SDA Group Submission Assignment 3

CD

For your group submission, upload one .pdf in Canvas 😀

If you follow all installation instructions given in Canvas and still have big issues with R Markdown installation, you can exceptionally submit a .pdf generated in a different way (e.g., by copying code 'manually' into a text editor), which should still include all relevant explanations, code snippets and graphics.

Always remember that there are several ways to code something, and that the useful `R` functions provided for each task are just some suggestions. Have fun!

Exercise 1

Although, in reality, one typically does not know the *true* distribution from which a sample originates, here we will simulate a sample from a *known* distribution to better compare different kernel density estimators.

a. Set a seed of your choice (for reproducibility) and generate a random sample of size $n = 20$ from a t distribution with 3 degrees of freedom. Plot the following side-by-side:

   - an histogram (scaled to density) of the sample, with the kernel density estimators corresponding to different kernel choices (Gaussian, Epanechnikov, rectangular and triangular) superimposed as colorful lines. Add a legend explaining which color corresponds to which kernel choice;
   - an histogram (scaled to density) of the sample, with the kernel density estimators corresponding to different bandwidth choices (default, 0.3 and 1.5) superimposed as colorful lines. Add a legend explaining which color corresponds to which bandwidth choice.

   Hint: look into the help page `?"density"` to see how to change kernel and/or bandwidth for the estimator. As always, set plotting options such that nothing is cropped.

b. Comment on what you observe: how do different kernel choices compare to each other? How do different bandwidth choices compare to each other? Which choice (kernel or bandwidth) seems to have a bigger influence on the estimator?

c. Write a function `h_opt(x)` that, for a sample `x`, computes $\hat{h}_{opt}$ as in (the first equality of) Equation (4.3) in the Syllabus. The function should use, for the estimator $\hat{\sigma}$, the minimum between sample standard deviation and sample interquartile range divided by 1.34 (see page 45). Use the function to compute $\hat{h}_{opt}$ for the sample generated in a., and compare it with the default bandwidth used by `R` in this case.

   Hint: look into `?"density"` to see what the function returns (under `Value`) and how to extract the bandwidth used.

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The graphical output required by a.
- Your comment for b.
- Your code for the function in c. and the result of the function call

Useful `R` functions

`rt`, `par(mfrow = c(1, 2))`, `hist`, `density`, `legend`, `length`, `sd`, `IQR`, ...

Exercise 2

A total of $n = 130$ VU students (all enrolled in an unspecified statistics class) decide to record the time (in minutes) each of them spent waiting for the elevator bringing them to the 14^{th} floor of the main building (HG). These (unfortunately *fictitious*) data are collected in file `waiting_times.RDS` on Canvas, and we assume that they are realizations $x_1, \ldots, x_n$ of i.i.d. random variables with *unknown* (continuous) distribution F (and density f).

a. Given the nature of the variable we are studying, for which values of t do you expect $f(t) = 0$?

b. Using the given sample, find a suitable density estimator $t \mapsto \hat{f}(t)$, taking also into account what you established in a. Use just one of the available approaches discussed in Section 4.4 of Syllabus/in the slide set for Unit 4. Plot the histogram (scaled to density) of the data and add a colorful line corresponding to the density estimator.

   Hint: if you would like to use the log-transformation and first find a suitable kernel density estimator $\hat{f}_y$ based on the log-transformed sample $y_1 = \log(x_1), \ldots, y_n = \log(x_n)$, you can then obtain the density estimate $\hat{f}_x$ for the original sample based on[1] (where it is assumed that the density is estimated at 512 equally spaced points – this is the default, see `?"density"`):

```
y_seq <- seq(min(y), max(y), length.out = 512)
lines(exp(y_seq), (density(y, ..., from = min(y_seq), to = max(y_seq))$y)/exp(y_seq))
```

c. Waiting times are typically modeled by an exponential distribution. Using a QQ-plot (as in A2, you can use `EnvStats::qqPlot`), assess whether this is distribution could be appropriate to model the sample.

d. Note that if $x_1, \ldots, x_n$ are realizations of $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Exp}(\lambda)$, the maximum likelihood estimator of the rate parameter λ is $\hat{\lambda}_{MLE} = 1/\bar{x}$, where $\bar{x}$ is the sample mean. Use this fact to estimate the rate parameter from the data, and reproduce the plot of b. with an extra line depicting the theoretical density of an exponential distribution with rate $\hat{\lambda}_{MLE}$. How does this exponential density compare to the density estimator you found in b.?

Deliverables

- For all subtasks, to aid in correction: all relevant code
- A short explanation for a.
- The code and graphical output required by b.
- The graphical output (and related judgement) required by c.
- The graphical output (and related answer) required by d.

Useful `R` functions

`density`, `hist`, `lines`, `EnvStats::qqPlot`, `curve`, `dexp`, ...

[1] This is due to $F_y(t) = F_x(\exp(t))$ for the corresponding cumulative distribution functions and, consequentely, $f_y(t) = f_x(\exp(t)) \cdot \exp(t)$ for the corresponding densities.