the exact value, given by `norminv(1-0.05)` (alo write it down).

Other distributions

Some common choices of distribution functions have their own names. They are not only just functions in a mathematical sense, but have also been found suitable when modelling random phenomena in science and technology. The distributions are listed in almost any basic text-book in mathematical statistics.

Some of the distributions are implemented in the Statistics Toolbox. You have already encountered the distribution function when $X \in \mathrm{N}(0;1)$; it is often denoted by Φ:

$$F_X(x) = \Phi(x) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

An easy way to obtain new distributions is scaling random variables or adding constants to them: Suppose that the distribution function $F_X(x)$ of some stochastic variable X is known. If a new stochastic variable Y is defined as $Y = aX + b$, where $a > 0$ and b are constants, we can perform the following calculation

$$F_Y(y) = \mathsf{P}(Y \leq y) = \mathsf{P}(aX + b \leq y) = \mathsf{P}(X \leq (y-b)/a) = F_X((y-b)/a)$$

to obtain the distribution function for Y.

Question 6: What is the distribution function F_Y of Y if $a < 0$? Write it down in the terms of F_X. What is the distribution function F_Y of Y if $a = 0$? Sketch a plot of F_Y in the latter case.

This transformation is governed by two *parameters* a and b; other distributions or transformations of distributions might be governed by other sets of parameters. When analysing real-world data, one often knows from experience which type of distribution is suitable to describe the data. What remains is then to try to *estimate* the parameters out of data.

The normal distribution, for instance, is characterised by two parameters, m and σ^2 (or σ): if X is standard-normal, i.e. $X \in \mathrm{N}(0;1)$, then for $Y = \sigma X + m$ we have that $Y \in \mathrm{N}(m;\sigma^2)$, where m and σ^2 (or σ) are the parameters.

The Gumbel distribution and the Weibull distribution

Two important distributions, with which we will meet up again in the course, are the Gumbel distribution (also called type I extreme value distribution, or double exponential distribution) and the Weibull

distribution. A Gumbel distributed random variable X has the distribution function

$$F_X(x) = \exp(-\mathrm{e}^{-(x-b)/a}), \quad -\infty < x < \infty.$$

Here, b is a location parameter and $a > 0$ is a scaling parameter. If X belongs to a Weibull distribution, we have

$$F_X(x) = 1 - \exp(-((x-b)/a)^k), \quad x \geq b \tag{3}$$

where k is a shape parameter, b is a location parameter and $a > 0$ is a scaling parameter.

Question 7: Choose some values of the parameters and plot the function $1 - \exp(-((x-b)/a)^k)$. How does $1 - \exp(-((x-b)/a)^k)$ behave when $x < b$? Does it explain restriction $x \geq b$, and why?

Let us make some plots of these distribution functions. Try the following two cases of a Gumbel distribution:

```
>> help evcdf
>> figure
>> x=-4:0.05:6;
>> a=2; b=0; F1=evcdf(x,b,a);
>> a=1; b=1; F2=evcdf(x,b,a);
>> plot(x,F1,'b',x,F2,'r')
```

Two examples of a Weibull distribution are drawn by typing

```
>> help wblcdf
>> figure
>> x=linspace(0,6,200);
>> a=1; k=1; F1=wblcdf(x,a,k);
>> a=2.3; k=1.8; F2=wblcdf(x,a,k);
>> plot(x,F1,'b',x,F2,'r')
```

Note that the Matlab routine `wblcdf` models only the case when $b = 0$, cf (3).

3 Expectation and variance of a random variable

For a random variable X, the *expectation*, sometimes called the *mean* and denoted $\mathsf{E}(X)$, gives the value of X "on average"; if the distribution of X had been the *mass* distribution of a physical thing, the expectation would have located the centre of gravity of that thing. The *variance* $\mathsf{V}(X)$ (or, rather, the *standard deviation* $\mathsf{D}(X) = \sqrt{\mathsf{V}(X)}$) of X can be regarded as a measure of the distribution's dispersion. For a set of important distributions, $\mathsf{E}(X)$ and $\mathsf{V}(X)$ have been explicitely derived (in terms of the distribution's parameters) and tabulated, see for example the textbook.

For a given data set $x_1, \ldots, x_n$ (sample), in most cases we do not know the distribution from which the sample is taken, and hence not the mean and variance of that distribution. The sample mean, often denoted $\bar{x} = (\sum_{i=1}^n x_i)/n$, and the sample variance, often denoted $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$, are then the corresponding measures of location and dispersion. If the number n of observations increases, we may expect that these quantities become closer to $\mathsf{E}(X)$ and $\mathsf{V}(X)$ respectively. Let us examine this in Matlab by means of simulated data, the distribution of which we can control:

Consider the Weibull distribution,

$$F_X(x) = 1 - \exp(-((x-b)/a)^c), \quad x \geq b.$$

The mean and variance are given by

$$\begin{aligned} \mathsf{E}(X) &= b + a\,\Gamma(1+\frac{1}{c}), \\ \mathsf{V}(X) &= a^2\,\Gamma(1+\frac{2}{c}) - a^2\left(\Gamma(1+\frac{1}{c})\right)^2, \end{aligned}$$

where

$$\Gamma(p) = \int_0^\infty x^{p-1}\mathrm{e}^{-x}\,\mathrm{d}x. \tag{4}$$

is the gamma function. Choose for example $a = 1,5$, $b = 0$, and $c = 2$. To calculate expectation and variance, one needs the gamma function in (4) which is implemented in Matlab as `gamma`; hence

```
>> a=1.5; b=0; c=2;
>> EX=b+a*gamma(1+1/c)
>> VX=a^2*gamma(1+2/c)-a^2*(gamma(1+1/c))^2;
>> DX=sqrt(VX)
```

Now, simulate a sample of 50 observations and find the sample mean and standard deviation by the commands `mean` and `std` respectively:

```
>> x=wblrnd(a,c,1,50);
>> mean(x), std(x)
```

Question 8: Compare the values estimated from the samples with the theoretical values `EX`, `DX` that you have also obtained above. Write down the values for $\mathsf{E}(X)$, $\bar{x}$, $\mathsf{D}(X)$, $d(x)$. Are the theoretical and empirical values consistent with each other? Simulate larger samples of, say, 200, 1000, and 5000 observations respectively. What happens when the number of observations increases?

4 Probability plots

Assume that we have a set of observations x_1, x_2, ..., x_n. Before we estimate any parameters, we must convince ourselves that the observations originate from the right *family* of distributions, e.g. normal, Gumbel, or Weibull. One way to get a rough idea of which family of distributions may be suitable, is to display the observations in a *probability plot*[2]: If you suspect that the data originate from, for instance, a normal distribution, then you should make a *normal probability plot*; if you instead suspect a Gumbel distribution, then make a *Gumbel probability plot*. If, in the plot, the observations seem to line up well along a straight line, it indicates that the chosen distribution for the probability plot indeed might serve as a good model for the observations. Statistics Toolbox provides `normplot` (for normal distribution), `wblplot` (for Weibull distribution); but unfortunately there is no probability plot for Gumbel distribution, so we have created one and named it `gumbplot` (available in `labfiles`). Acquaint yourself with the above-mentioned commands, for example

[2]Before the computer age, the observations were plotted manually into diagram-forms printed on sheets of paper; therefore we now and then will use the expression "to plot data in a certain probability paper" even if we are referring to computer-displayed diagrams.

```
>>  dat1=randn(2000,1);  % Attention: Normal distribution!
>>  normplot(dat1)
>>  wblplot(dat1)
>>  dat2=rand(3000,1);  % Attention: Uniform distribution!
>>  normplot(dat2)
>>  gumbplot(dat2)
>>  dat3=wblrnd(2,2.3,1,3000); % Attention: Weibull distribution!
>>  wblplot(dat3)
>>  gumbplot(dat3)   % Attention: Gumbel distribution!
>>  dat4=gumbrnd(1,2,1,3500);   % Available in labfiles
>>  gumbplot(dat4)
```

Experiment more with the number of observations; change also distributions!

Question 9: What happens when you plot the data in the "wrong" distribution plot?

Measurements of significant wave heights in the Atlantic Ocean

In the field of oceanography and marine technology, statistical extreme-value theory has been used to a great extent. In design of offshore structures knowledge about "extreme" conditions is important.

In the numerical examples above, we used artificial data, simulated from a distribution which we could control. We will now consider *real* measurements from the Atlantic Ocean. The data set contains so-called significant wave heights (in meters), that is, the average of the highest one-third of the waves.

Now, load the data set `atlantic.dat` and read about the measurements; then find the size of data, and plot it:

```
>> atl=load('atlantic.dat');
>> help atlantic
>> size(atl)
>> plot(atl,'.')
```

One knows that, roughly speaking, the registered so-called significant wave-heights behave, statistically, as if they were maximum wave-heights; therefore one can suspect them to originate from a Gumbel distribution, for instance. Below we will make different probability plots.

```
>> normplot(atl)
>> normplot(log(atl))
>> gumbplot(atl)
>> wblplot(atl)
```

Question 10: Which distribution might be a satisfactory choice?