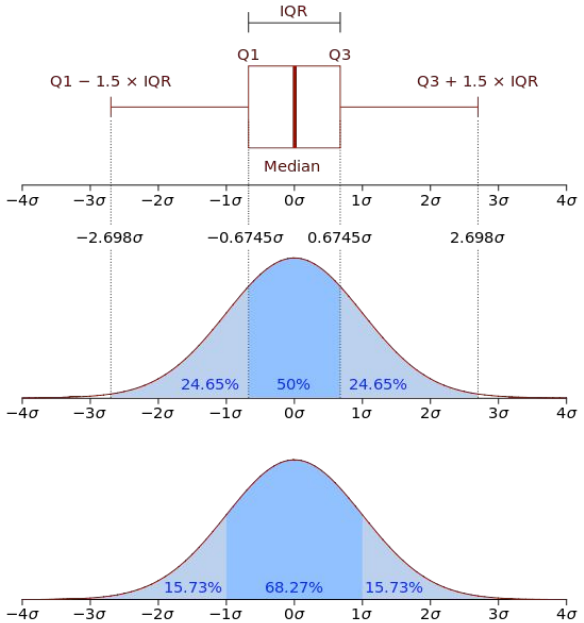
Data Visualization
W7-1

Quiz

- What do you find interesting in today's VotW?
- Explain every element (e.g., box, line, …) of a box plot. What do they represent and how do we calculate them?
- In histogram, _____ represents the frequency of the data.
- What are the issues when bins are too wide or too narrow in drawing histograms?

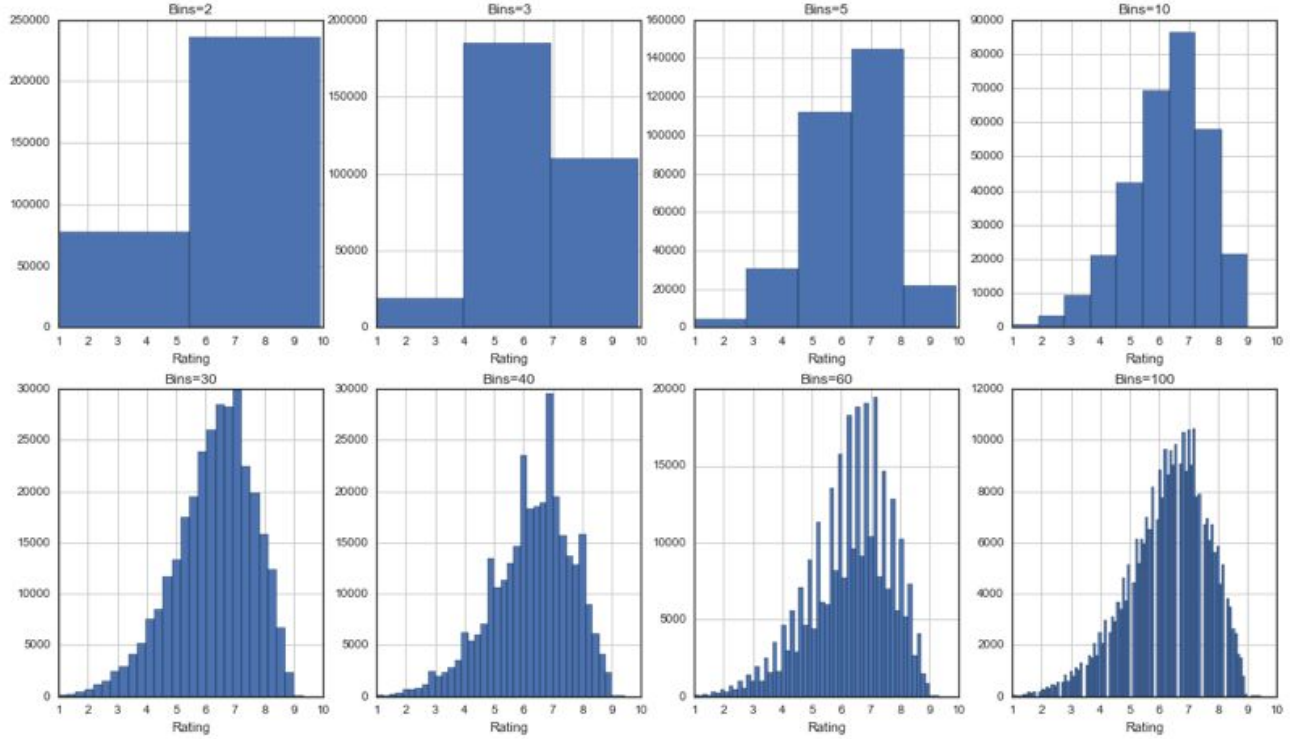
Let's construct a box plot



1. We want to know the median (the central value in the data).
2. What's the range that capture about the half of the data points?
3. How about most of the data points? (Many ways to estimate this!)
4. What do we do with the rest of the data points?

In histogram, **area**, not the height, should represent the frequency!

Too wide bins → large bias (inaccurate)
Too narrow bins → large variance (overfit to the data/noise)



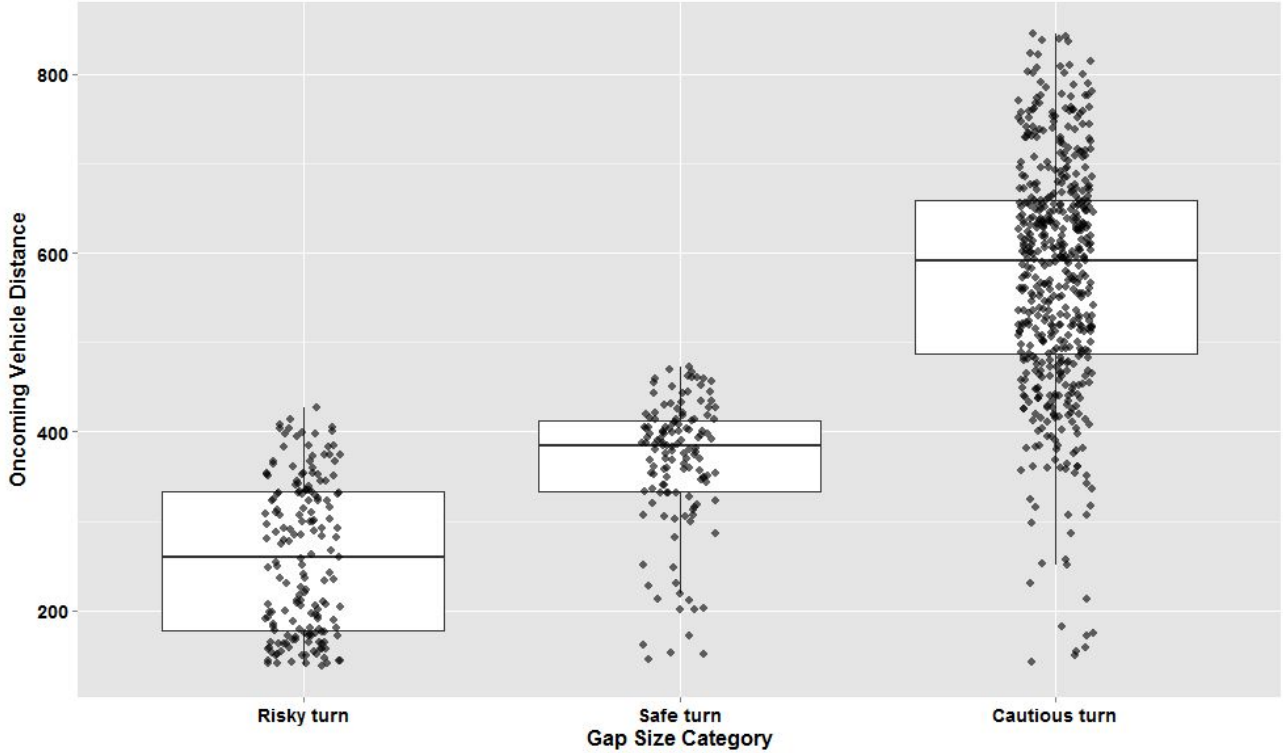
Histogram is cheap and easy.

Try multiple histograms

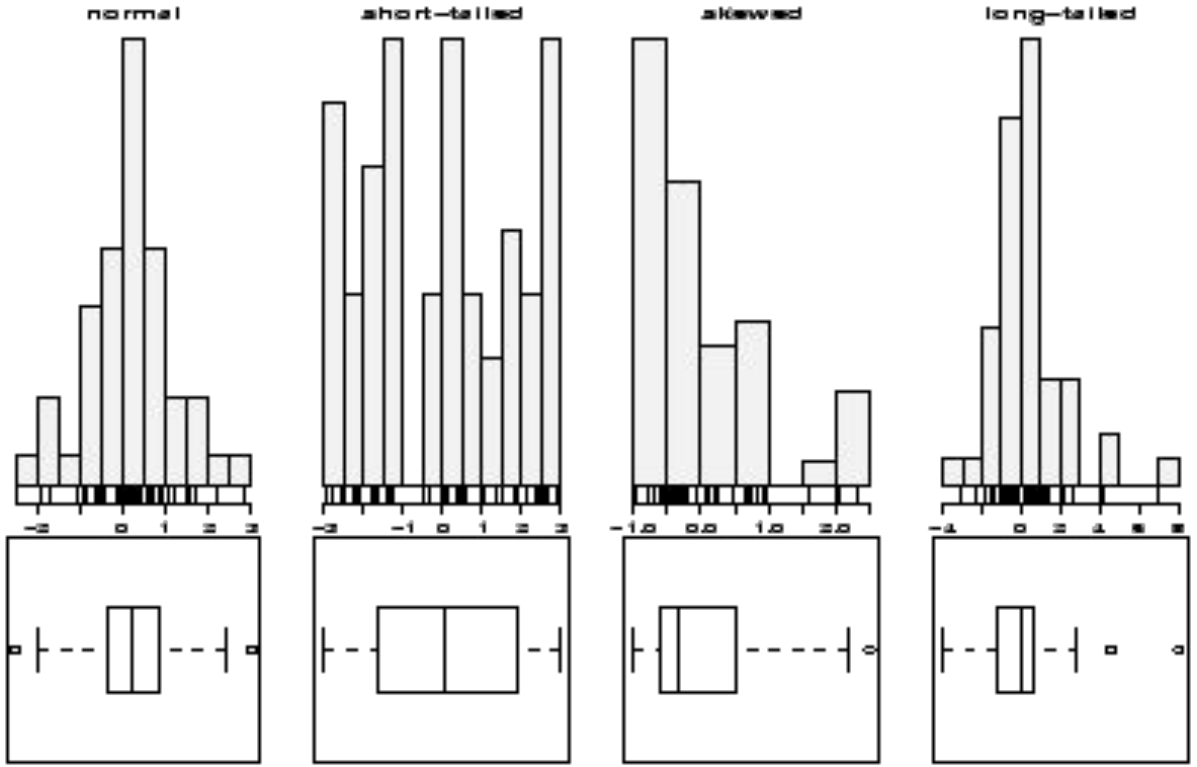
Boxplot? histogram?

Which should I use?

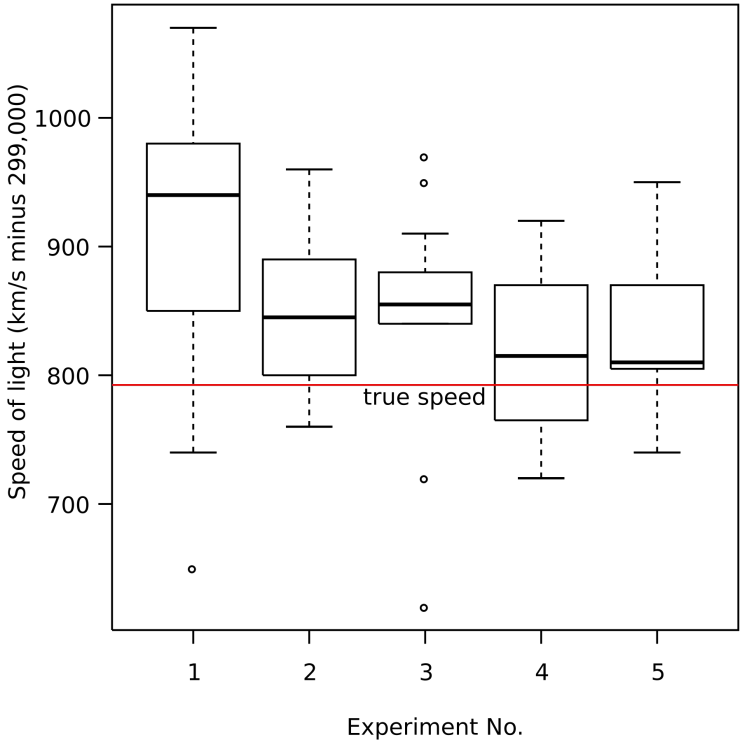
It’s easy to combine a boxplot with scatterplot



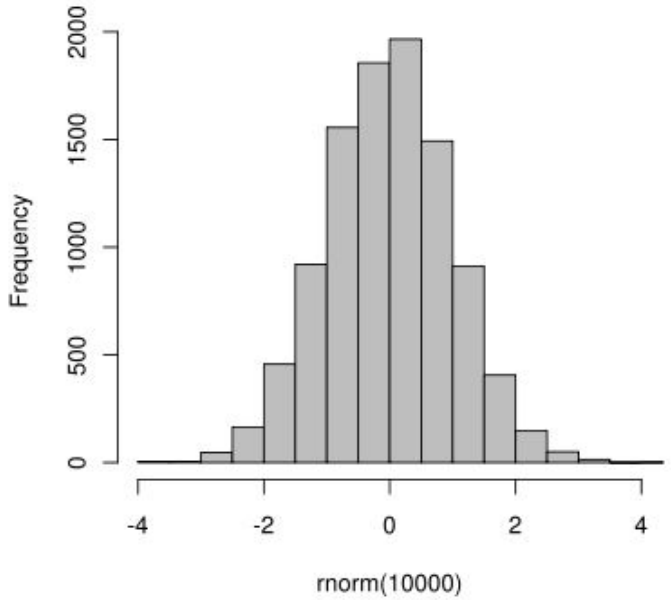
Boxplots hide details



Boxplots are succinct and useful for comparison

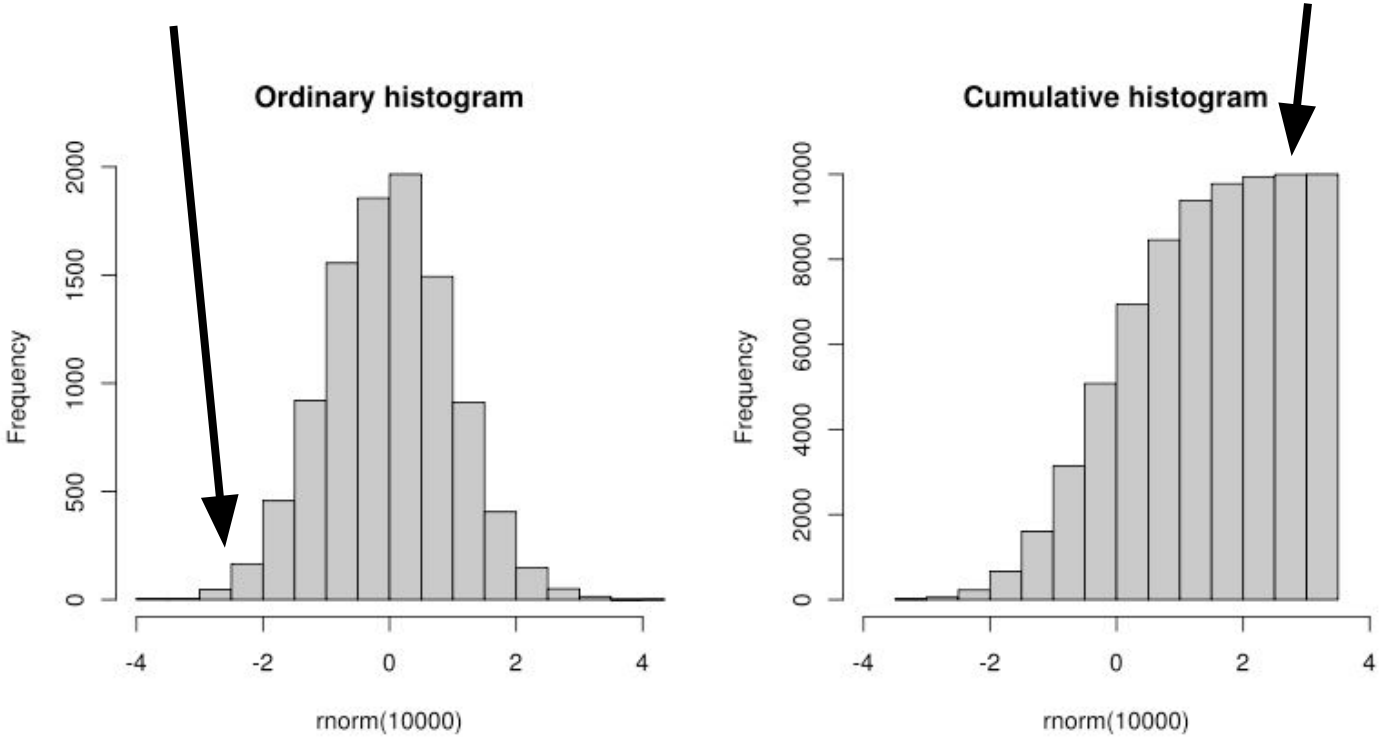


Q: Can we design a **new way to draw histogram** so that you can easily spot the location of median
(or even any arbitrary percentile points)?

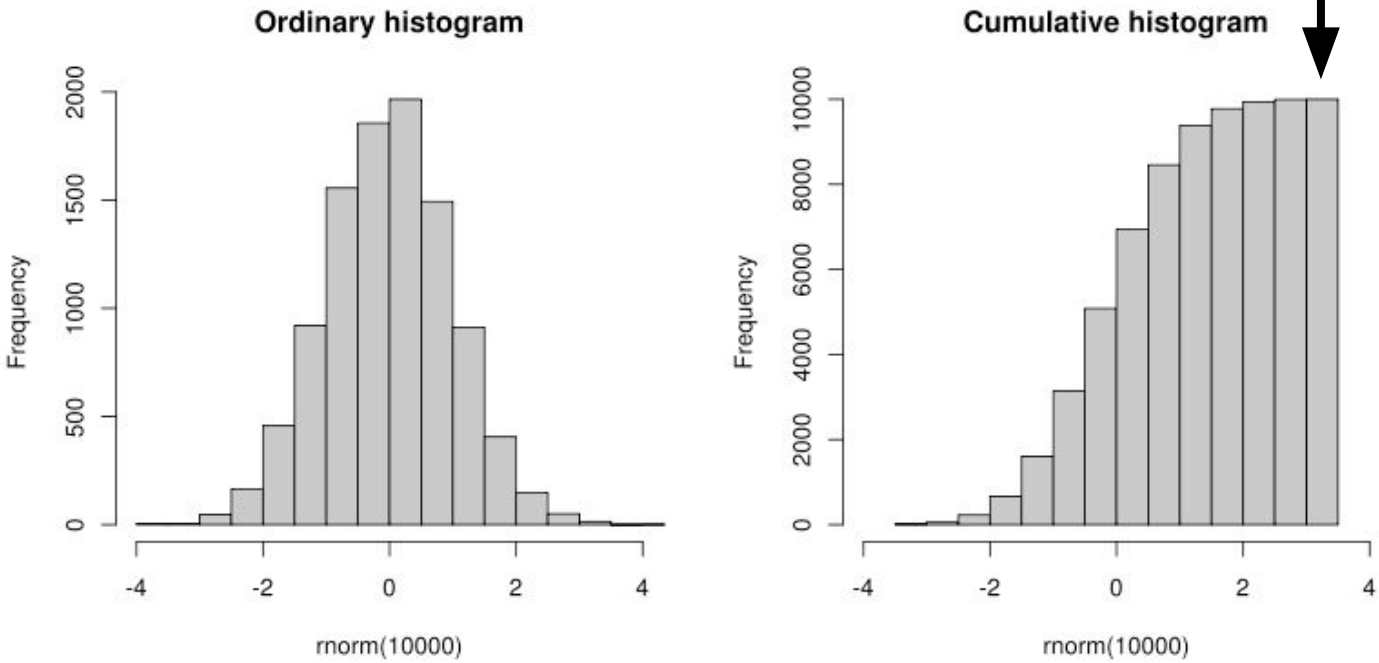


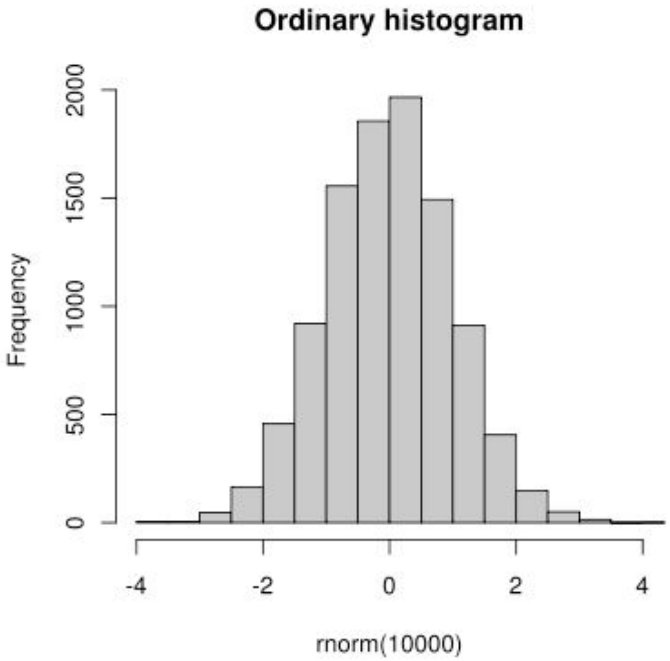
"The number of data points within this bin"

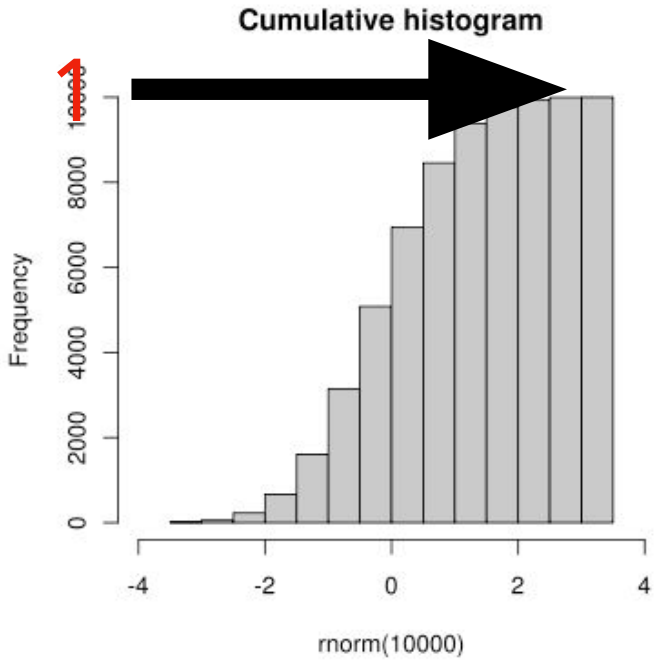
"The number of data points we have seen so far"

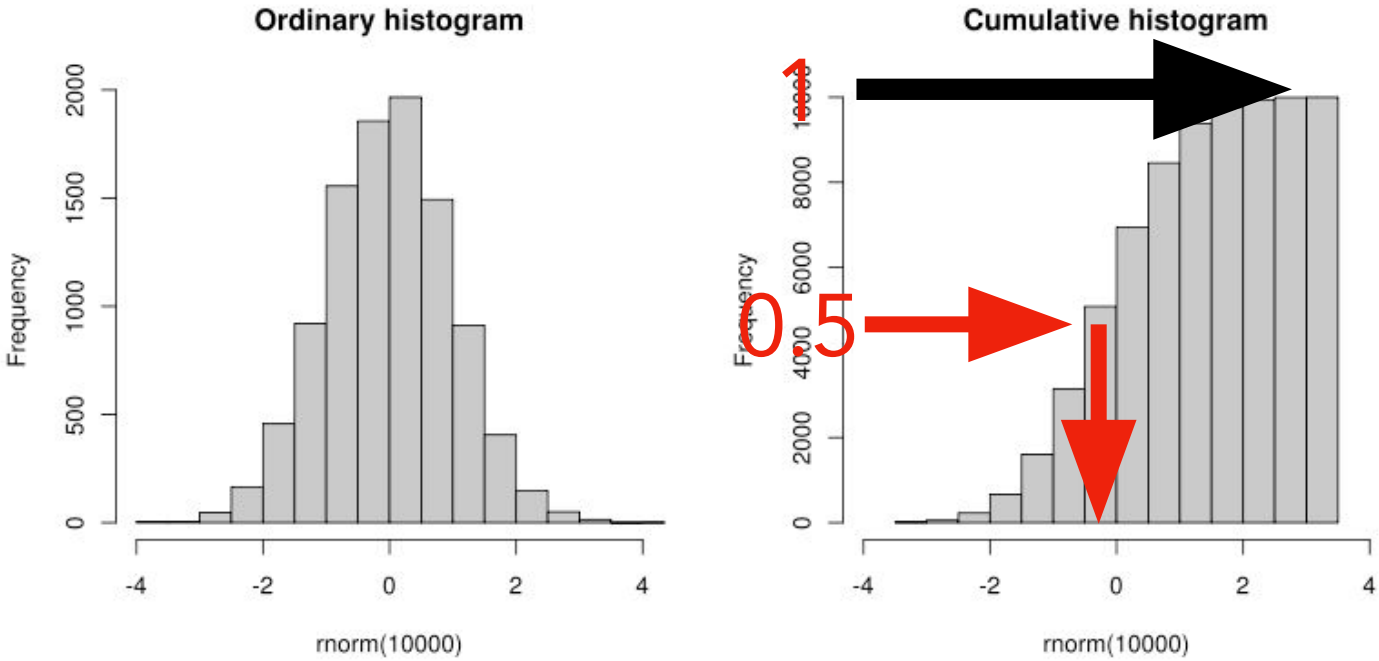


The total number of data points

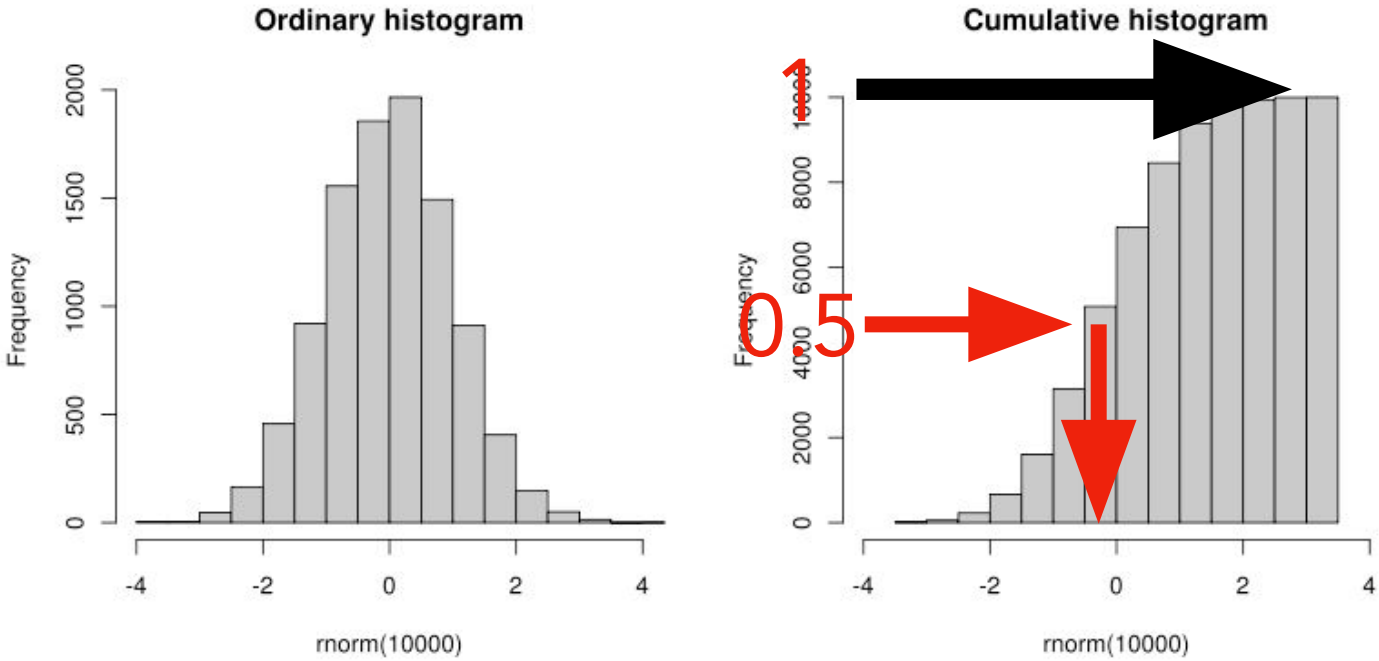


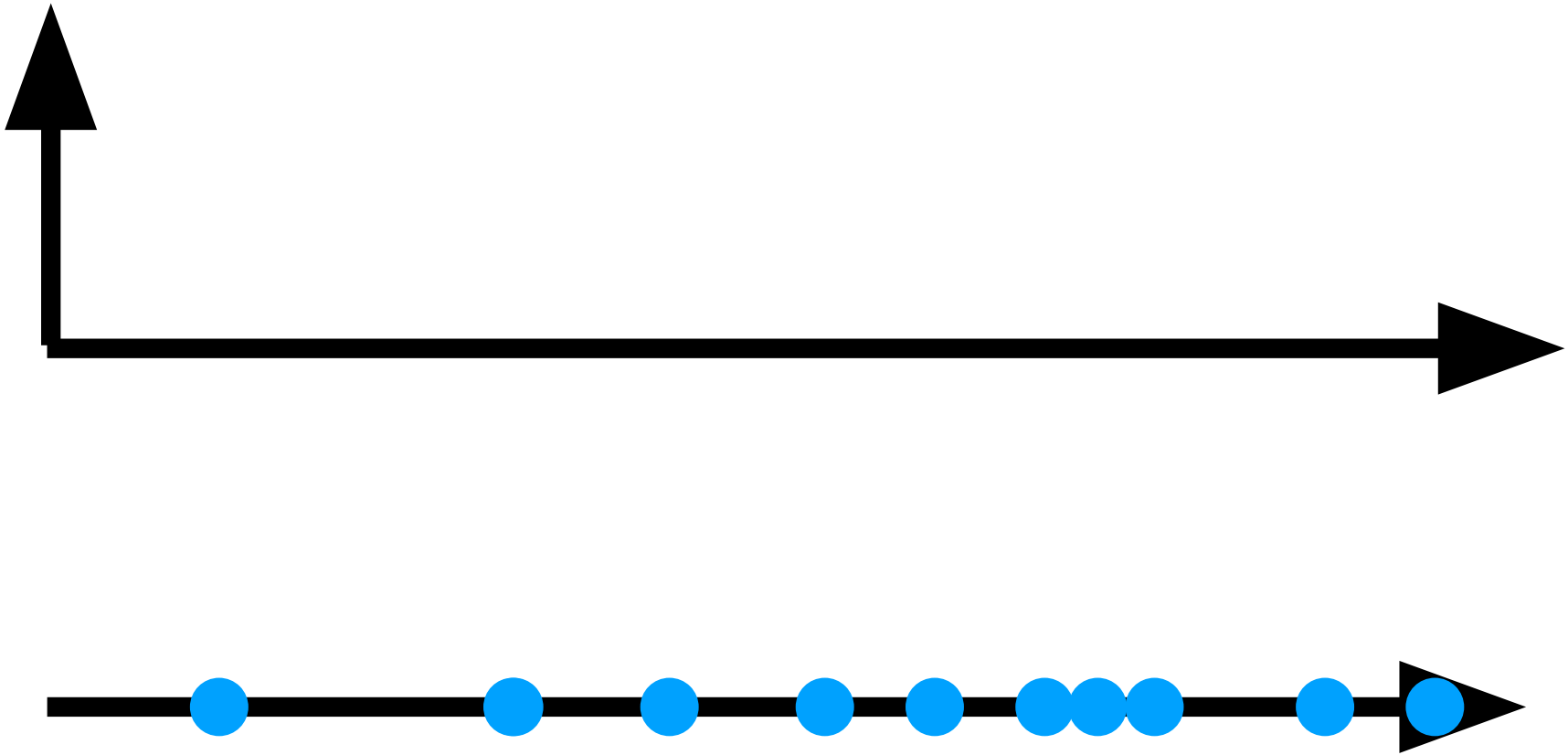

Ordinary histogram
Frequency
0
500
1000
1500
2000
-4
-2
0
2
4
rnorm(10000)

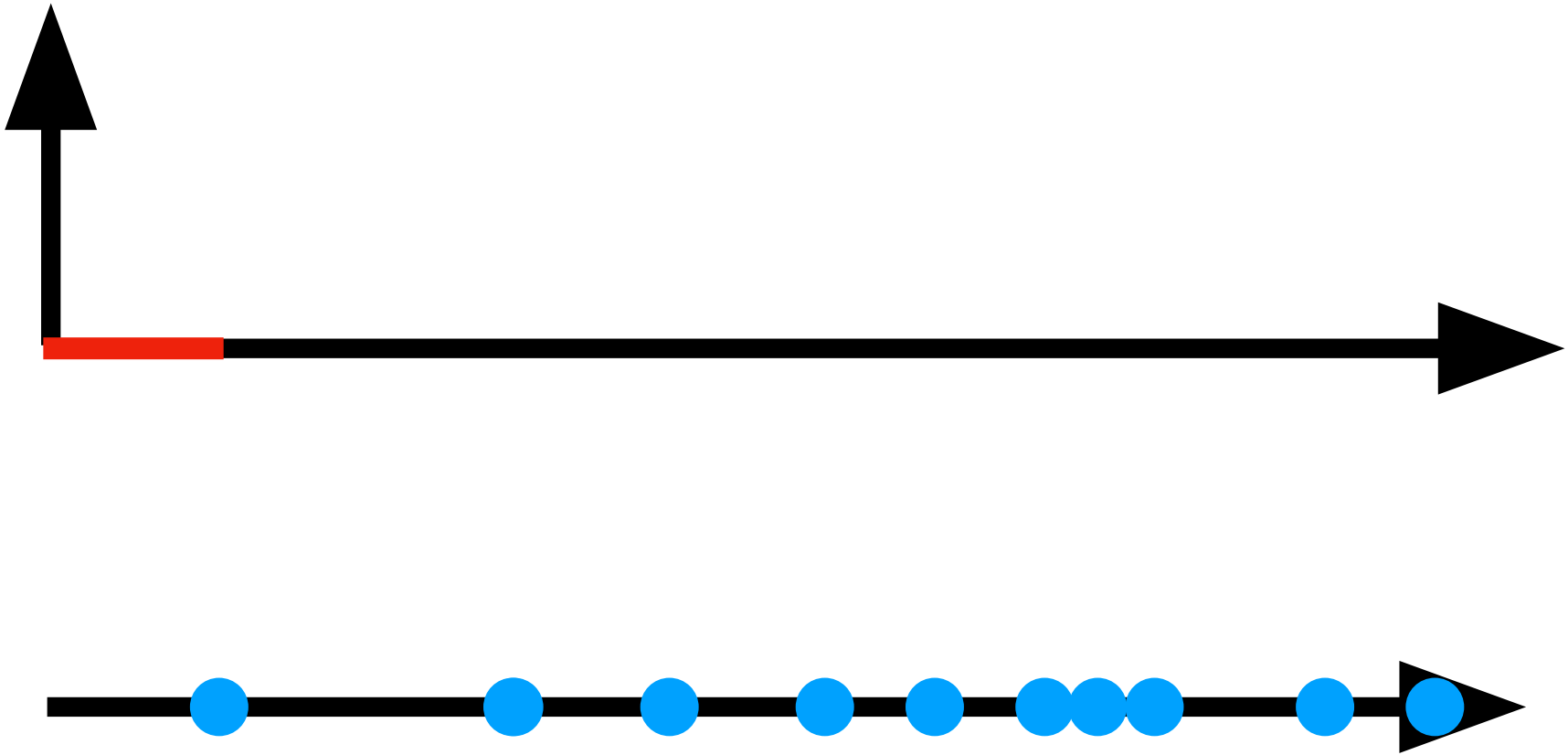

Cumulative histogram
Frequency
0
2000
4000
6000
8000
-4
-2
0
2
4
rnorm(10000)

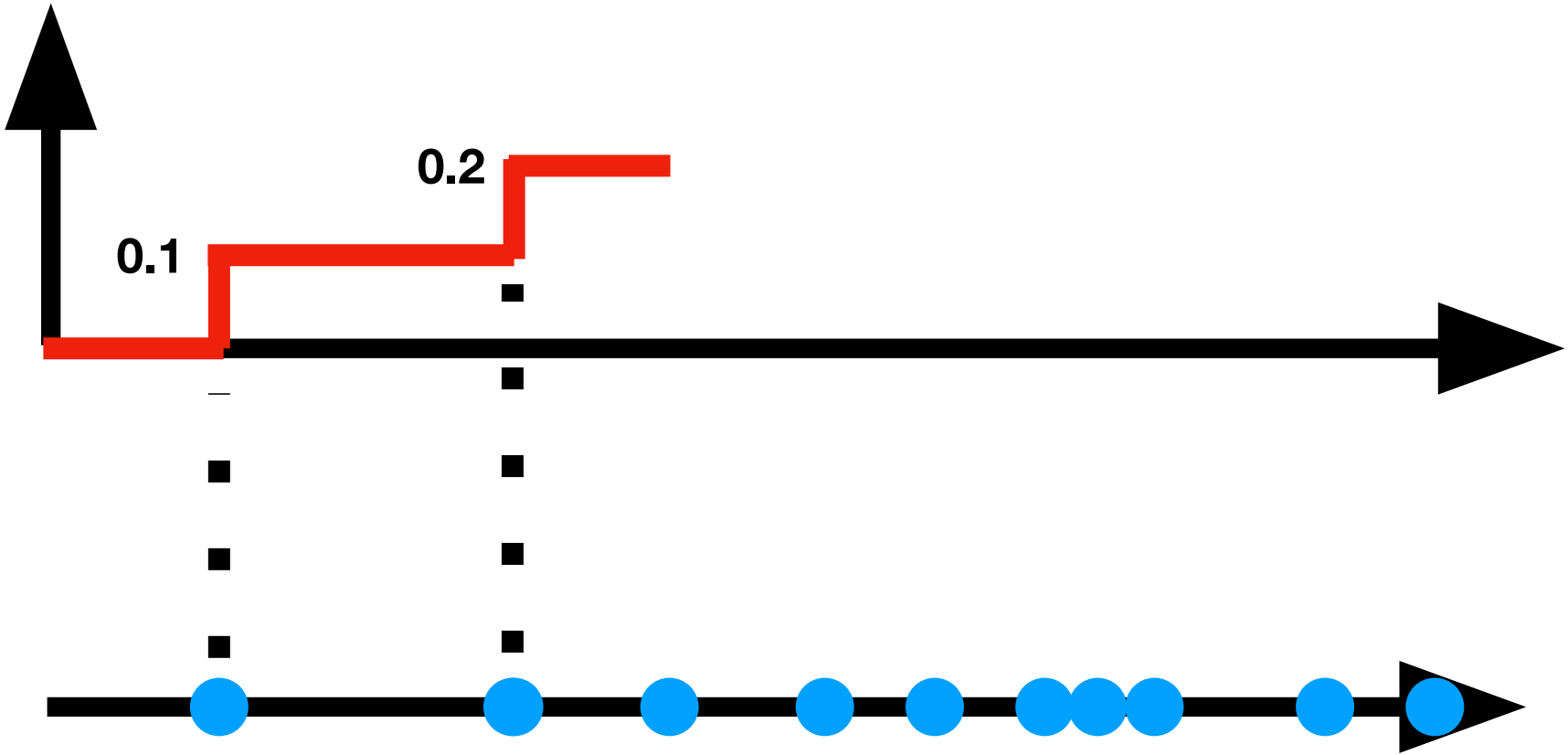

Ordinary histogram
Frequency
0
500
1000
1500
2000
-4
-2
0
2
4
rnorm(10000)
Cumulative histogram
Frequency
0
2000
4000
6000
8000
10000
-4
-2
0
2
4
rnorm(10000)
0.5

Can we improve it even further???
(to make it have a higher resolution)








0.2
0.1

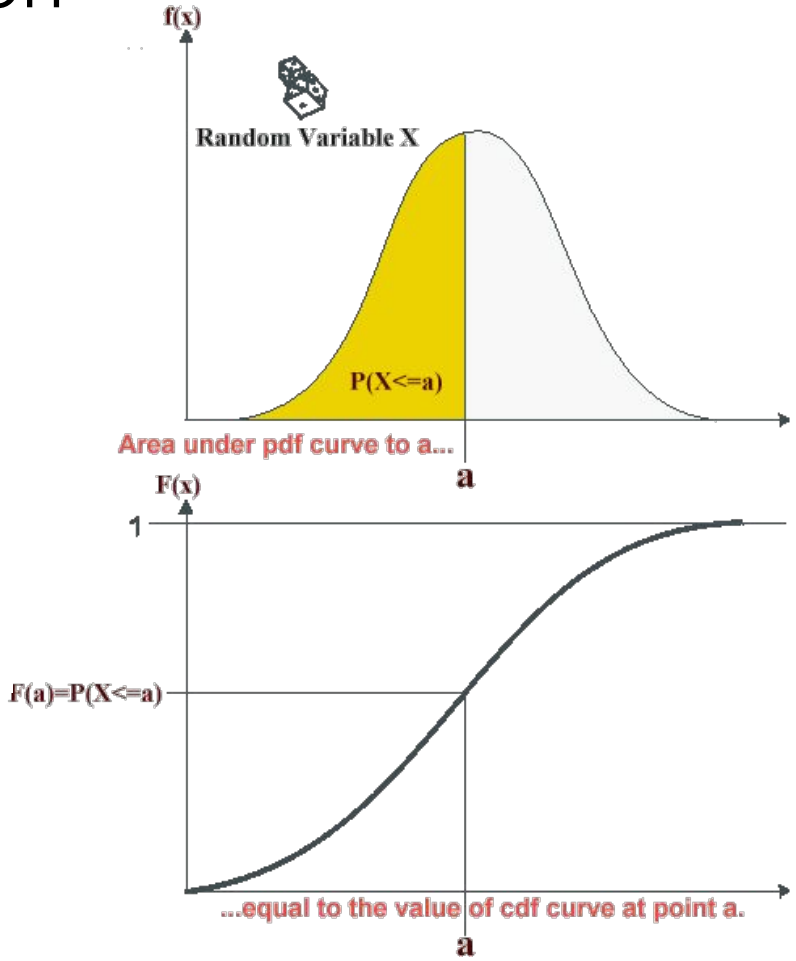
CDF (Cumulative Distribution Function)

$$F_X(x) = P(X \leq x)$$

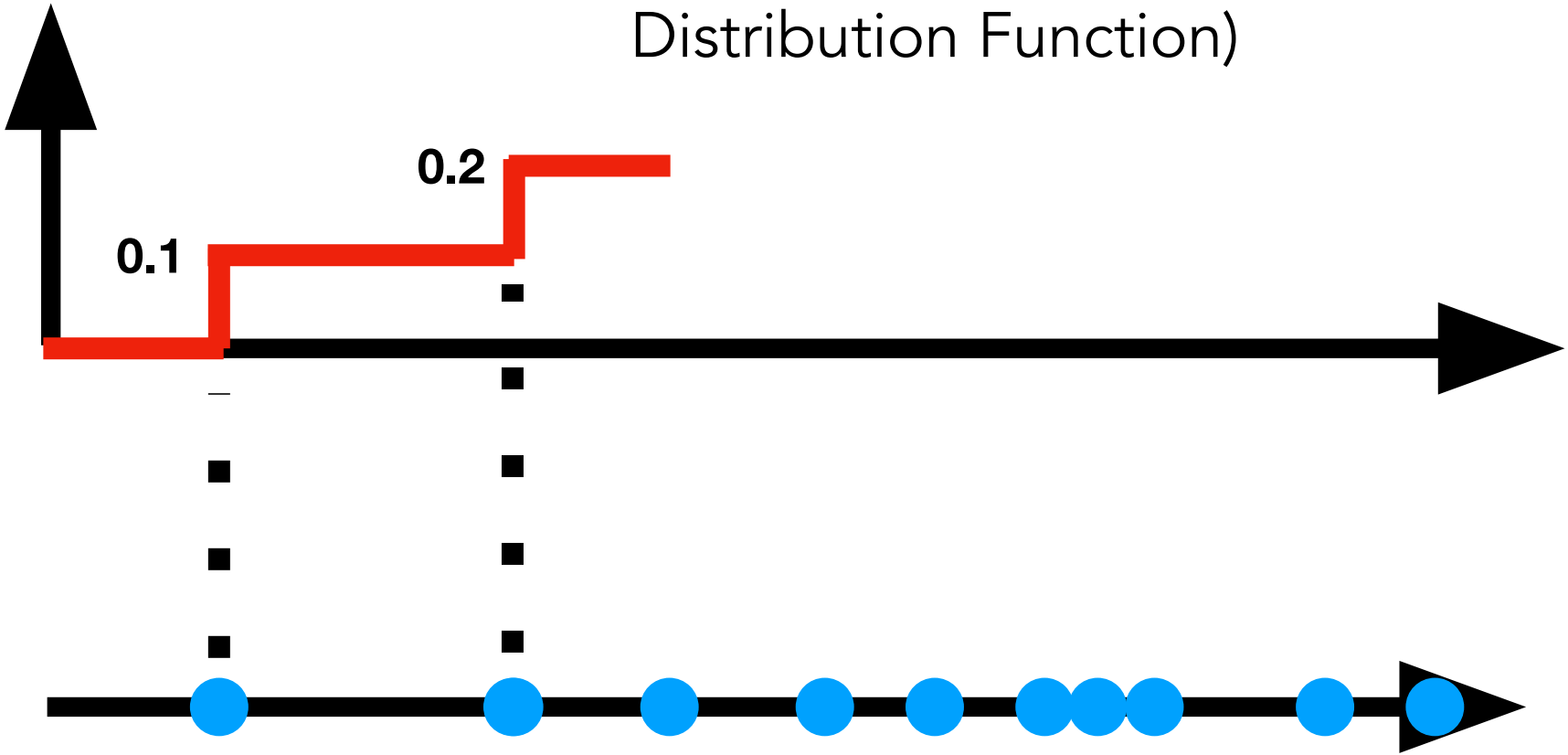
$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

Prob. density function (PDF)

We don't need any bins!!



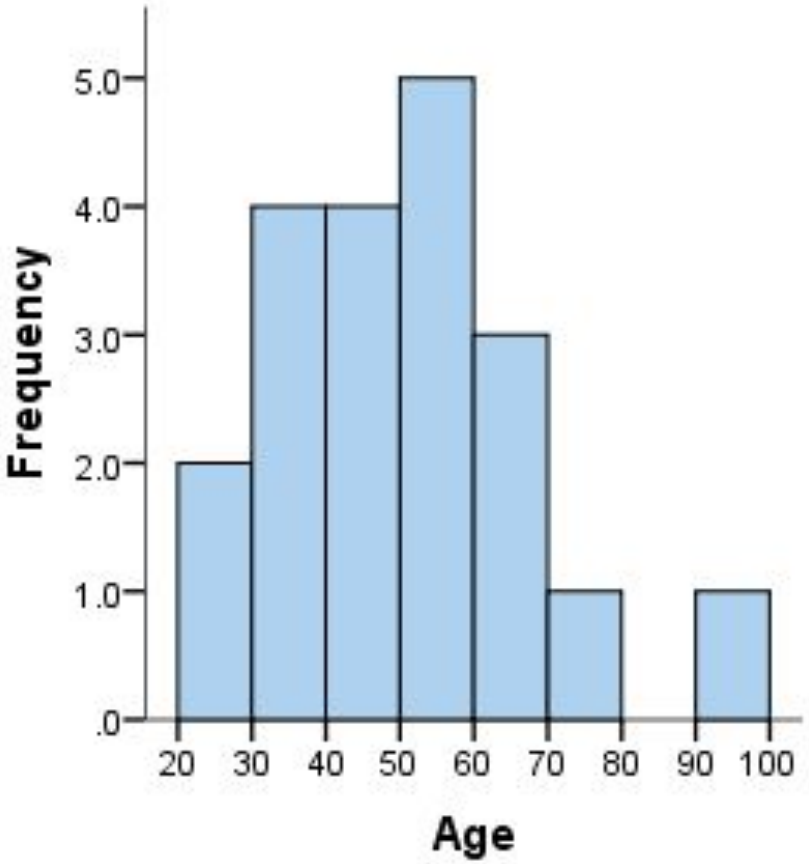
Empirical CDF (Empirical Cumulative Distribution Function)



Let’s draw the empirical CDF for the following data

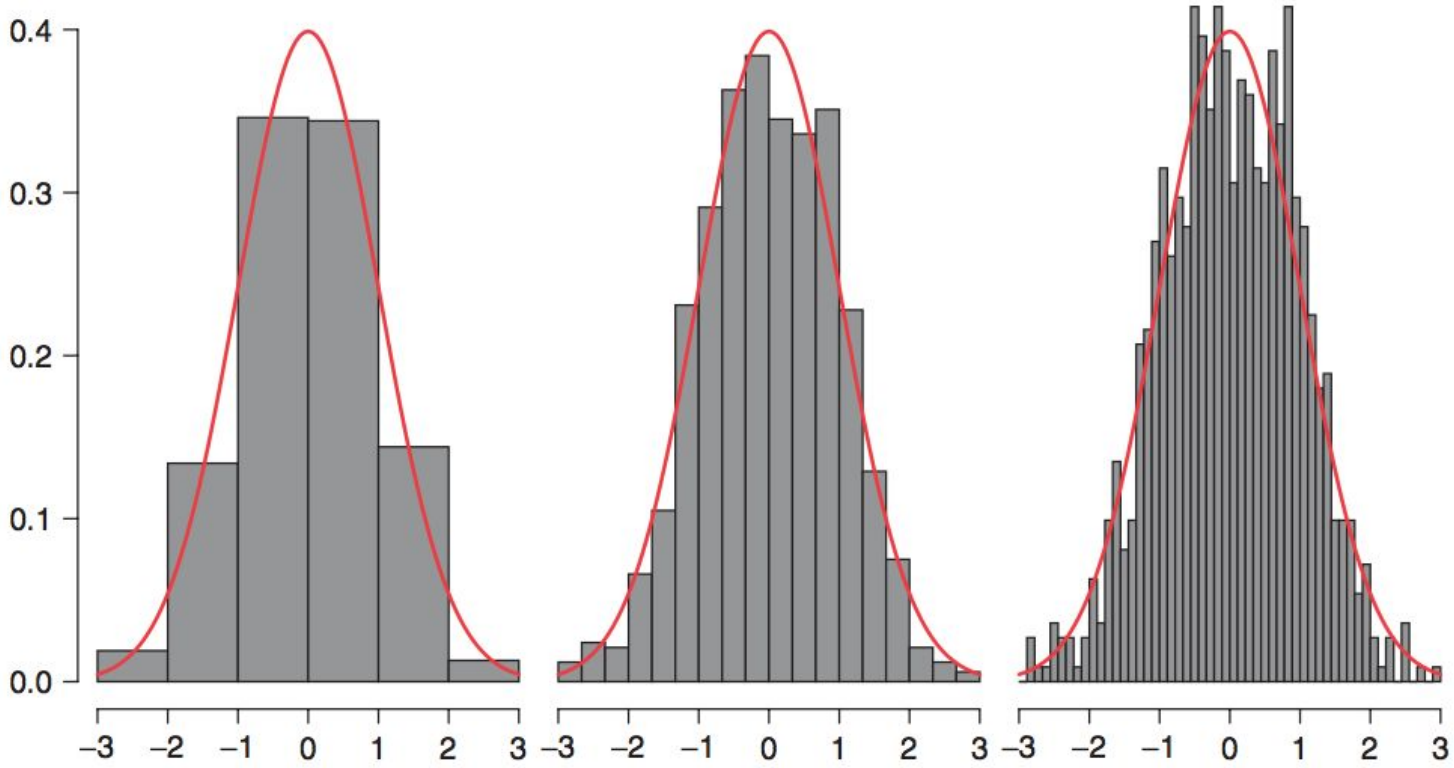
3, 5, 5, 8, 10

Estimation

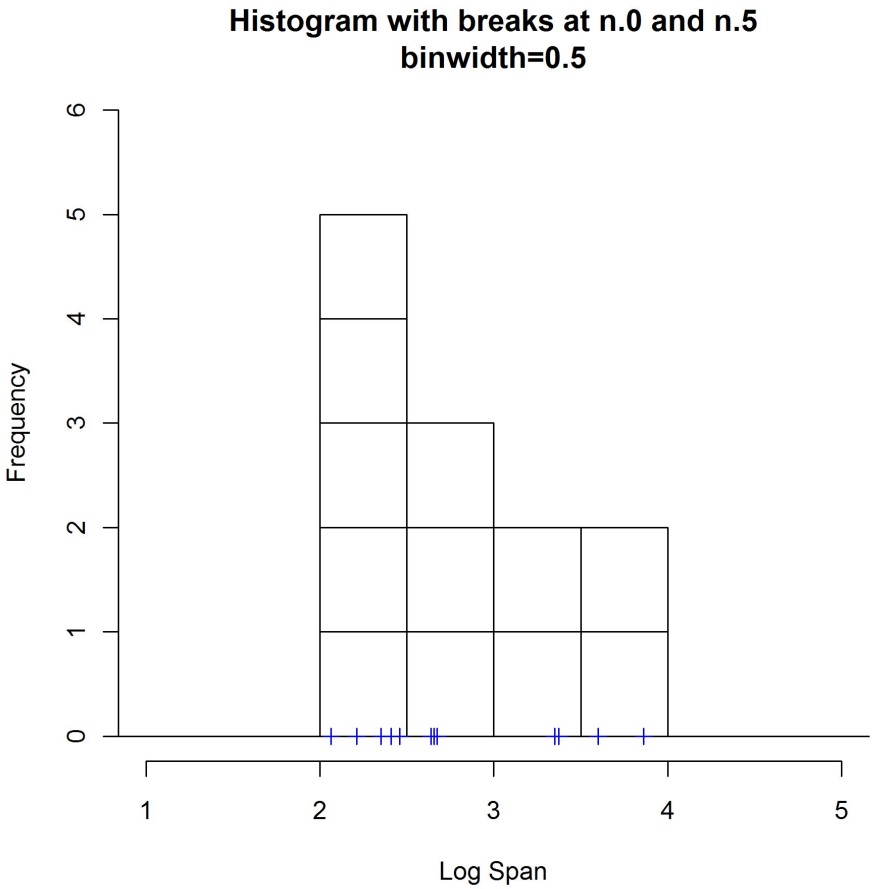


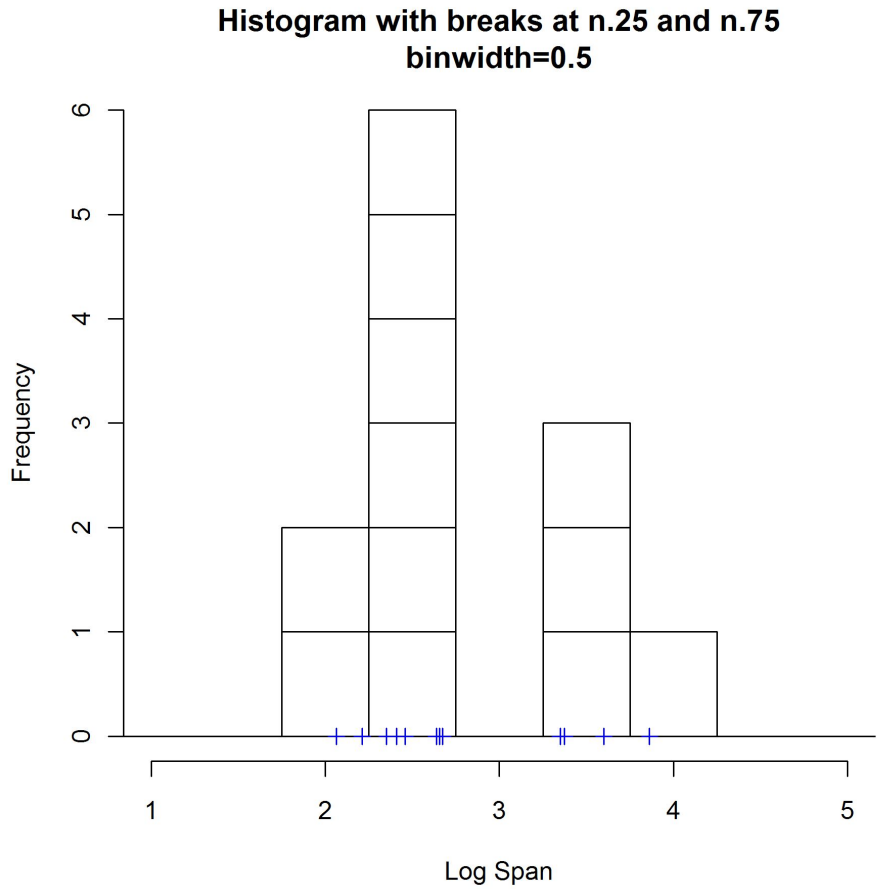
Simple

Transparent

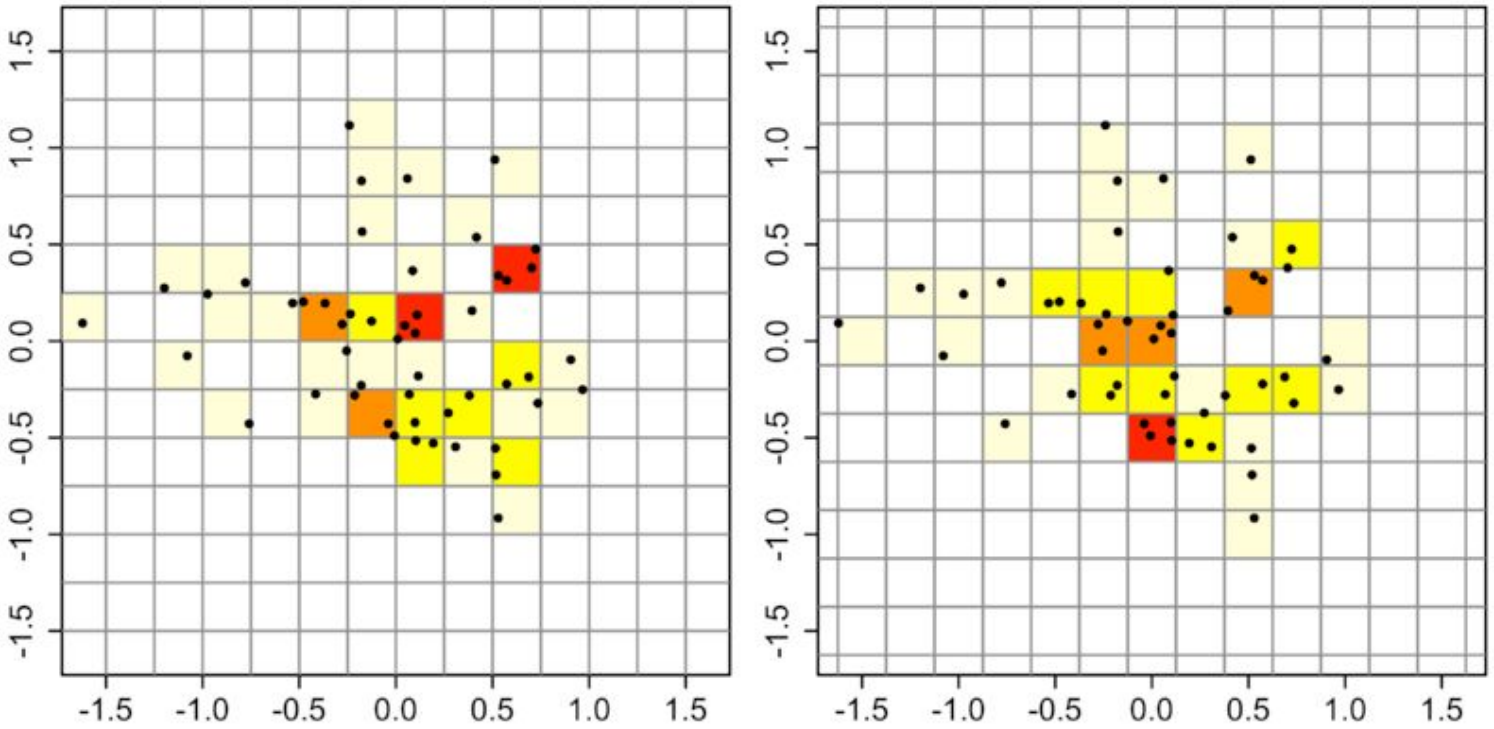


David W. Scott, “Scott’s Rule”, Wiley Interdisciplinary Reviews: Computational Statistics, **DOI:** 10.1002/wics.103



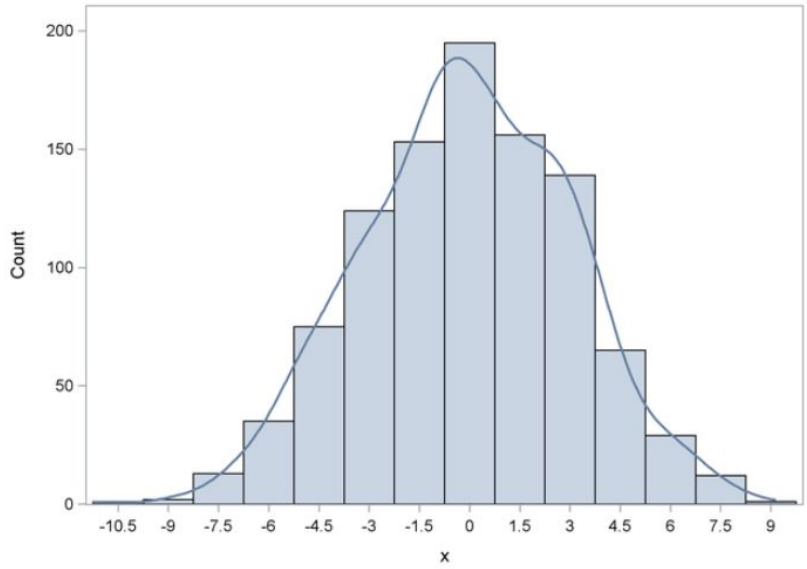


Tarn Duong, An introduction to kernel density estimation,
http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5

A histogram is a “model” that infers the underlying distribution of the data.

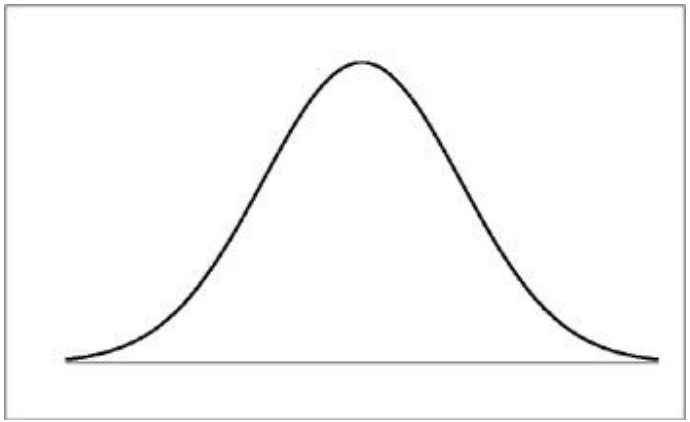
Can we directly infer the underlying (smooth) distribution from data? How can we do that?



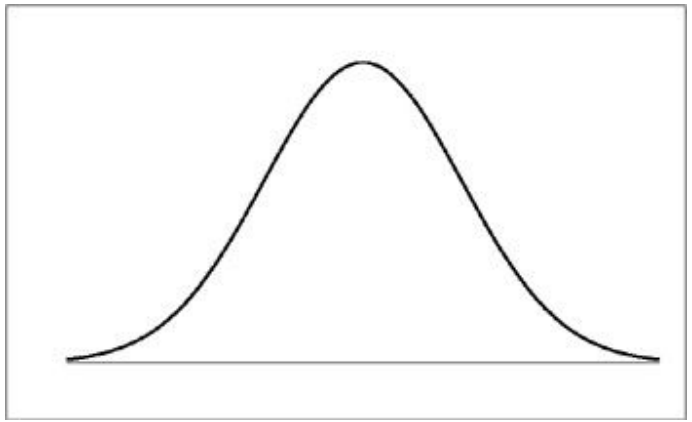
SAS User guide

1. Parametric approach

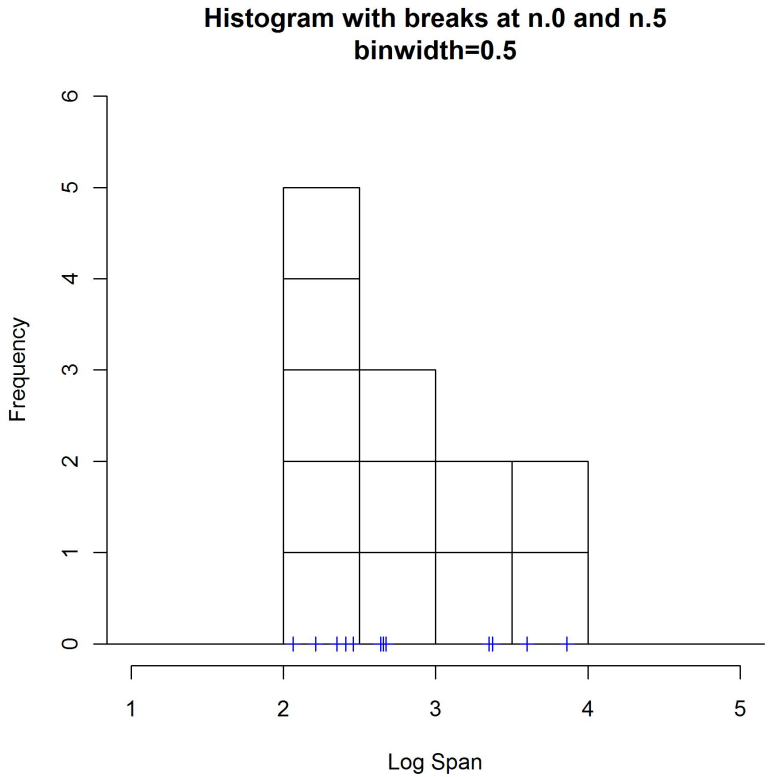
$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

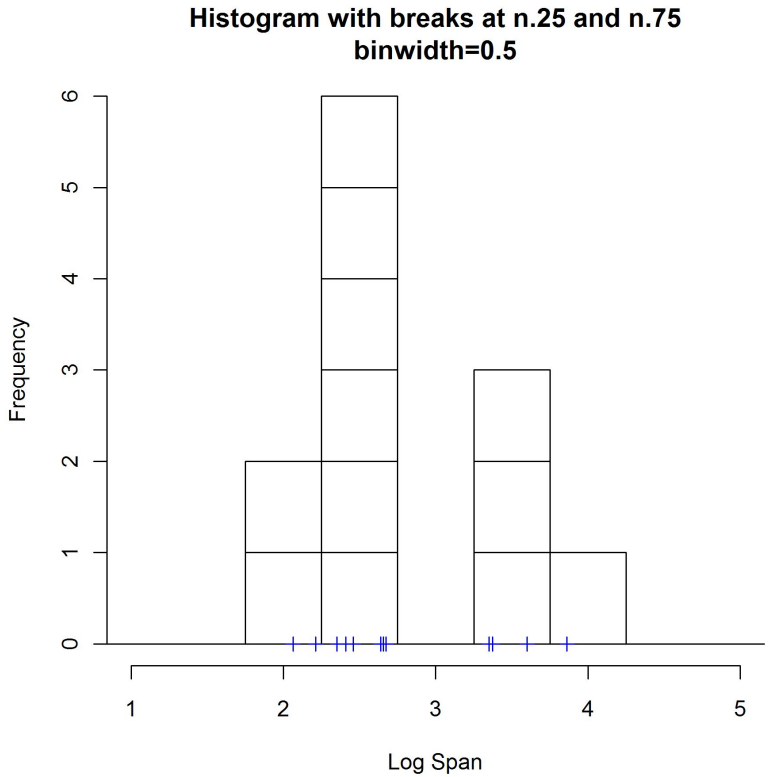


$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

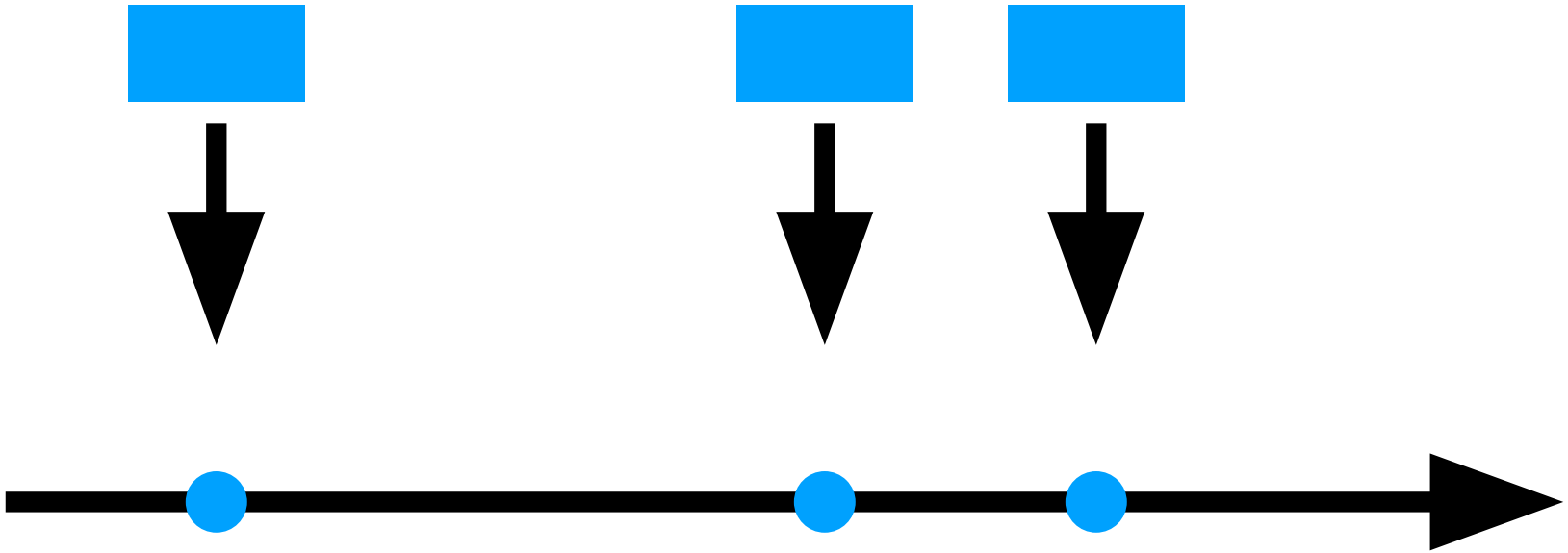


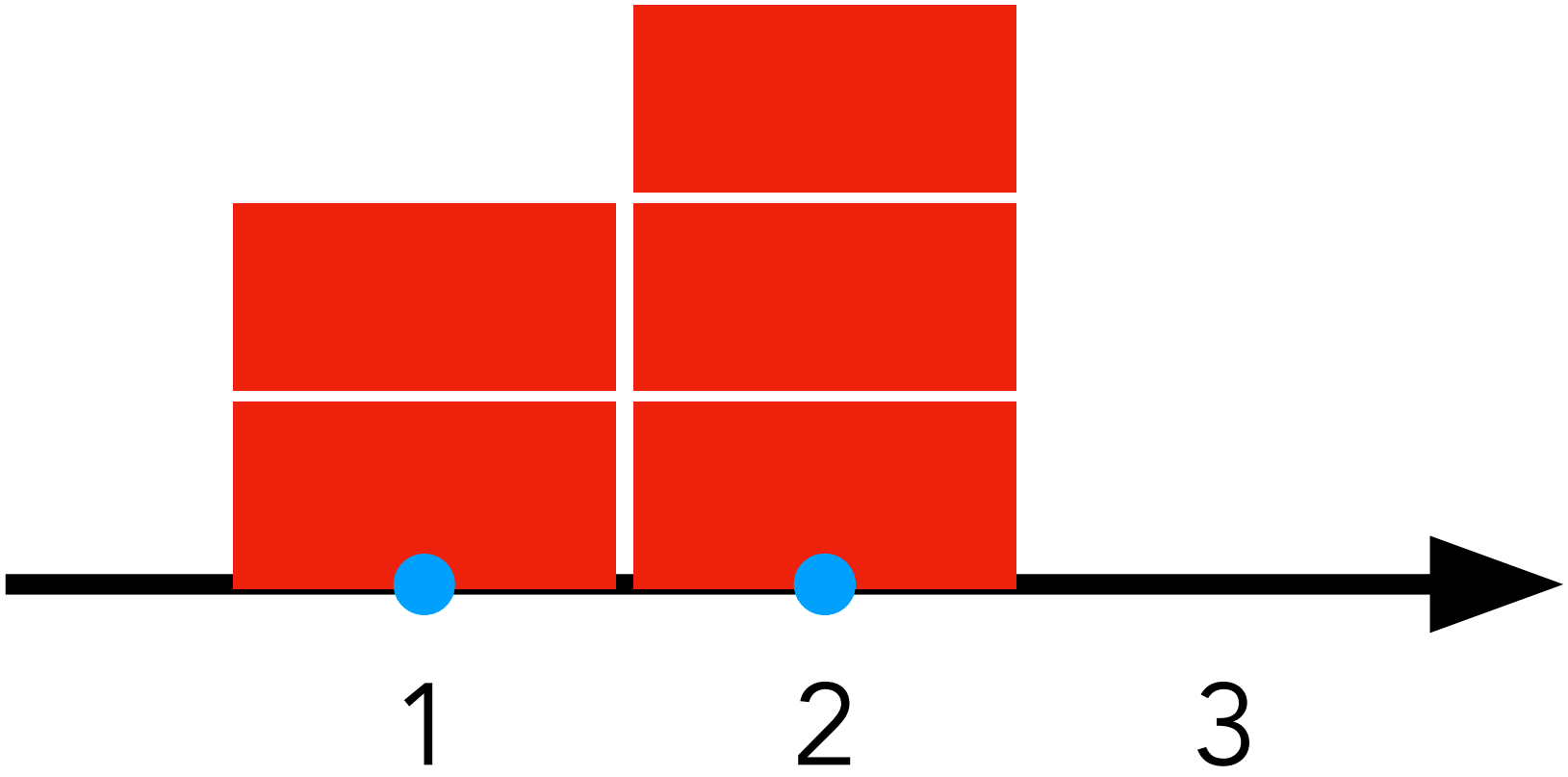
2. Non-parametric approach

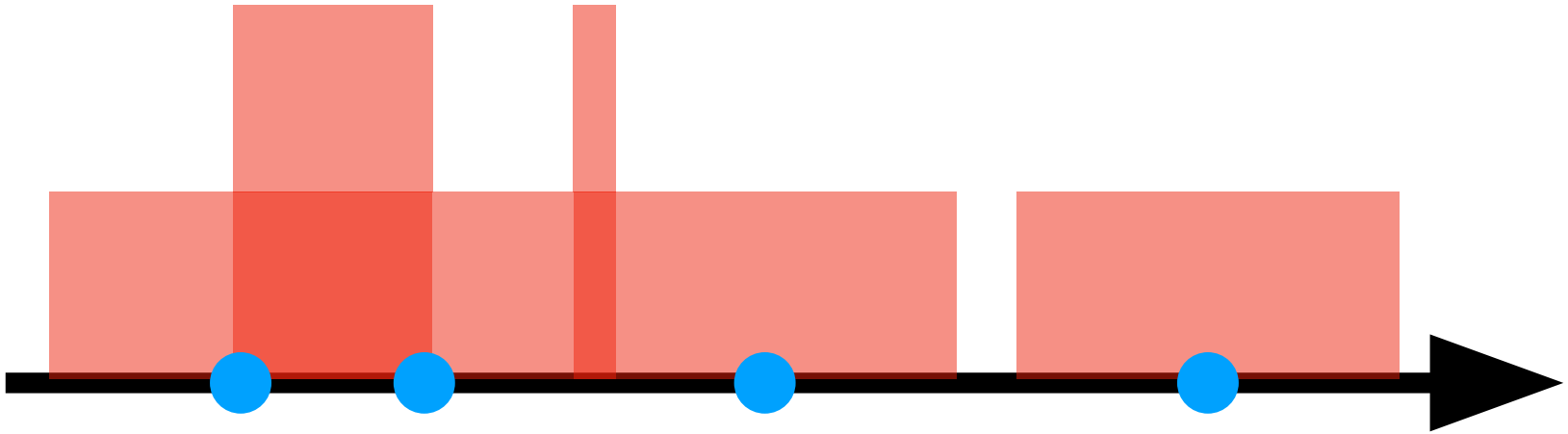


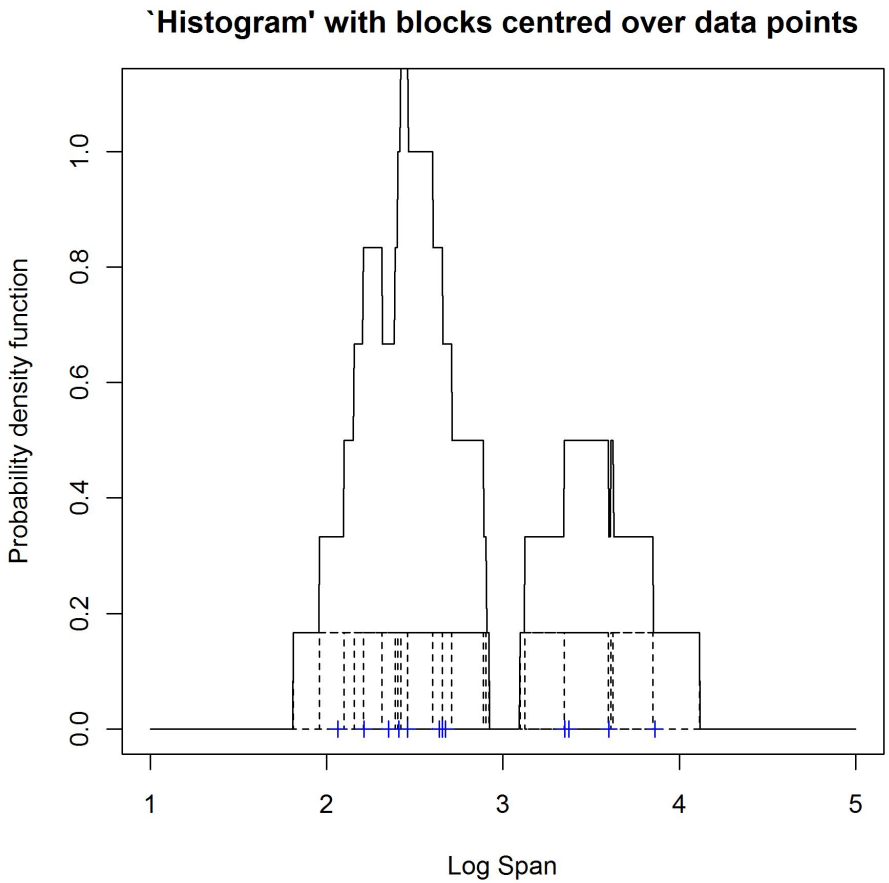


Tarn Duong, An introduction to kernel density estimation,
http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

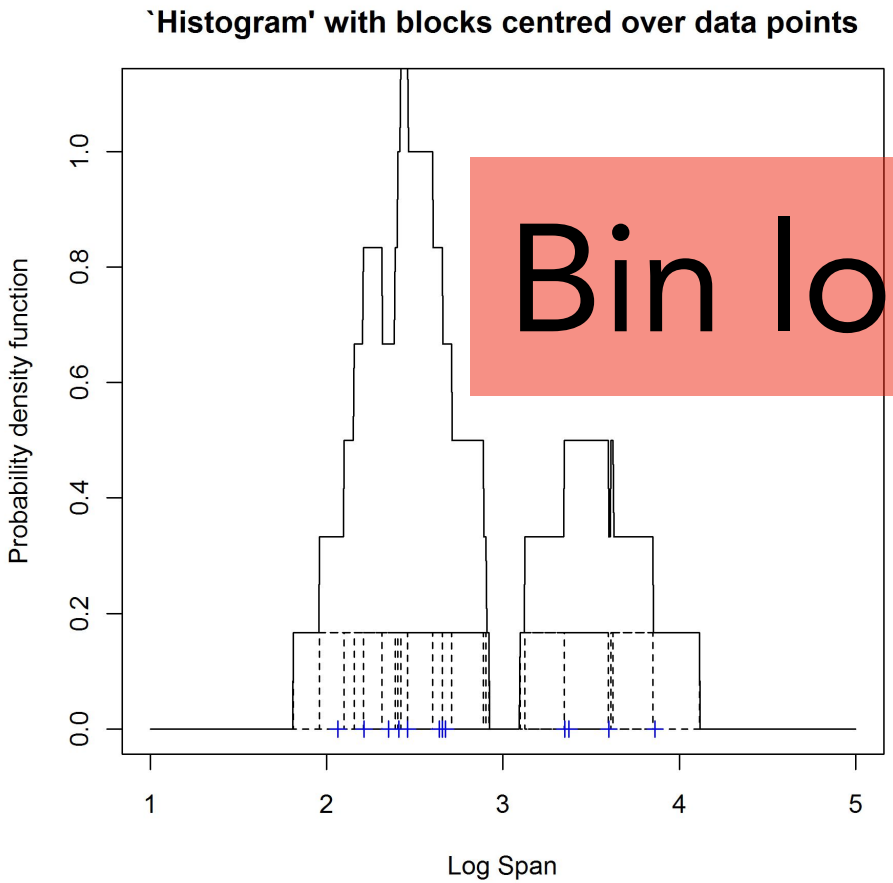



1
2
3



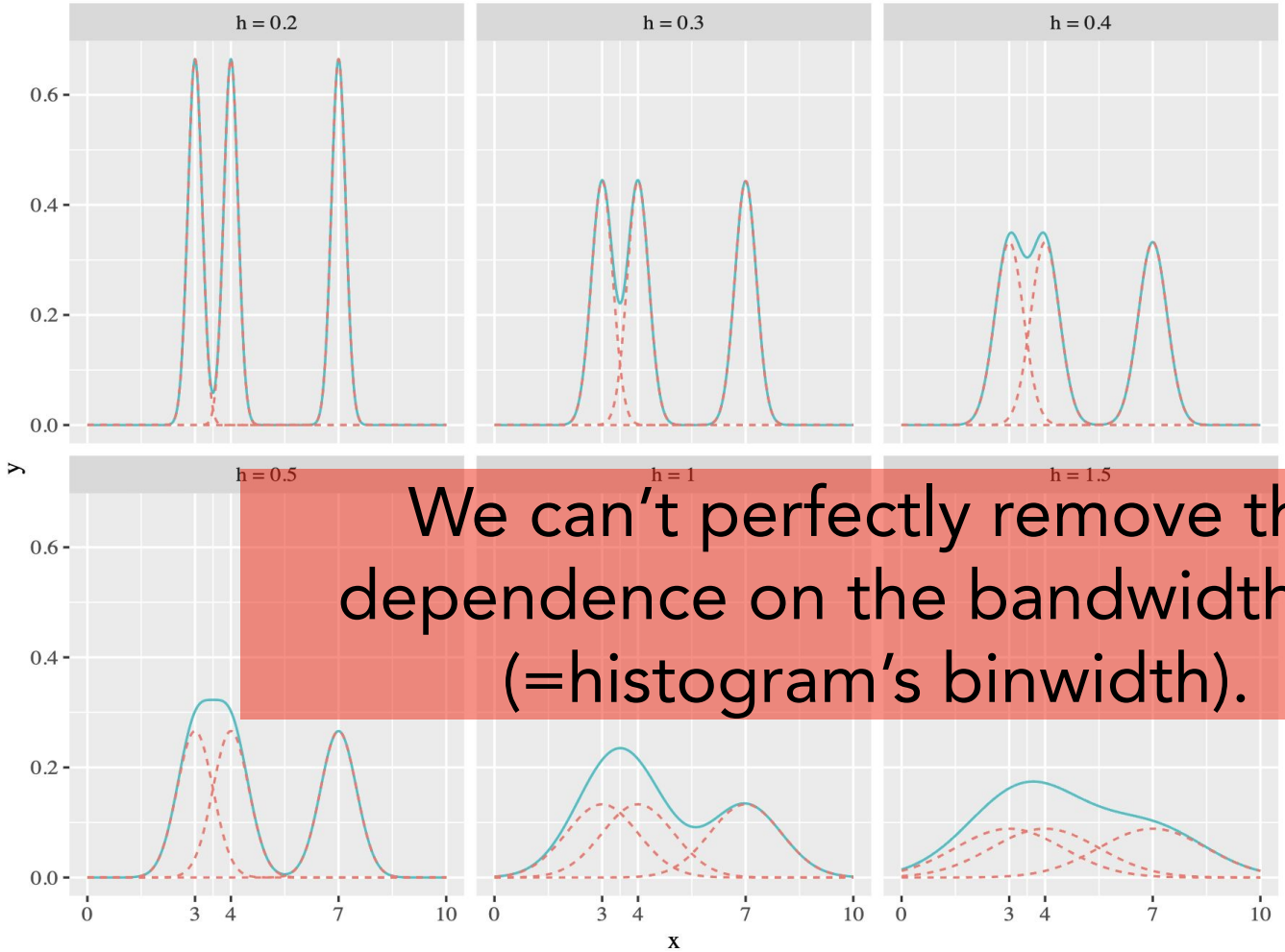


Tarn Duong, An introduction to kernel density estimation,
http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

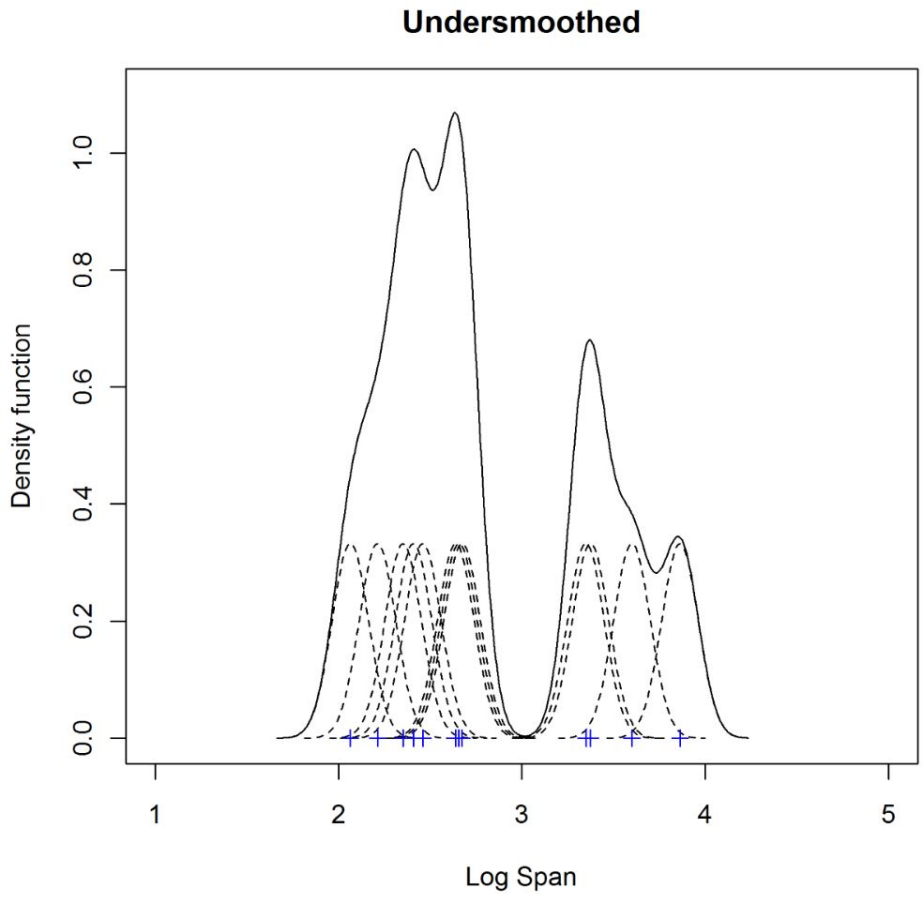


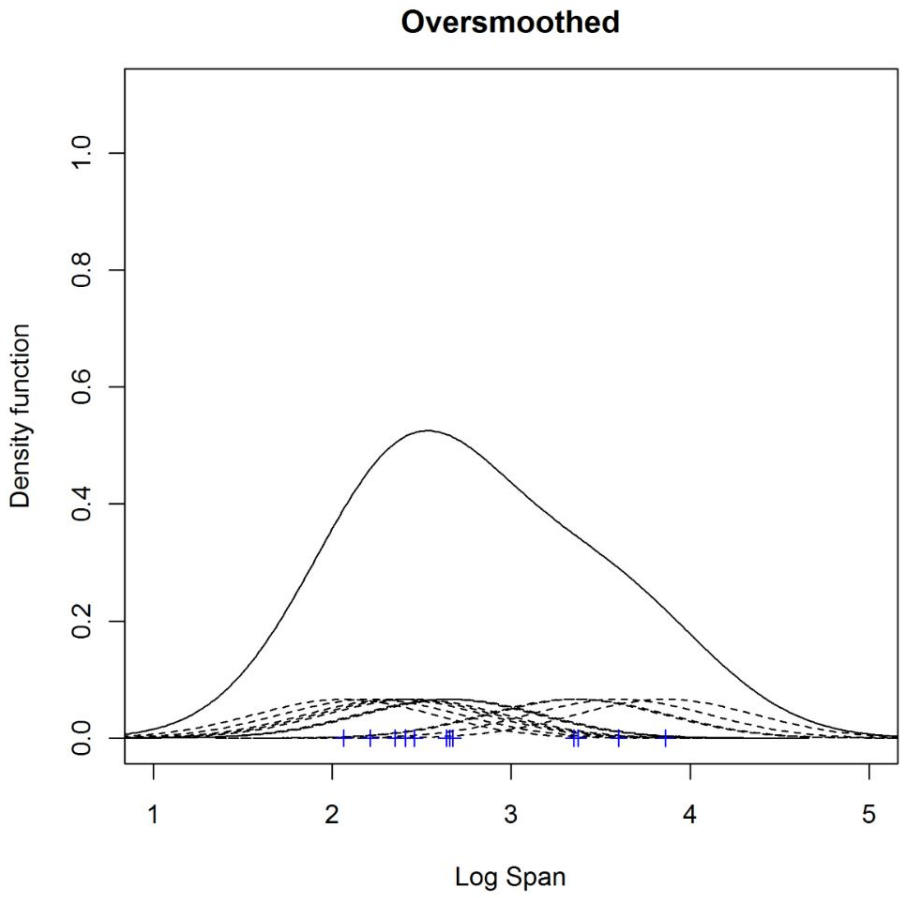
Bin location

Tarn Duong, An introduction to kernel density estimation,
http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

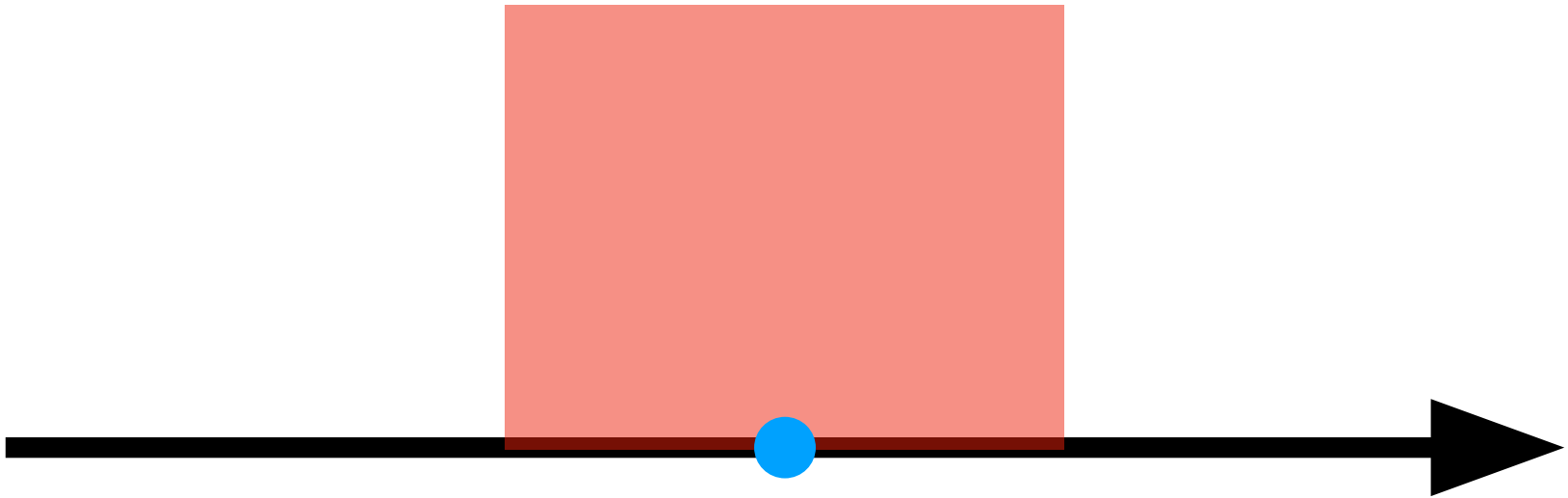


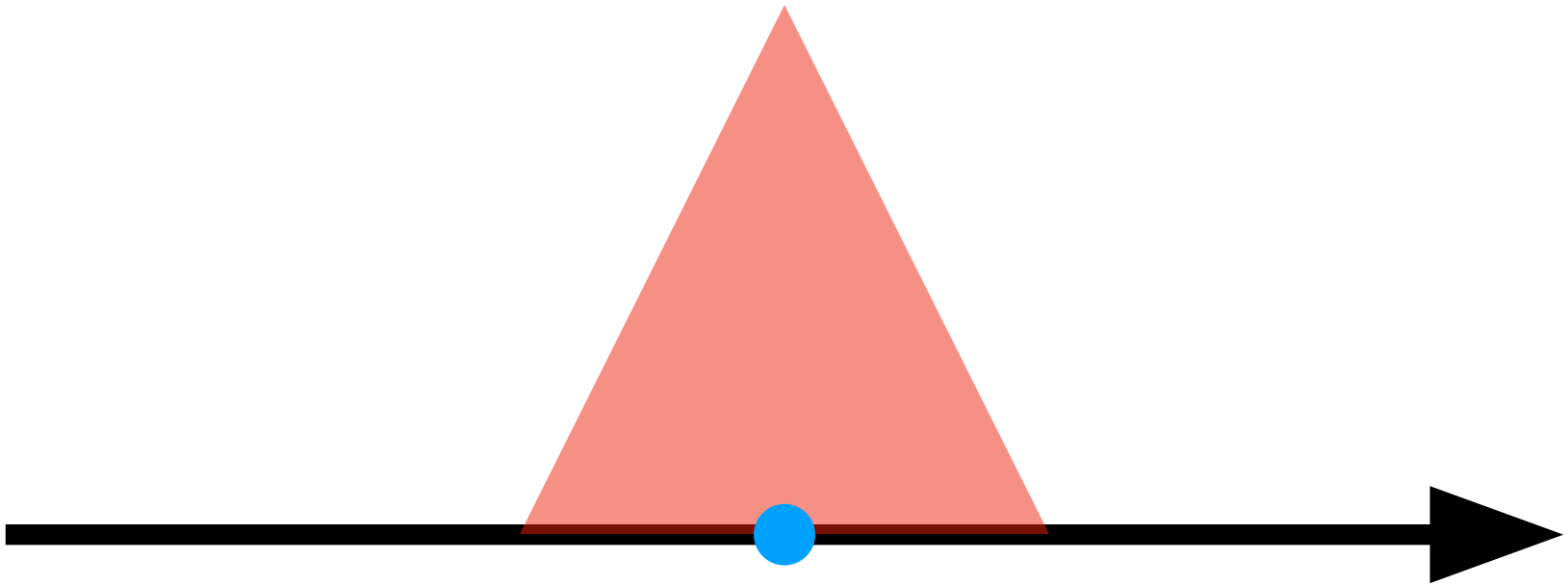
We can't perfectly remove the dependence on the bandwidth yet (=histogram's binwidth).

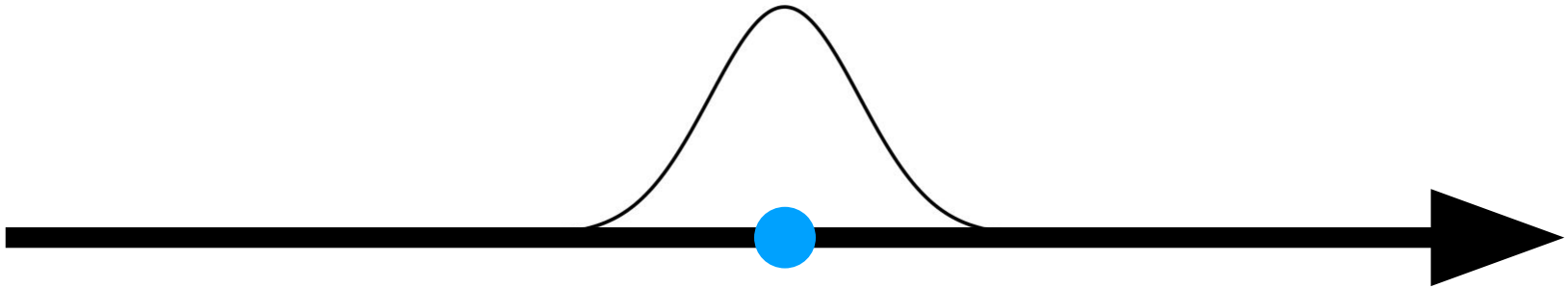

Undersmoothed
Density function
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Log Span

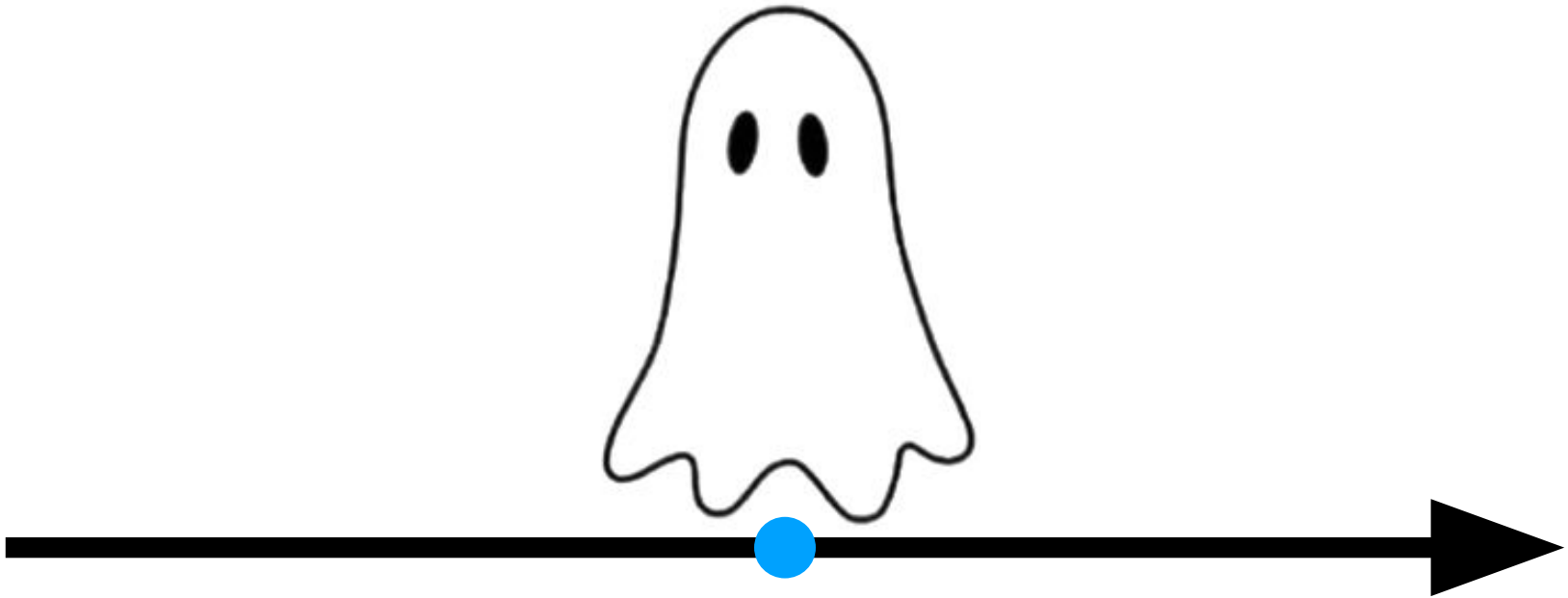

Oversmoothed
Density function
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Log Span

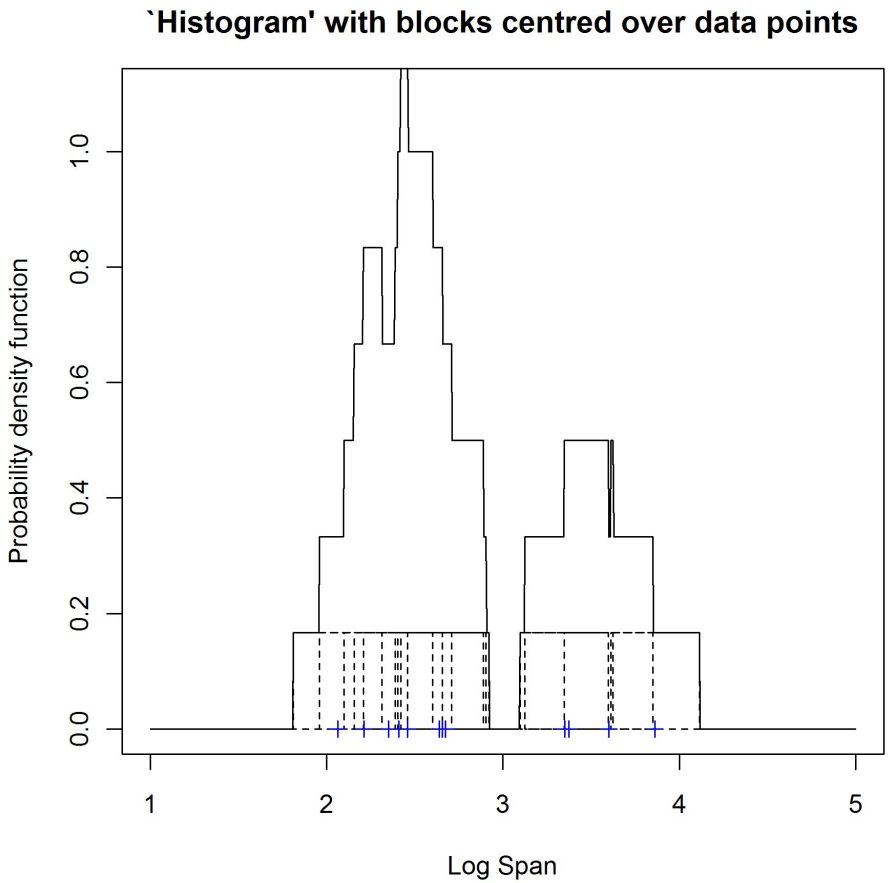
Less arbitrary
More data-driven



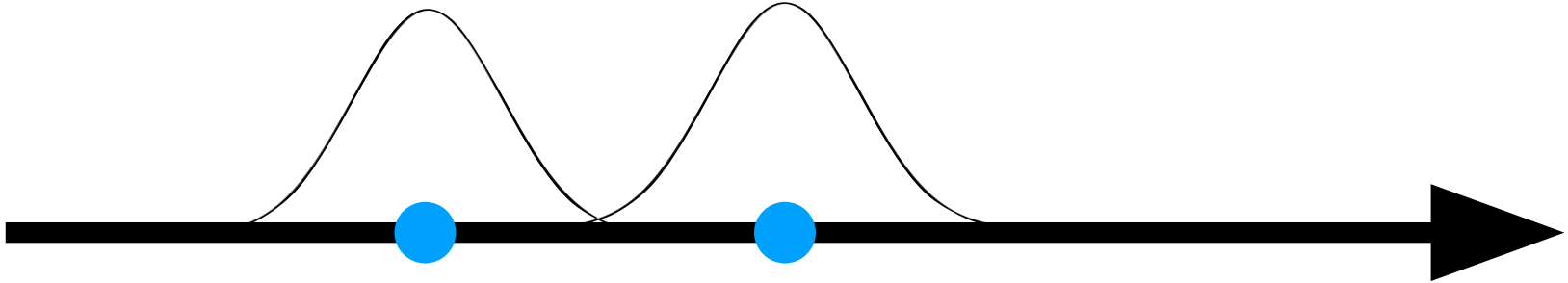




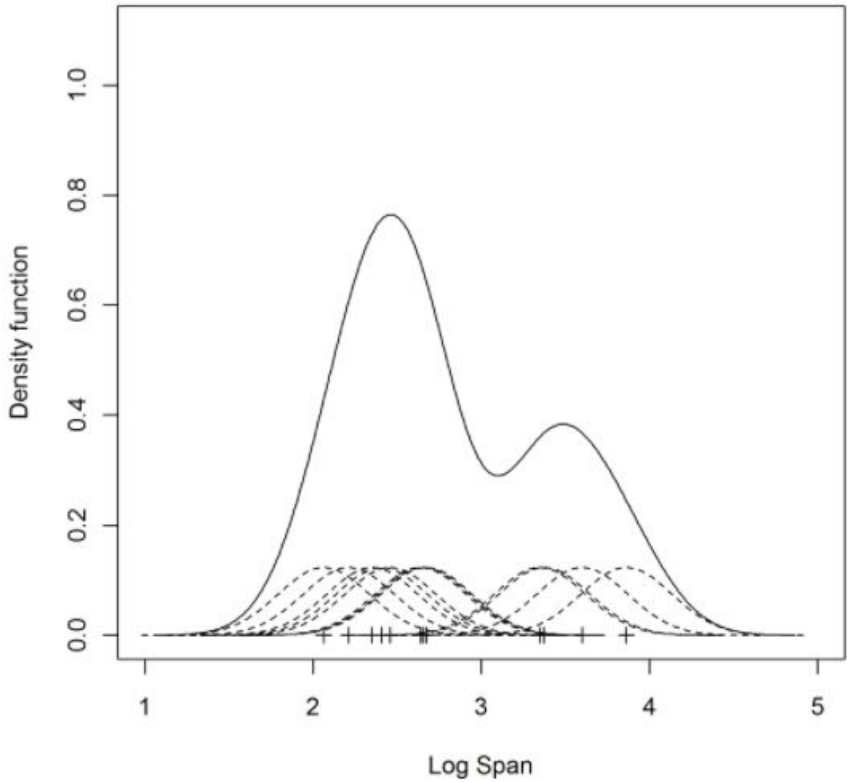




Tarn Duong, An introduction to kernel density estimation,
http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/



Kernel Density Estimation (KDE)



Tarn Duong, An introduction to kernel density estimation, http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

"Kernels"