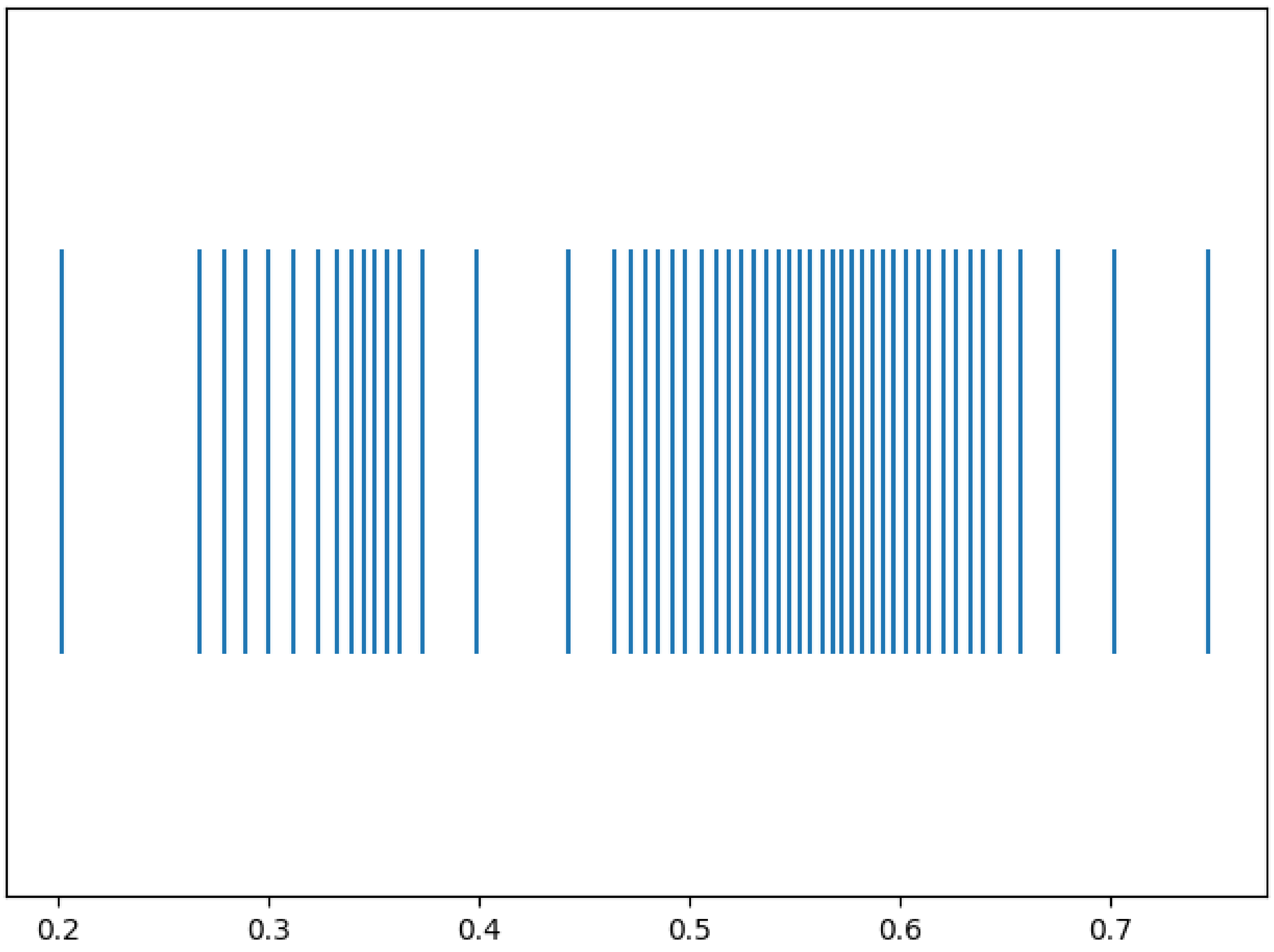
Paweł Szewczyk

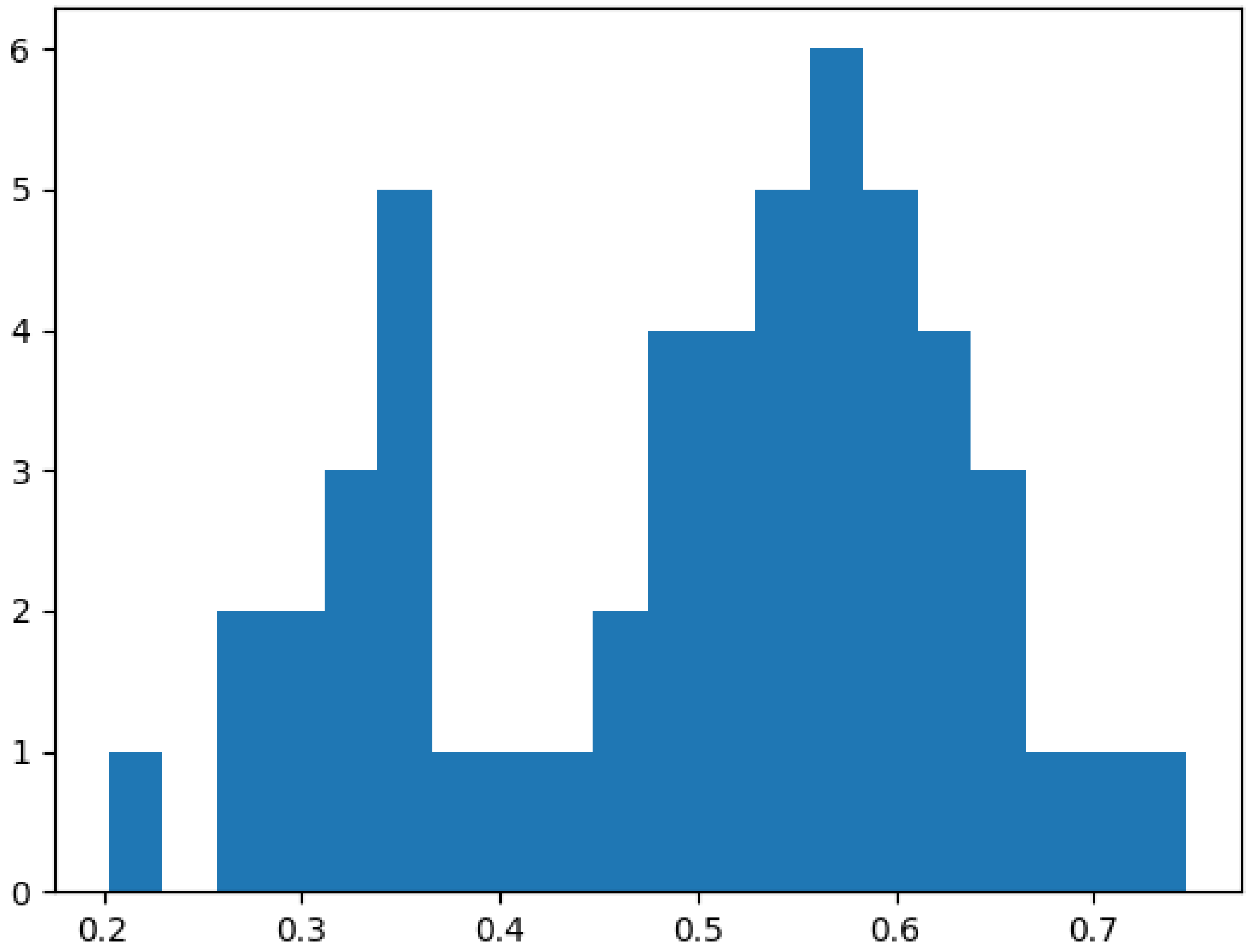
Data analysis recipes: Choosing the binning for a histogram

Data analysis recipes: Choosing the binning for a histogram[1]

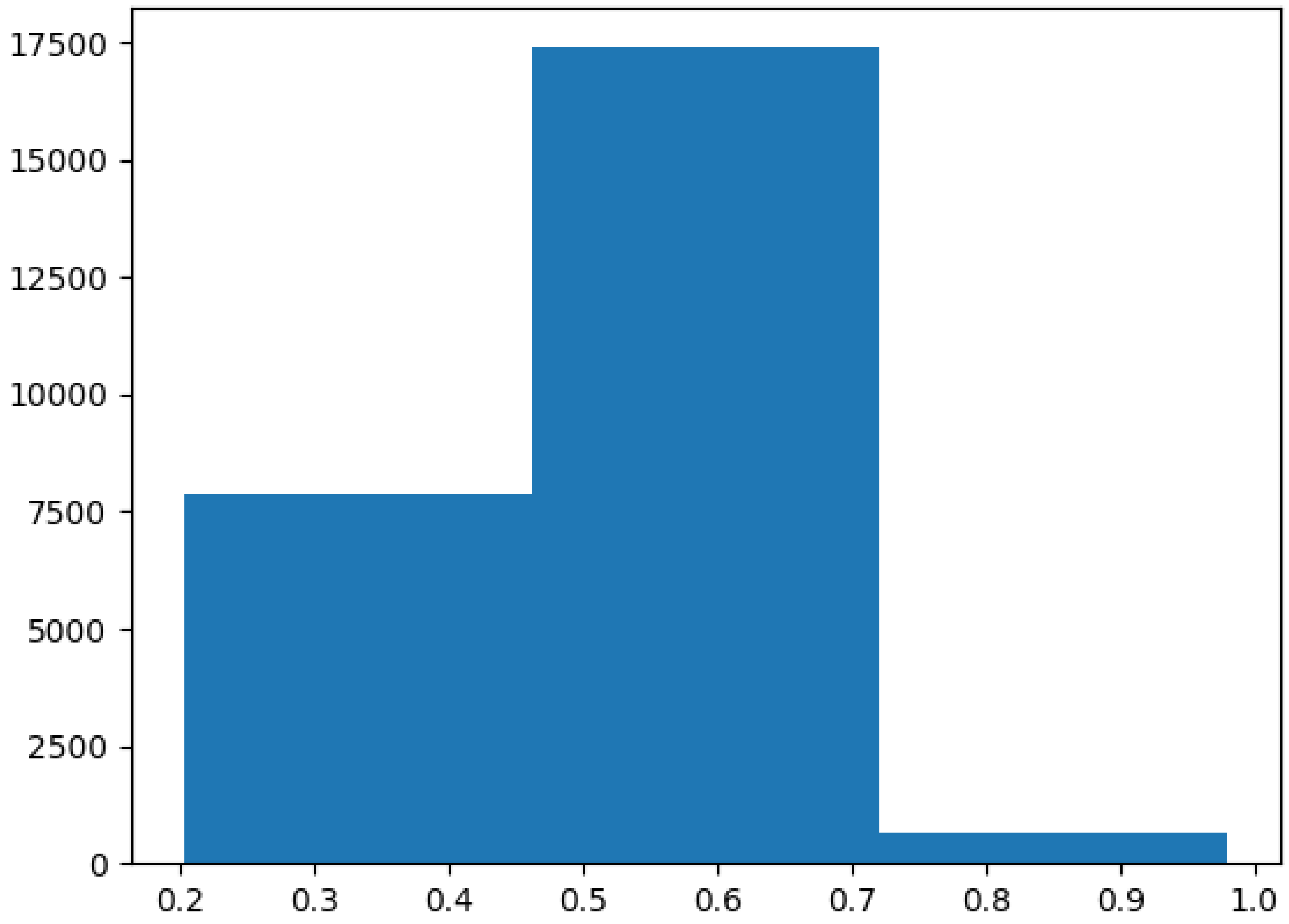
David W. Hogg
Center for Cosmology and Particle Physics, Department of Physics
New York University
david.hogg@nyu.edu

Histograms

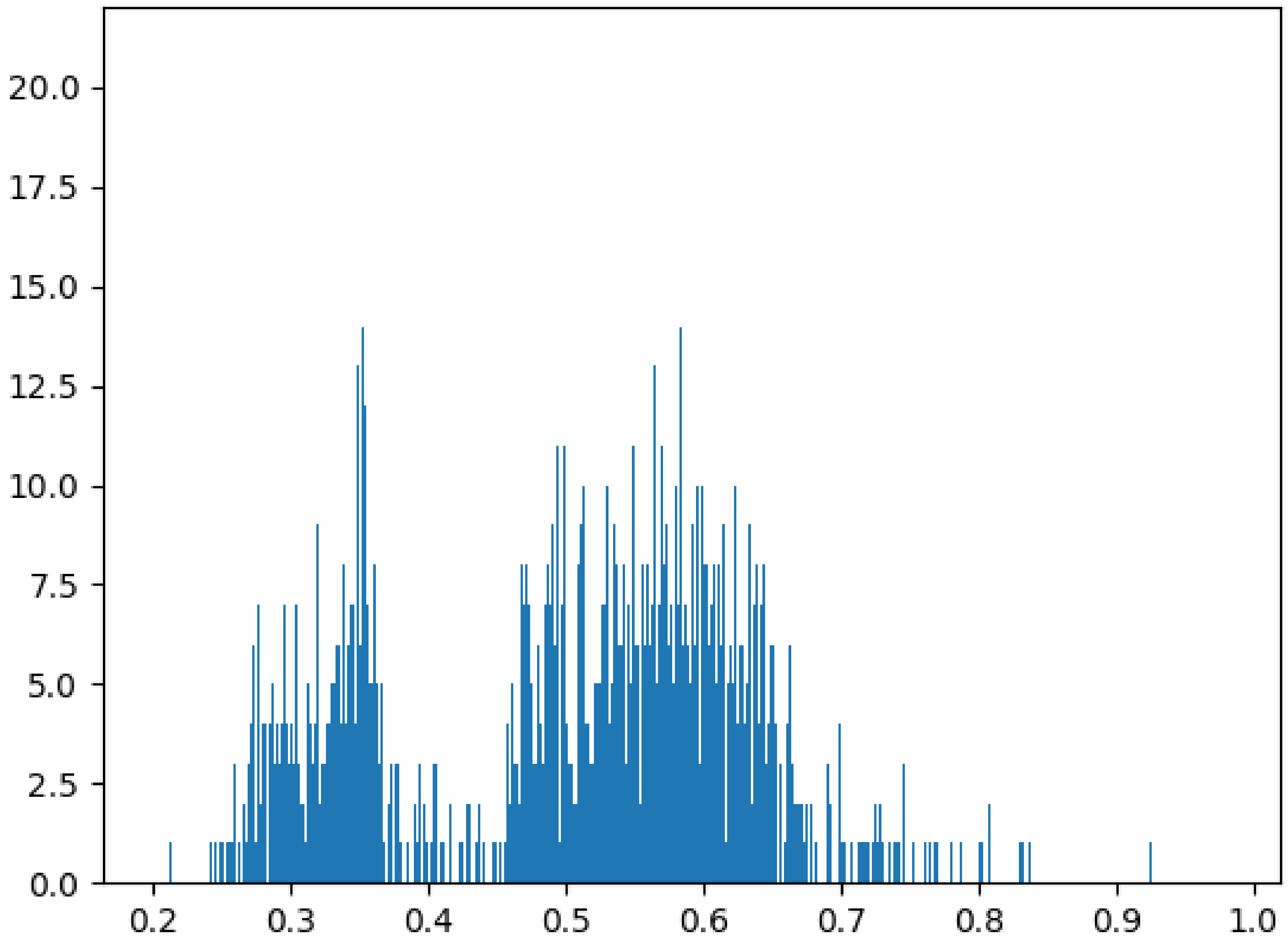




Large bins



Empty bins



Histogram vs probability distrubution

- Probability p that the data lands in bin i
- Naive approach:

$$p(i) = \frac{N_i}{\sum_k N_k}$$

- Empty bins give p = 0

Probability distribution:

$$\tilde{f}(\mathbf{x}) = \frac{p\left(i\left(\mathbf{x}\right)\right)}{\mathbf{\Delta}_{i(\mathbf{x})}}$$

Histogram vs probability distrubution

- Probability with smoothing constant:

$$p(i) = \frac{N_i + \alpha}{\sum_k [N_k + \alpha]}$$

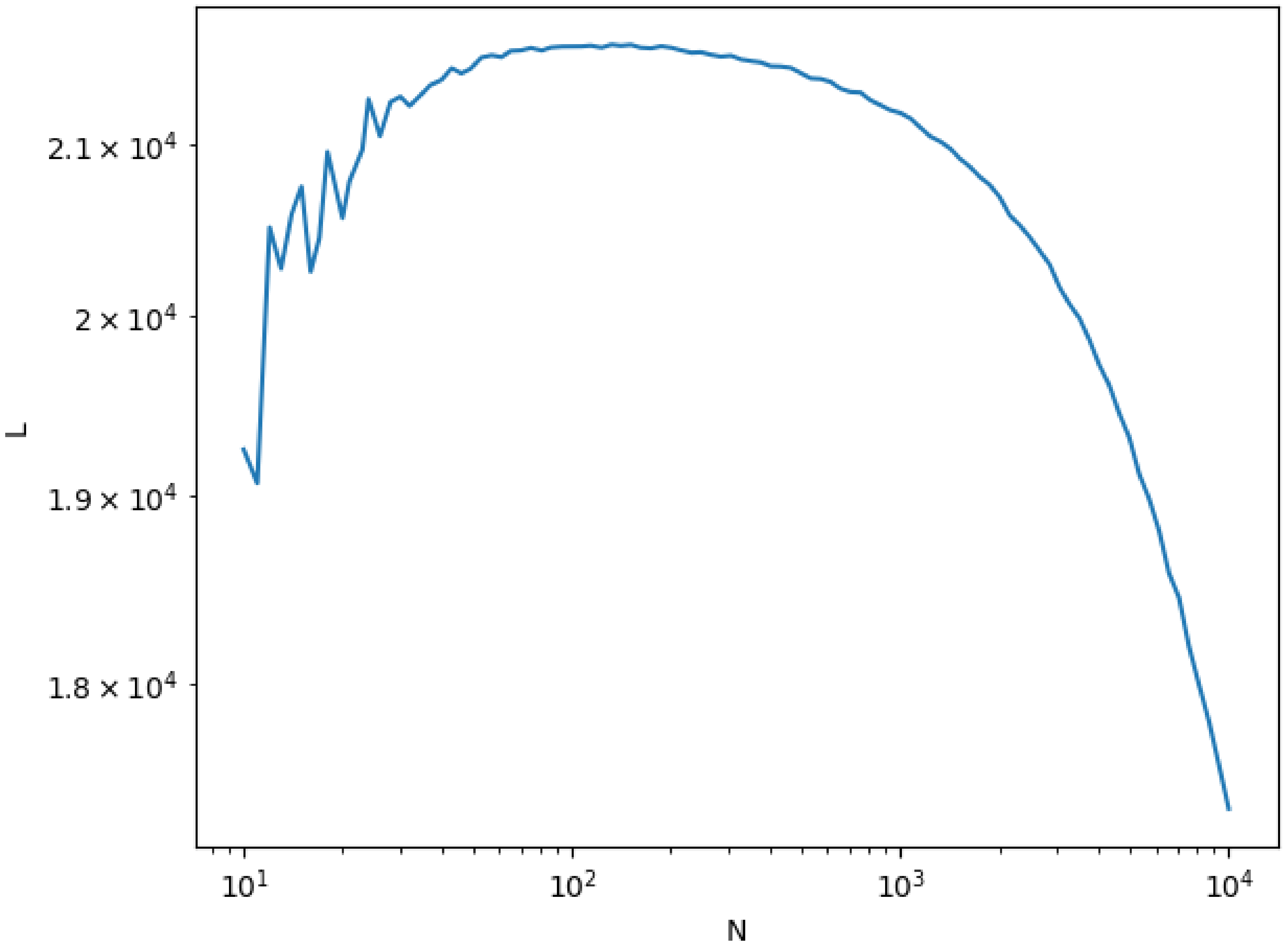
- Weighted data

$$p(i) = \frac{W_i + \alpha}{\sum_k [W_k + \alpha]}$$

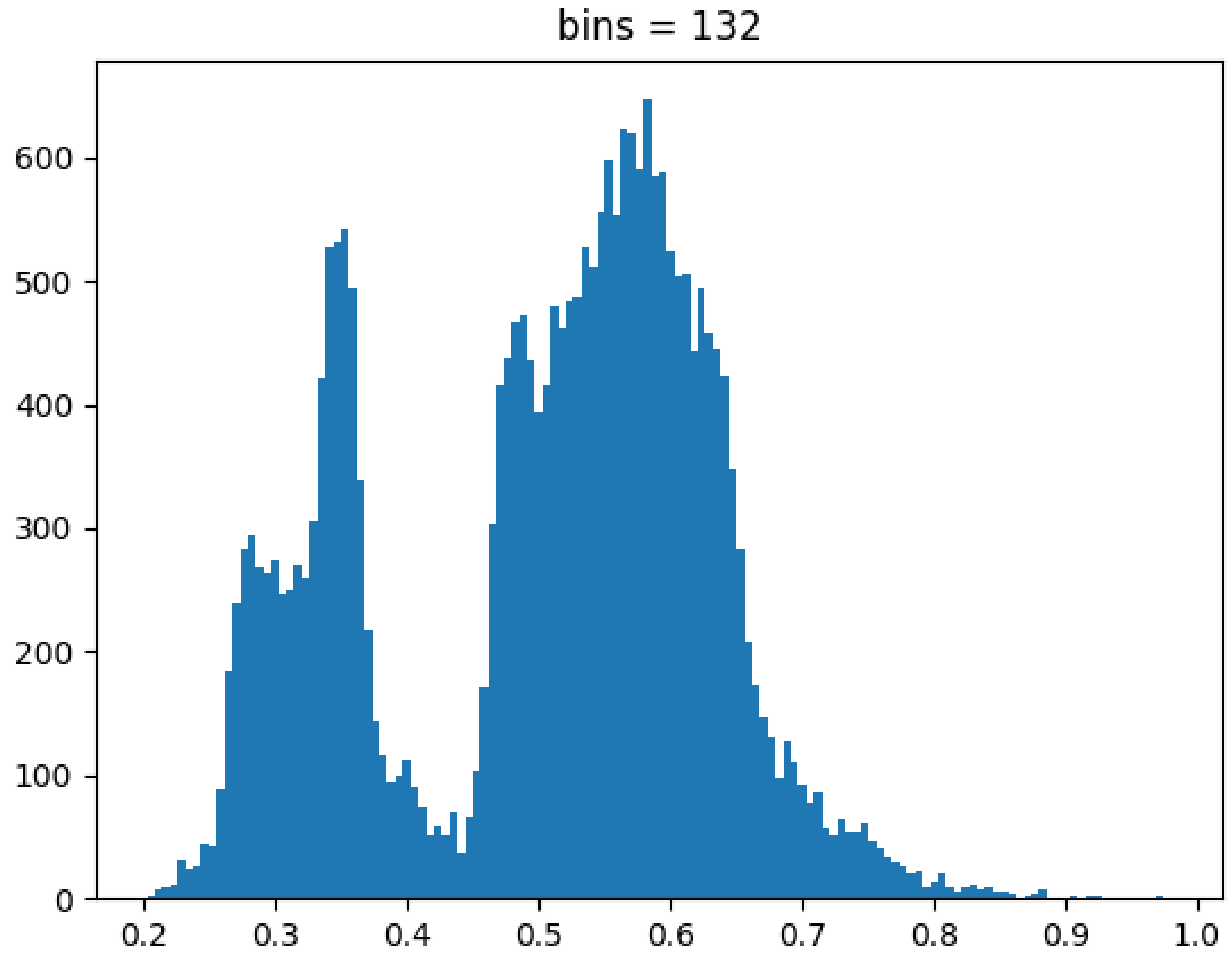
Likelyhood function

$$L = \sum_i N_i \ln\left(\frac{N_i + \alpha - 1}{\boldsymbol{\Delta}_i \left[\sum_k [N_k + \alpha] - 1\right]}\right)$$

Likelyhood vs number of bins



Best bin size



Multi-dimensional case