

# Fundamentals of Statistical Modeling (VT20)

Andrea Discacciati  
Karolinska Institutet  
Stockholm, Sweden

## Lab 3

Load the dataset and the `mlci` command

{{1}}

### Exercise 1

I retrieved data on age at death among females in Switzerland in 2016 from <http://www.mortality.org> (variable `age`) ( $n = 33,638$ ). There are no censored observations (we know the age at death for all individuals). Plot an histogram of age at death. What can we say about the distribution?

{{2}}

Assume that  $f(\text{age})$  follows a generalized extreme values distribution. Estimate the parameters  $\mu$  and  $\sigma$ . Constrain  $\sigma$  to be positive.

Remember: we're assuming that the variable `age` is Standard-Exponential-distributed after we apply the transform  $G(y)$ . The pdf of a Standard Exponential distribution is  $f_{SE}(u) = \exp(-u)$ .

{{3}}

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

{{4}}

### Exercise 2

Inflate the probability of death during the first year of life ( $\text{age} < 1$ ), while constraining it to be between 0 and 1. How do you interpret the coefficient  $\eta$ ?

Note: we can probably improve the fit of this model by making it more flexible, for example using restricted cubic splines. This is described in the Extra material for Lab 3.

{{5}}

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

{{6}}

### Exercise 3

Assume now that all ages above 100 years were recorded as 100 years (those ages are right-censored at 100 years) (variable `age100`).

Plot an histogram of age at death. Note the spike at  $\text{age} = 100$  due to the censored observations.

{{7}}

Assume that  $f(\text{age})$  follows a generalized extreme values distribution. Estimate the parameters  $\eta$ ,  $\mu$  and  $\sigma$ . Constrain  $\eta$  to be between 0 and 1. Constrain  $\sigma$  to be positive. Take into account right-censoring in age-at-death. You'll need to generate an event/censoring indicator variable, first.

{{8}}

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

{{9}}

## Exercise 4

We know that age at death was actually recorded in integer years. The exact age at death is therefore unknown to us. We only know it happened between  $\lfloor \text{age} \rfloor$  and  $\lfloor \text{age} + 1 \rfloor$  years.

Estimate the parameters  $\mu$  and  $\sigma$ . Constrain  $\sigma$  to be positive. Take into account interval-censoring and right-censoring at 100 years.

{{10}}

## Exercise 5

We measured how many times a random sample of 722 subjects were admitted to the hospital in 2016. Plot an histogram of the variable  $y$ .

{{11}}

Assume that  $f(y)$  follows a Bernoulli-Poisson Mixture model. It's similar to the Bernoulli-Negative-Binomial Mixture model, but the density is:

$$(\beta + (1 - \beta) * f_{\text{Poi}}(0))^{I(y=0)} \times ((1 - \beta) * f_{\text{Poi}}(y))^{I(y>0)},$$

where  $f_{\text{Poi}}(y)$  is the pmf of a Poisson distribution ([https://en.wikipedia.org/wiki/Poisson\\_distribution](https://en.wikipedia.org/wiki/Poisson_distribution)) (see Stata's `poissonp()` function). Estimate the model's parameters. Remember to constrain the parameters to their parameter space.

{{12}}

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

{{13}}

## Exercise 6

We consider  $Y$  the interval-censored version of a latent (unobserved) variable  $Y^*$ . Assume that  $Y^*$  follows a gamma distribution. Estimate its parameters.

{{14}}

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

{{15}}

Which model seems to fit better the data? Tabulate the observed and model-based predicted proportions.

{{16}}

## Extra

Let's refit the model in Exercise 3, but this time we use the optimization function `optimize()` (which is the function that `mlexp` calls behind the curtains). `optimize()` is part of Mata, Stata's matrix programming language.

{{17}}