Distributions, simulations and model fit

Computer lab 1

Write your name here

2024-03-26

The purpose with this lab is for you to - practice on the concepts density function and distribution function - practice to draw random numbers from a distribution - practice to use different graphical methods to evaluate possible distributions that your observations can come from

Table of contents

Data material soil samples

In the forest area ASA test park in Småland, 94 different pits are dug in the ground, and from each pit soil samples are taken where, among other things, aluminum content and calcium content are measured (mg/g). The data is in the file `data/lab1_files/jordprov.Rdata` which contains the two variables `al` and `ca`.

Open the file in RStudio, for example by clicking on it under Files. You can see the measured concentrations by clicking on the icon furthest to the right on the `jordprov` row. Alternatively, you can type `View(jordprov)` in RStudio's console window. To access the variable `al` in the dataset `jordprov`, you would write `jordprov$al`.

```r
load("data/lab1_filer/jordprov.Rdata")
# write your R-code here
```

1 Överblick av datamaterialet

Firstly, you would want to calculate some summary statistics for the data (mean, min, max, standard deviation, etc.) and visualize some overview figures. The command `mean(jordprov$al)` gives you the mean of the aluminum measurements.

Task 1.1

Perform an overview analysis of the aluminum and calcium levels using the following functions in R: `sd()`, `summary()`, `hist()`, `boxplot()`, `plot()`. Also, try `plot(soil_samples$al, soil_samples$ca)`.
When dealing with measurements, $x_1, x_2, \ldots, x_n$, a lot of information can be obtained by plotting the so-called empirical cumulative distribution function (ECDF). The data points, x_i, are sorted from smallest to largest. The proportion of data points less than or equal to x_i is then plotted against x_i. It forms a growing step function that takes a step of height $1/n$ for each data point. In R, you can plot this function using the command `plot.ecdf()`.

```r
# write your R-code here
```

Task 1.2

Plot the empirical distribution function for the aluminum levels. A grid is added to the figure if you write `plot.ecdf(jordprov$al, panel.first=grid())`. Use the figure to find out the proportion of measurements that are below 80 mg/g.

Answer:

Task 1.3

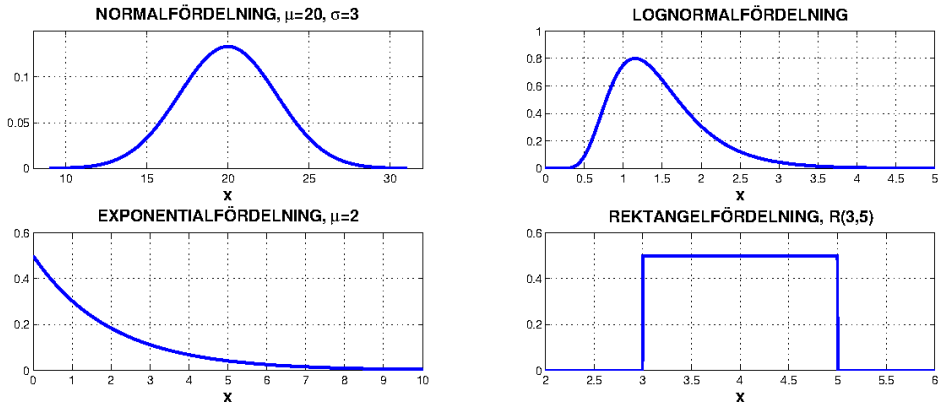
What aluminium is exceeded in 70% of measurements?

Answer:

Does a standard distribution fit to my data?

Now we want, using graphical methods, to investigate if the two datasets can be modeled with some standard distributions. Some continuous standard distributions that we encounter

in the course are the normal distribution, lognormal distribution, exponential distribution, and uniform distribution.



When looking at the histogram for aluminum levels, it doesn't seem unreasonable that they could be normal distributed, but a histogram is often a blunt instrument when trying to fit a standard distribution to data. A more useful method is to plot the data on a so-called probability paper or QQ plot. To illustrate the method, we are helped by seeing how it works on samples where we really know the distribution, so we need to know how to generate random numbers from different distributions.

2 Simulate random variables in R

In R there are functions for drawing numbers from standard distributions. You find some examples below.

Distribution	Function	Example
Normal	`rnorm(antal,mean,stddev)`	`rnorm(20,3,1)`
Exponential	`rexp(antal,1/mean)`	`rexp(50,0.5)`
Uniform	`runif(antal,min,max)`	`runif(30,-2,5)`
Binomial	`rbinom(antal,n,p)`	`rbinom(10,5,0.2)`
Poisson	`rpois(antal,mean)`	`rpois(25,4)`

If you, for example, want to simulate 100 random numbers from a normal distribution with a mean of 20 and a standard deviation of 3 and store them in the variable `norm1`, you would do it like this: `norm1 <- rnorm(100, 20, 3)`.

Task 2.1

Plot the 100 normal distributed random numbers in `norm1` as a histogram. Does the histogram resemble the theoretical normal distribution in the figure above? What happens if you instead create a histogram of 1000 simulated normal distributed random

numbers?

Answer:

Task 2.2

Generate 1000 random numbers from an exponential distribution with a mean of 2 and store them in the variable `exp1`. Create a histogram and compare it with the density function of the exponential distribution plotted in the figure above.

Task 2.3

Do the same for a uniform distribution $R(3, 5)$, i.e., draw 1000 random numbers (call them `unif1`), visualize them with a histogram, and compare with the theoretical density distribution.

```
# write your R-code here
```

QQ-plot

Often, there is the question of whether data in a sample can be modeled with a theoretical standard distribution. This was the case, for example, with the aluminum levels in the soil that you studied earlier. A graphical method when attempting to fit your data to a distribution is to use a so-called QQ-plot, where Q stands for quantile. The values in the dataset (sample) are compared with those expected from a certain theoretical distribution. If the data matches the expected values, the points in a QQ-plot will lie along a straight line. Conversely, if the QQ-plot shows significant deviation from a straight line, the distribution we tested does not fit our data well. To test if a sample might come from a normal distribution, the command is `qqnorm(name of the sample)`.

Task 2.4

Try how the QQ-plot looks like when you fit the sample with 100 normal distributed random numbers (`norm1`) to a normal distribution using the command `qqnorm(norm1)`. What happens when you also write `qqline(norm1)`?

Answer:

Since we know that `norm1` contains normal distributed random numbers, the fit should naturally be good. However, it should be noted that one cannot expect a perfect straight line in the plot, as we are dealing with random numbers. A slight deviation at both ends of the line is not uncommon.

Task 2.5

What does it look like if we fit exponentially distributed random numbers `exp1` or uniformly distributed random numbers `unif1` to a normal distribution using the commands `qqnorm(exp1)` and `qqnorm(unif1)`, respectively?

Answer:

3 Model for aluminum levels

We study the aluminum levels in the 94 samples again.

Task 3.1

Decide if the levels can be modelled by a normal distribution. Use the method QQ-plot with the command `qqnorm(jordprov$al)`. Does it seem to be a good fit?

Answer:

Previously, you used R to calculate the mean (m) and standard deviation (s) for the aluminum levels. You can use these values as estimates of the expected value μ and the standard deviation σ in the fitted distribution. Given that you believe a normal distribution fits the data well, *we can now set up a model for our observations.*

Model: X = "aluminum level in a soil sample"; X is normal distribution with expected value μ and standard deviation σ. We estimate these parameters as $\hat{\mu} = m$ ("sample mean") and $\hat{\sigma} = s$ ("standard deviation for the sample").

Task 3.2

Calculate the estimated parameters and insert them into your model by writing

```
m = mean(jordprov\$al)
s = sd(jordprov$al)
hist(jordprov$al,probability=TRUE)
xx <- seq(40,120,by=0.2)
lines(xx,dnorm(xx,m,s))
```

Now you can compare the density function from the model with the histogram of the sample. They should be quite similar for the model to be considered good.
With this model, you can calculate probabilities and make predictions about future measurements. For example, you can calculate the probability that a new aluminum measurement will exceed 80 mg/g, or determine the aluminum level that will be exceeded by 10% of future measurements.

Task 3.3

Suppose you want to calculate the probability that the aluminum level in a sample exceeds 80 mg/g, in other words, you want to calculate $P(X > 80)$. In exercises, you have done this using calculators and/or tables. In R, the command for calculating the cumulative distribution function of the normal distribution $P(X \leq x)$ is `pnorm()`. Explain why the desired probability is calculated using the command `1 - pnorm(80, m, s)`.

Answer:

Task 3.4

The aluminum level that will be exceeded by 10% of future measurements is a quantile in the normal distribution. Use R to calculate the quantile with the command `qnorm(0.1,m,s)`.

Answer:

4 Model for calcium levels

Task 4.1

Try using a QQ-plot to see if the calcium levels can also be modeled with a normal distribution. Does it look good?

Answer:

Another standard distribution commonly used for bio data is the lognormal distribution. Measurements can be modeled with a lognormal distribution if the **logarithms** of the measurements fit well with a normal distribution. This means that no special QQ-plot is needed for this distribution, one can use `qqnorm(log(stickprovsnamn))`.

Task 4.2

Investigate if the calcium levels can be modeled with a lognormal distribution. Does it look good?

Answer:

Now you need to estimate the probability that a calcium measurement exceeds 30 mg/g. First, you need to set up a model for calcium measurements. Assume that calcium values follow a lognormal distribution. Then, use the model to make calculations.

Task 4.3

If Y = calcium level, we are looking for $P(Y > 30)$, which is equivalent to $P(\log(Y) > \log(30))$. If Y follows a lognormal distribution, then $\log(Y)$ follows a normal distribution. Set up a model for $\log(Y)$ and estimate its parameters. What distribution do the estimated parameters have?

Answer:

Task 4.4

Calculate the probability that `log(calcium level)` exceeds log(30). Provide the command you use to perform this calculation."

Answer: