Statistical Theory and Modeling, 7.5 hp

Home assignment - Part 1

Mattias Villani

2025-04-03

Table of contents

Problem 1 - Exponential distribution and Numerical integration

The exponential distribution, $X \sim \text{Expon}(\beta)$ has (probability) density function

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \text{ for } x \geq 0$$

In this parameterization, the parameter β is called a **scale** parameter, and here $\mathbb{E}(X) = \beta$. This is the parameterization used in the course book.

R (and Wikipedia) instead uses the alternative parameterization with a **rate** parameter λ and the density function

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0.$$

In this parameterization $\mathbb{E}(X) = \frac{1}{\lambda}$. So, the connection between the two parameterization is that $\lambda = \frac{1}{\beta}$.

We will use the parameterization in the course book with the scale parameter β. If you want to simulate 10 random numbers from the $X \sim \text{Expon}(\beta)$ with $\beta = 2$ you have to use the command `rexp(n = 10, rate = 1/2)`, since $\beta = 2$ implies $\lambda = 1/2$. The names of the arguments can be left out so `rexp(10, 1/2)` also works (but then you have to write the arguments in that exact order).

Problem 1a)

A good way to check which parameterization is actually used in a given programming language is to simulate a large number of random numbers (also called **draws**) from the distribution and then compute the usual sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

of those random numbers. According to the **law of large numbers**, this sample mean should be close to the "theoretical"/population mean of $\mathbb{E}(X)$ in the given parameterization.

Simulate $n = 10000$ random numbers from the exponential distribution with rate $\lambda = 2$ to verify that R is indeed using the rate parameterization.

Problem 1b)

Simulate 200 draws (random numbers) from the $X \sim \text{Expon}(\beta = 2)$ distribution. Plot a histogram of the draws (use 30 histogram bins/cells) and overlay the theoretical probability density function (pdf) for the $\text{Expon}(\beta = 2)$ distribution as a curve. Note that you have to use the argument `freq=FALSE` in the `hist` function, otherwise the vertical scale will be *counts* within each bin in the histogram, and you want really want the height of the histogram bars to represent the *density*.
[*Hint*: evaluate the `dexp` density function over a fine grid of x-values to plot the pdf.]

Problem 1c)

Overlay two more pdf curves: one for $\text{Expon}(\beta = 1)$ and the other for $\text{Expon}(\beta = 3)$. Use different colors. Which of the three pdf curves fit the data (histogram) best? Why do you think that is?

Problem 1d)

The **empirical** cumulative density function (cdf) from a sample with n observations is given by

$$\hat{F}_n(x) = \frac{\text{number of elements in the sample } \leq x}{n}$$

Plot the empirical cdf for the $n = 200$ observations that you simulated in Problem 1b); see https://en.wikipedia.org/wiki/Empirical_distribution_function for a little information about the empirical cdf, if you are curious. Overlay the cdf from the three distributions above: $\text{Expon}(\beta = 1/2)$, $\text{Expon}(\beta = 1)$ and $\text{Expon}(\beta = 5)$. Which distribution seems to fit best?

Does it match with your conclusion from Problem 1c)?
[*Hint*: the `sort` function might be handy for the empirical cdf, and don't forget about the so called `p`-functions in R.]

Problem 1e)

Compare the *sample median* from the $n = 200$ observations to the theoretical medians for each of the above three distributions. Explain both how:

- a sample median is defined and
- how a median of a *statistical distribution* is defined.

[Hint: recall the so called `q`-functions in R].

Problem 1f)

Verify by numerical integration that the Expon($\beta = 2$) density in R really fulfills the required property of any density $\int_{-\infty}^{\infty} f(x)dx = 1$. This entails doing a rectangle sum approximation as in the definition of the integral in Lecture 2 (do *not* use a built-in function or a package for numerical integration). Start with a rectangle width of $\Delta x = 0.5$ and then lower it until the integral seems to have converged.

Problem 1g)

Compute the expected value of the exponential distribution with $\beta = 2$ using numerical integration, i.e. using similar technique as in Problem 1f). Verify your result from the rectangle sum by using R's built in numerical integration routine `integrate` (see `?integrate` for the documentation).
(Here we actually know the result, the expected value of Expon(β) is β, but numerical integration technique can be used for the expectation of *any* function, for example if you are interested in $\mathbb{E}(\log(X))$, when $X \sim \text{Expon}(\beta)$).
[*Hint1*: note that `integrate` requires the *function* f(x) to be integrated as input argument, so you have to define such a function before calling the integrate function. To remind you of how functions are written in R, a toy function in R is given below.]
[*Hint2*: don't forget that I asking you to compute the *expected value*, not just to integrate the density function]

```
# Just a toy function to show how functions are implemented in R.
myCubicFunction <- function(x){
  y = x^3
  return(y)
```

```
}
myCubicFunction(3)
```

```
[1] 27
```

Problem 2 - Probability models for count data

Problem 2a)

The file `bugs.csv` contains a dataset with the number of bugs and some other explanatory variables for $n = 91$ releases of several software projects. Here we will only analyze the variable `nBugs`, which we will store in a vector `y`, for simplicity. Load the data like this:

```
data = read.csv("https://github.com/StatisticsSU/STM/raw/main/assignment/bugs.csv",
                header = TRUE)
y = data$nBugs # number of bugs, a vector with n = 91 observations
```

You can ignore that some of the observations actually comes from the same project at different releases, and assume that the observations are *independent* and identically distributed. Consider first the model

$$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$$

where $n = 91$ here and $\overset{iid}{\sim}$ means that the observations are assumed independent and identically distributed (that is, each observation is assumed to come from the same Poisson distribution).

Since λ is the mean in the Poisson(λ) distribution, a reasonable *estimator* of λ is the sample mean $\bar{y}$. Plot a histogram of the data and overlay the density of Poisson distribution with $\lambda = \bar{y}$. Does this Poisson model fit the data well. If not, why?

[*Hint*: either use a histogram when plotting the data, or use `proportions(table(y))` to compute a table of proportions and then use `barplot` to plot a bar chart, which is suitable for discrete data. A histogram is easier, however.]

Problem 2b)

Let us now try to with a negative binomial model for the data. We will use the variant that counts *the number of failures* until r successes has been observed, and we will use the alternative parameterization with an explicit parameter μ for the mean. So the model is

$$Y_1, Y_2, ..., Y_n \overset{iid}{\sim} \text{NegBin}(r, \mu),$$

where each random variable Y_i can take on values in the set $\{0, 1, 2, ...\}$. Since μ is the mean, we will estimate it with the sample mean $\bar{y}$. Add the probability function from the negative binomial model for three different r values: $r = 1$, $r = 3$ and $r = 100$ (one curve for each) to the plot you did in Problem 2a). Which of these models do you prefer? Why? Which of the negative binomial models is closest to the Poisson model? Why?

[*hint*: note that R has the `dnbinom` function that can be called with the mean parameterization. For example, `dnbinom(1, size = 3, mu = 2)` give the probability $\Pr(Y = 1)$ when $Y \sim \text{NegBin}(r = 3, \mu = 2)$, so that the argument `size` is the parameter r.]

Problem 3 - Transforming random variables

Problem 3a)

This problem is to be done **after Lecture 6**.

Let $X \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$. We are now interested in the distribution of $Y = \exp(X)$. Obtain the distribution for Y by simulating 10000 draws. Plot a histogram with 100 bins.

Problem 3b)

Use the method of transformation (Section 6.4 in the course book) to show that the probability density for Y is given by

$$f(x) = \frac{1}{\sqrt{2\pi}x} \exp\Big(-\frac{1}{2}(\log(x) - \mu)^2 \Big)$$

Overlay a plot of this density in the histogram from Problem 3a).

[*hint*: you can use LaTeX to write math in Quarto file (Google it), but it is also OK to just do the math on paper, take a photo and include the photo]

Problem 3c)

Use (Monte Carlo) simulation with $m = 10000$ random draws to estimate $\mathrm{E}(Y)$, where, as before, $Y = \exp(X)$ and $X \sim \mathrm{Normal}(\mu = 0, \sigma^2 = 1)$. Check the convergence of the estimate by plotting the sequential Monte Carlo estimates for increasing Monte Carlo sample sizes of $10, 20, 30, \ldots, 9900, 10000$. Does the estimate seem to converge (settle down) to the true expectation, which happens to be $\mathrm{E}(Y) = \exp(\frac{1}{2})$? [How do I know that this is the true expected value? See this: https://en.wikipedia.org/wiki/Log-normal_distribution]