

# Fundamentals of Statistical Modeling (VT21)

Andrea Discacciati  
Karolinska Institutet  
Stockholm, Sweden

## Lab 2

Load the dataset and the `mlci` command

{{1}}

Install the `rcsngen` command (you need to be connected to the Internet)

{{2}}

### Exercise 1

This dataset contains information on the blood concentration of a biomarker ( $y$ ) in a random sample of 1432 subjects. Take a look at the histogram. What can we say about the distribution of this biomarker?

Plot also the histogram of  $\log(y)$ . How does the distribution of the biomarker after logarithmic transform look like?

{{3}}

{{4}}

### Exercise 2

We assume that  $f(y)$  is gamma (see Lab 1). Estimate the parameters  $\alpha$  and  $\beta$  using the `gammaden()` function. Fix the location parameter  $g$  (the third argument of the `gammaden()` function) to be equal to 0. Constrain  $\alpha$  and  $\beta$  to be positive.

Note: the parameters  $\alpha$  and  $\beta$  are not interpretable. We can reparametrise the gamma distribution so that one parameter is equal to its mean. This is described in the Extra material for Lab 2.

{{5}}

Plot the estimated density  $\hat{f}(y)$  over the sample histogram

{{6}}

### Exercise 3

We assume that  $f(y)$  is log-normal distributed. That is, we assume that the biomarker is standard normal distributed after we apply the transform

$$G(y) = (\log(y) - \mu)/\sigma$$

The derivative of  $G(y)$  with respect to  $y$  is

$$G'(y) = g(y) = 1/(y\sigma).$$

Estimate the parameters  $\mu$  and  $\sigma$ . Constrain  $\sigma$  to be positive.

{{7}}

Compare the likelihood with that from the gamma model

Plot the estimated density  $\hat{f}(y)$  over the sample histogram

{{8}}

#### Exercise 4

We make the transform  $G(y)$  more flexible using polynomials. Consider the transform

$$G(y) = (\log(y) + \eta \log(y)^2 - \mu) / \sigma$$

The derivative of  $G(y)$  with respect to  $y$  is

$$G'(y) = g(y) = (1 + 2\eta \log(y)) / (\sigma y)$$

Estimate the parameters  $\mu, \sigma, \eta$ . Constrain  $\sigma$  to be positive.

{{9}}

Plot the estimated density  $\hat{f}(y)$  over the sample histogram

{{10}}

#### Exercise 5

Instead of a quadratic term, we add two restricted cubic splines transforms of  $\log(y)$ :  $V_2(\log(y))$  and  $V_3(\log(y))$ . We consider the transform

$$G(y) = (\log(y) + \eta_1 V_2(\log(y)) + \eta_2 V_3(\log(y)) - \mu) / \sigma$$

The derivative of  $G(y)$  with respect to  $y$  is

$$G'(y) = g(y) = (1 + \eta_1 v_2(\log(y)) + \eta_2 v_3(\log(y))) / (\sigma y)$$

Estimate the parameters  $\mu, \sigma, \eta_1, \eta_2$ . Constrain  $\sigma$  to be positive. Jointly test the 2 parameters  $\eta_1, \eta_2$  to assess whether adding the 2 RCS transforms improves the fit of this model with respect to the “basic” log-normal model (see Exercise 3).

{{11}}

Plot the estimated density  $\hat{f}(y)$  over the sample histogram

{{12}}

#### Exercise 6

Let's assess the goodness-of-fit of the log-normal model with RCS transforms (see Exercise 5) and of the log-normal model (see Exercise 3) using a quantile plot.

{{13}}

## Extra: Exercise 7

We now assume that  $f(y)$  is gamma-distributed after square root transform.

$$G(y) = \sqrt{y}$$

The derivative is

$$G'(y) = g(y) = 0.5/\sqrt{y}$$

Estimate the parameters  $\alpha$  and  $\beta$  using the `gammaden()` function. Fix the location parameter  $g$  to be equal to 0. Constrain  $\alpha$  and  $\beta$  to be positive. Compare the likelihood with that from the log-normal and gamma models

{{14}}

Plot the estimated density  $\hat{f}(y)$  over the sample histogram. Visually compare the estimated density from the lognormal + splines model with the density from the gamma model after square root transform. What do you conclude?

{{15}}