Karolinska Institutet

SciLifeLab

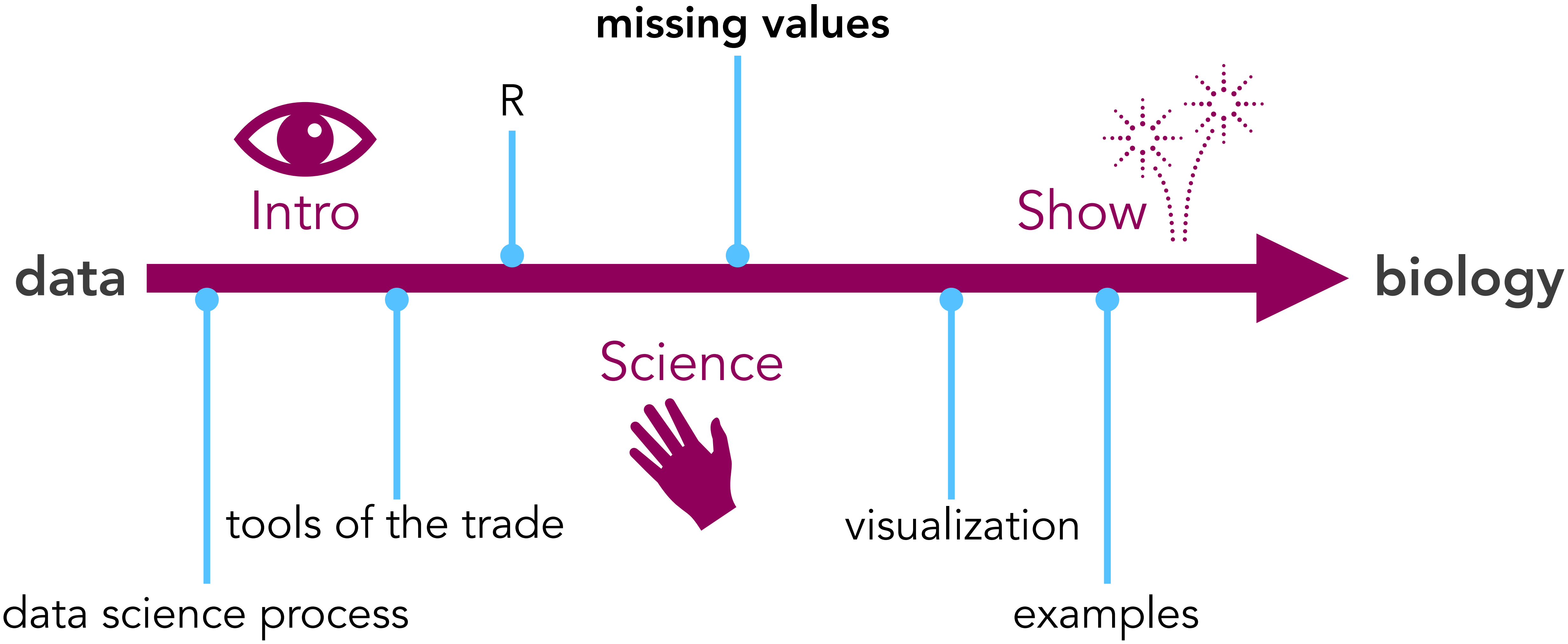
LehtiöLab

Missing value imputation

10 / 09 – 2019

Matthias Stahl

matthias.stahl@ki.se



data

data science process

Intro

tools of the trade

R

missing values

Science

visualization

Show

examples

biology

Three types of missing values

MCAR

missing completely at random

no cause to missingness

MAR

missing at random

missing values can be explained by the available data

NMAR

not missing at random

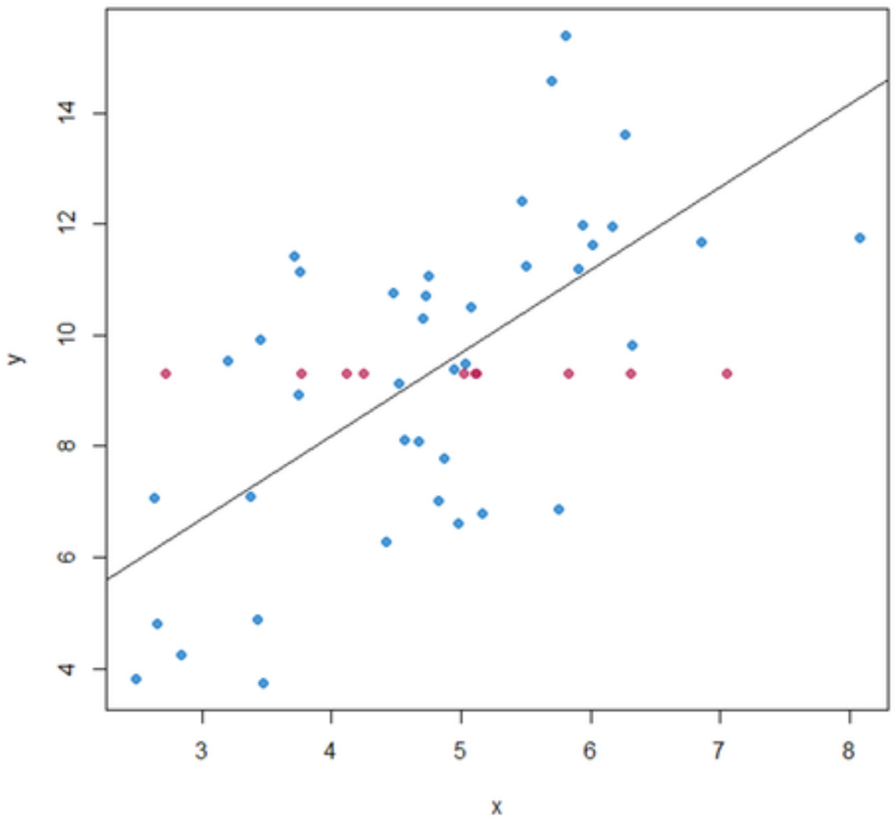
missing data can be ignored

missing data **cannot** be ignored

There are *numerous* methods to fill missing values.

Keep the sample statistics

impute with **mean** | **median** | **mode** of the data

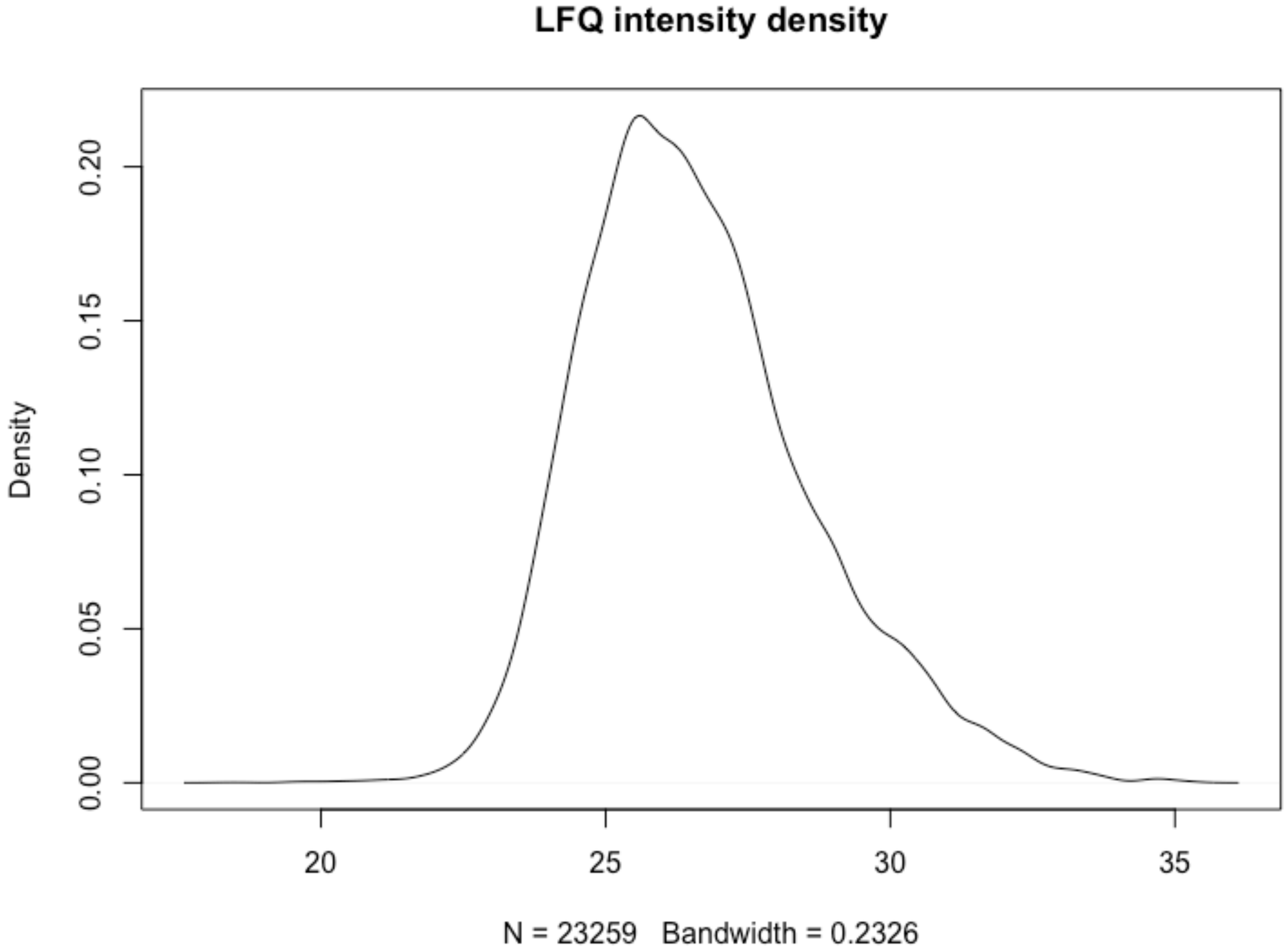


not optimal!

changes variance, introduces bias, …

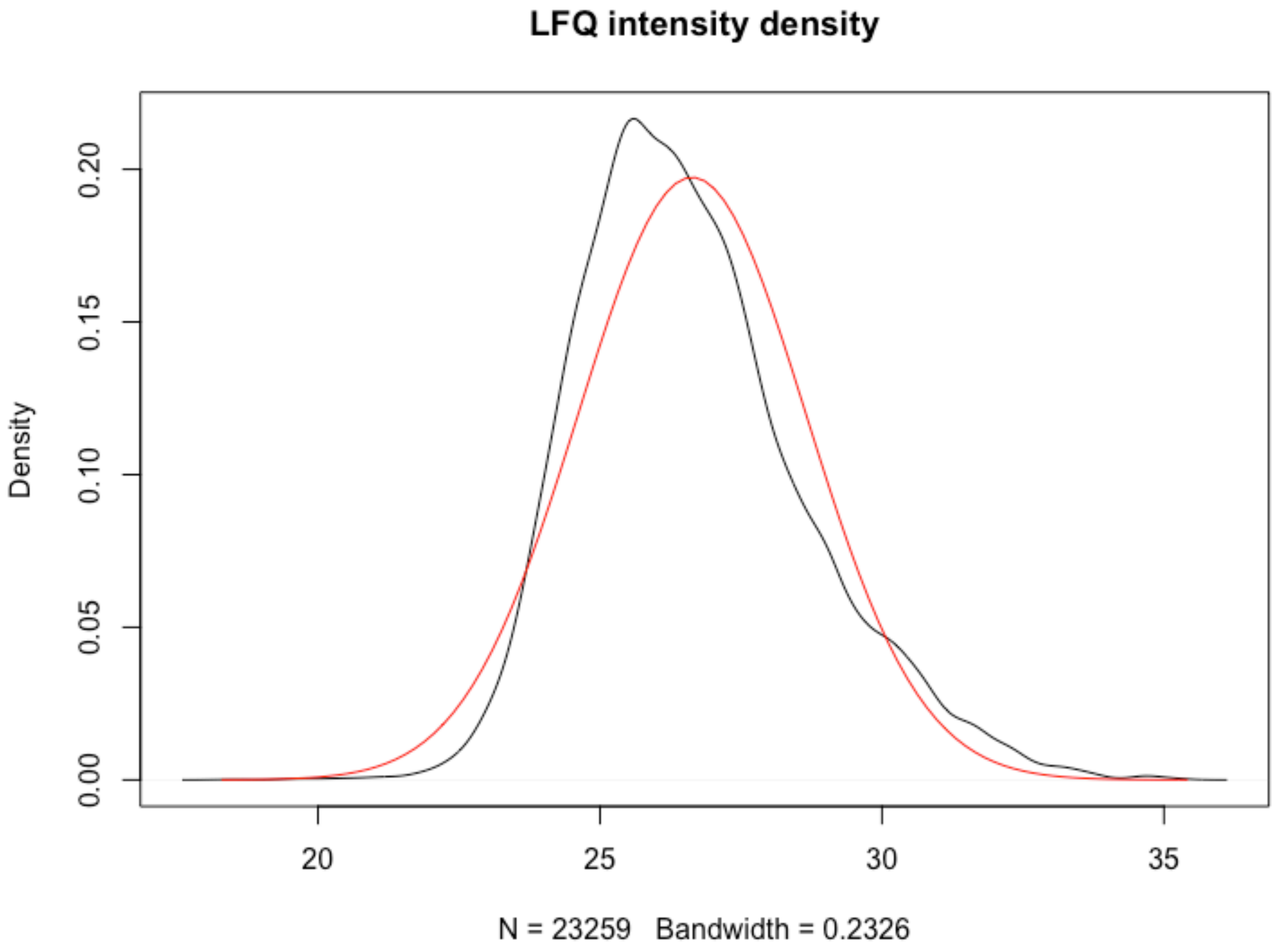
A usual dataset

Caro's dataset (HepG2 + **NC1** or DMSO)

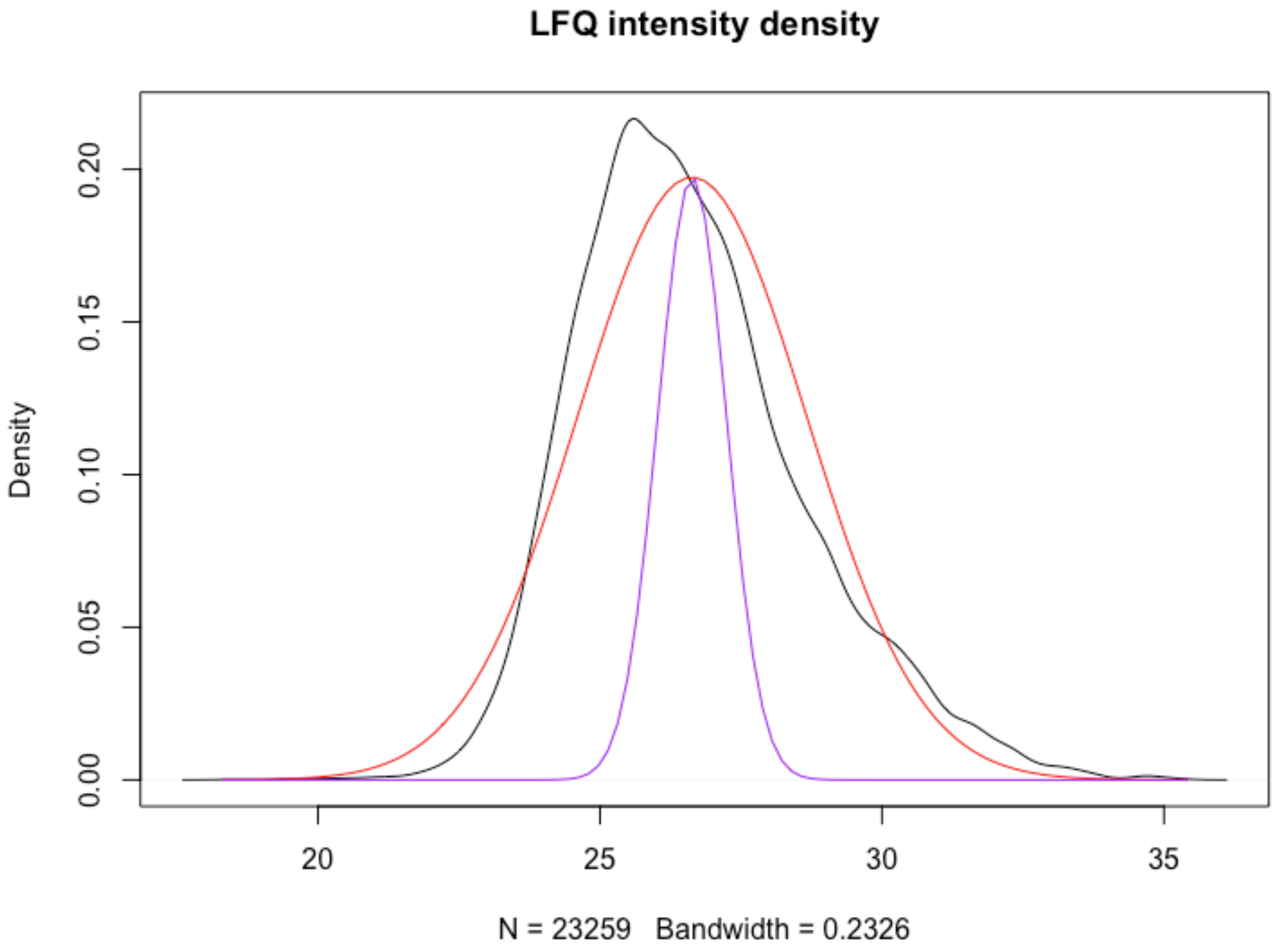


Infer a normal distribution

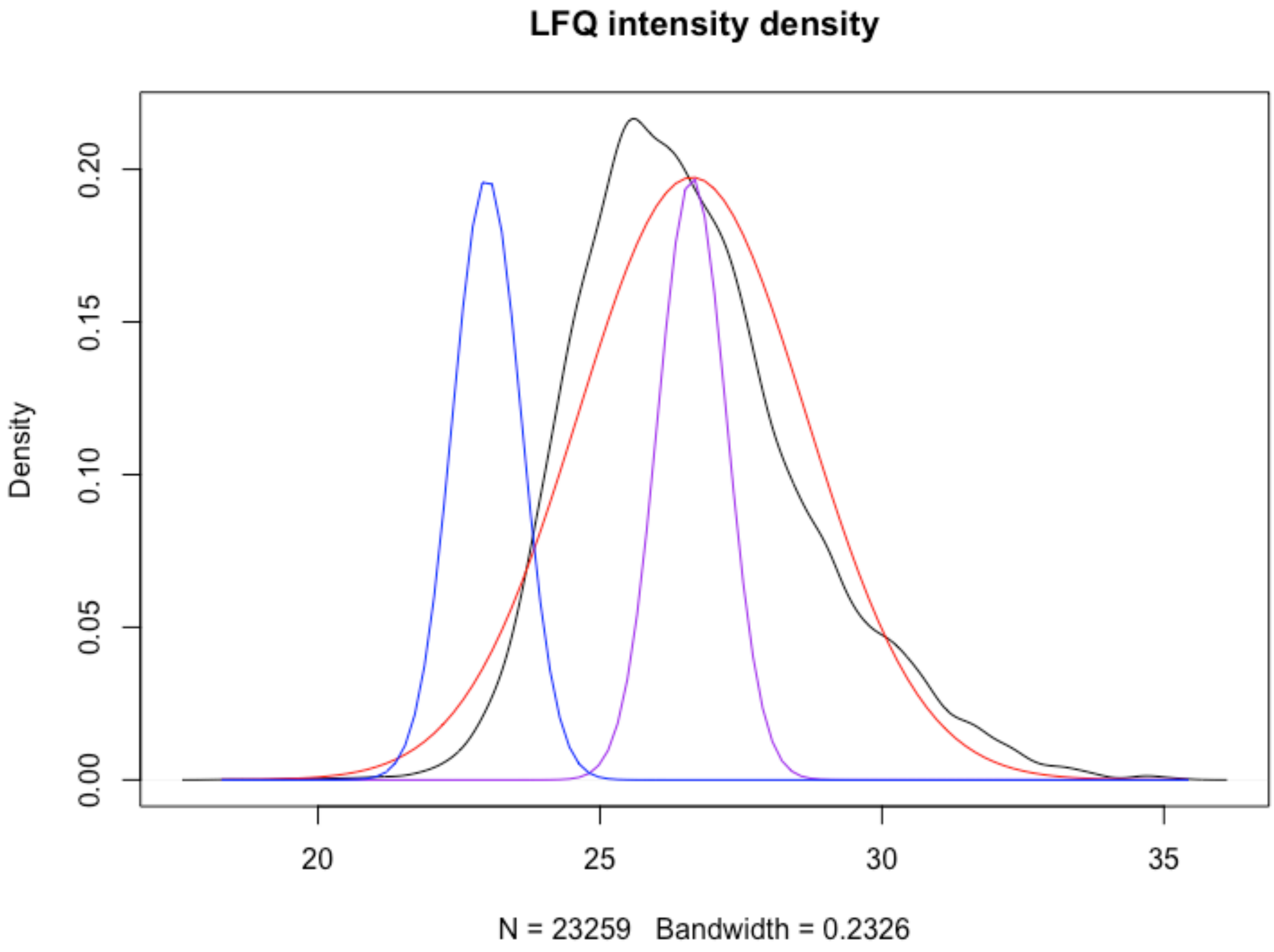
Disclaimer: This is how I understood Perseus does it. ;)



Make the distribution narrow

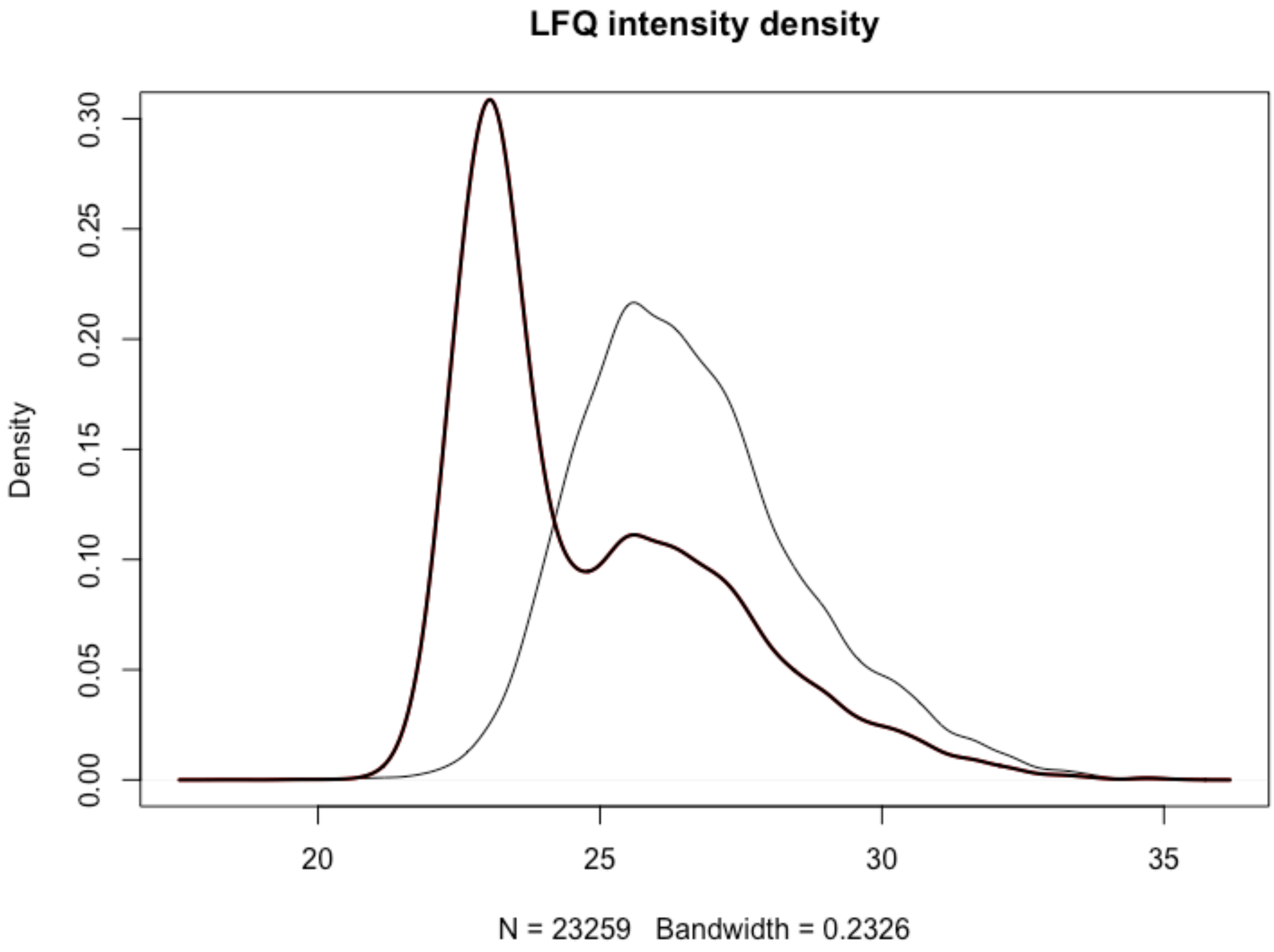


And shift it downwards



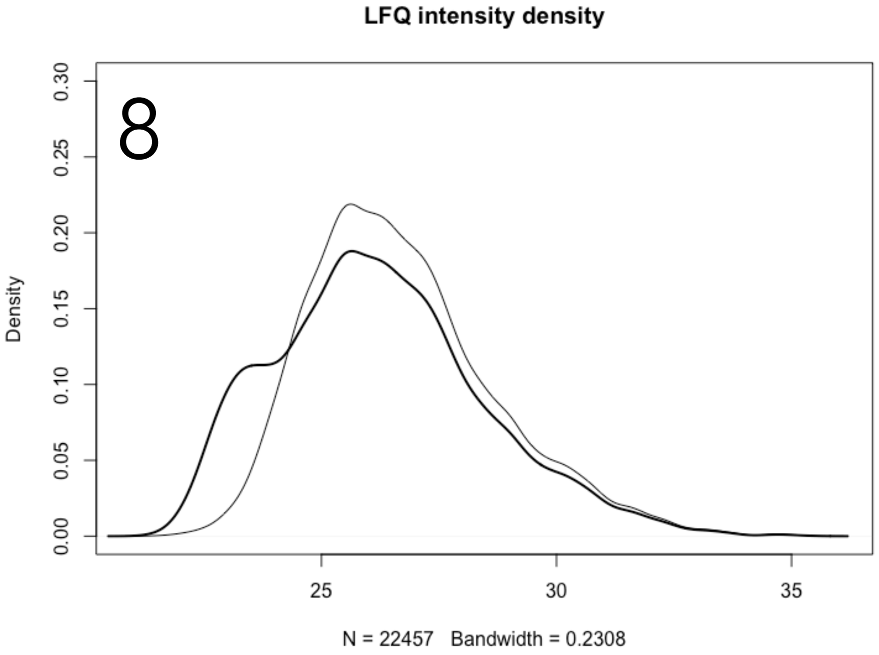
The new distribution models low intensity values

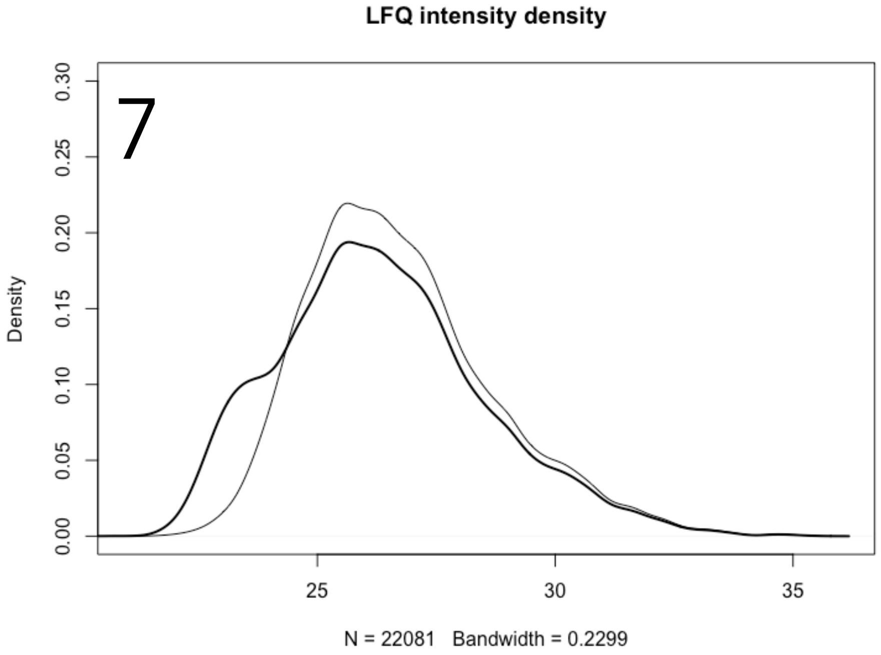
Take values from the new distribution

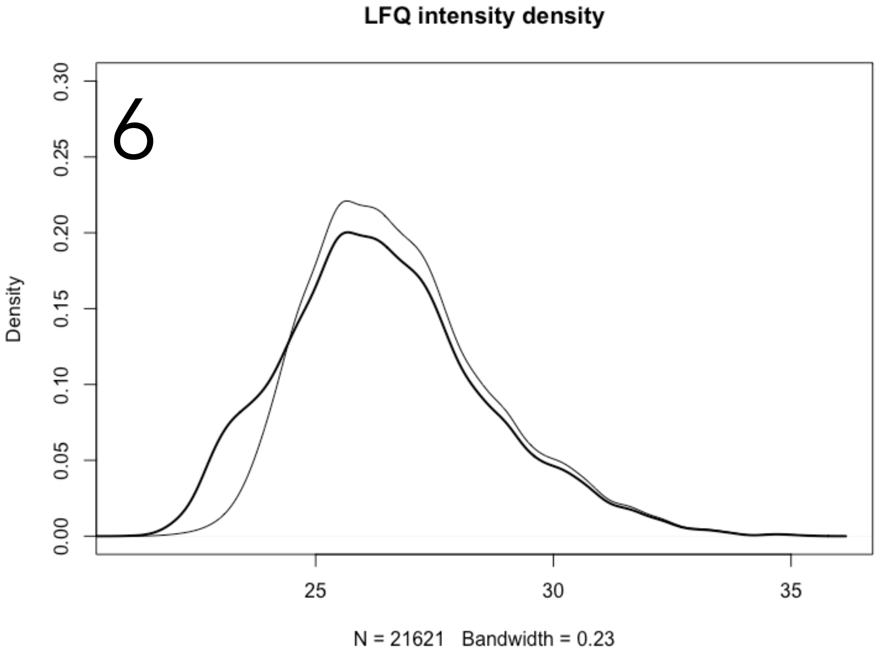


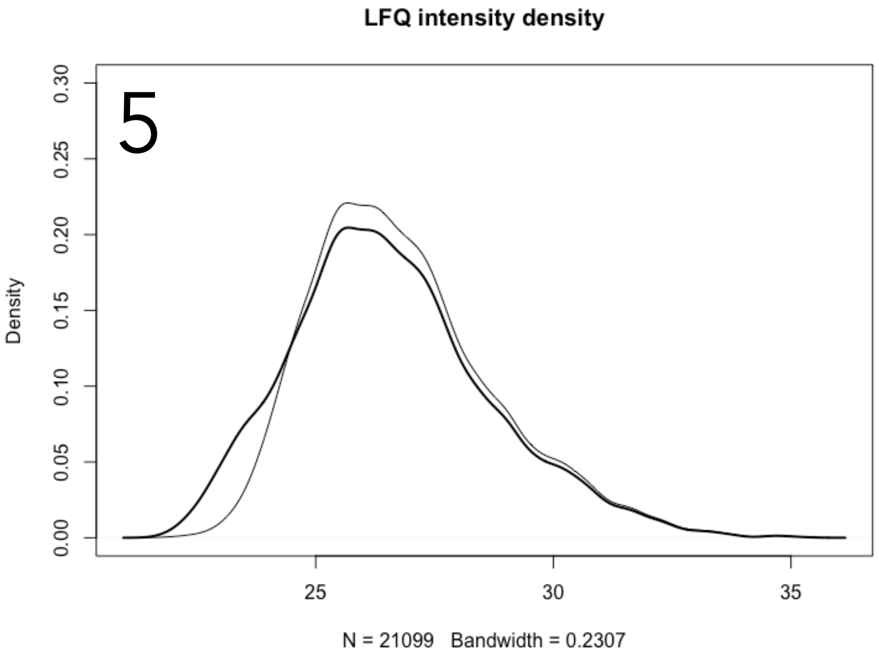
Prefilter your data!

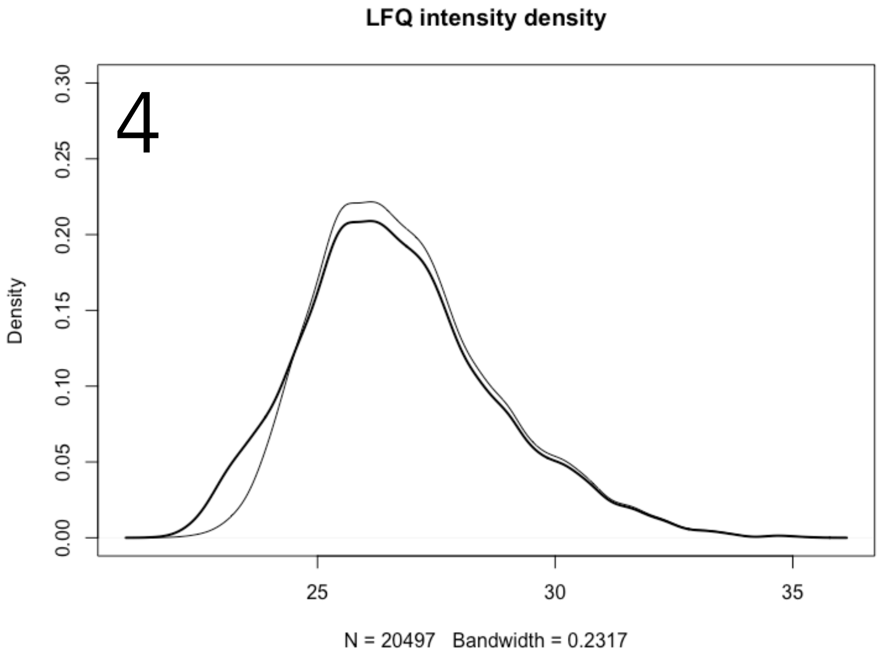
Max x missing values allowed

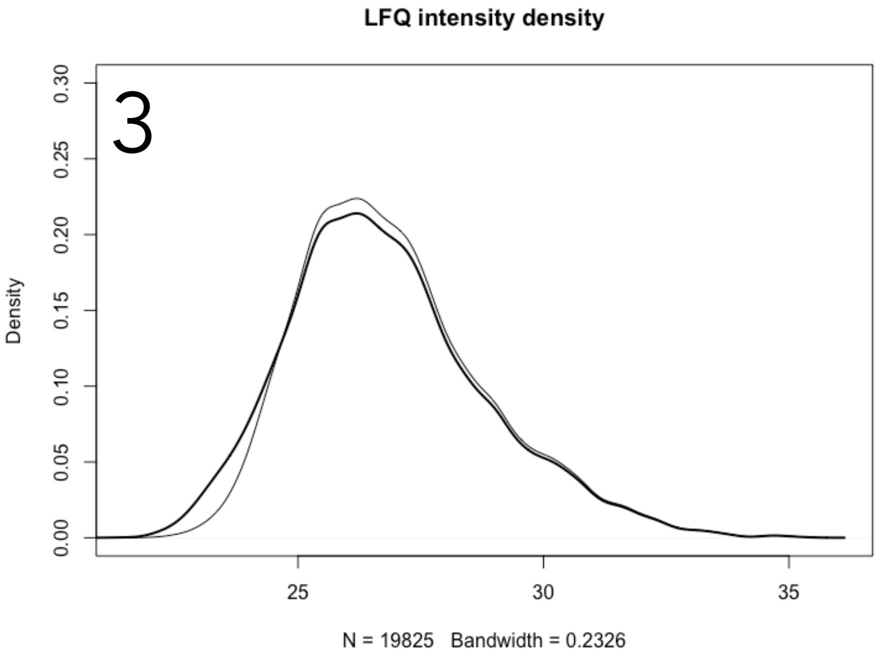












Stay in touch with your raw data!

Imputation guidelines

…according to my experiences

1. Try different parameters (filters, fittings).

2. Always keep the raw data in mind.
   Especially when interpreting your data.

3. Plot, plot, plot.

Now let’s do it in R!

Your tasks

Split in groups of two

Walk through the code for missing value imputation (`impute.R`)
Try to understand the lines

modify the filtering for number NAs allowed

modify the width of the new distribution

https://github.com/higsch/bioinfo-workshop

https://github.com/higsch/bioinfo-workshop