

Data Visualization

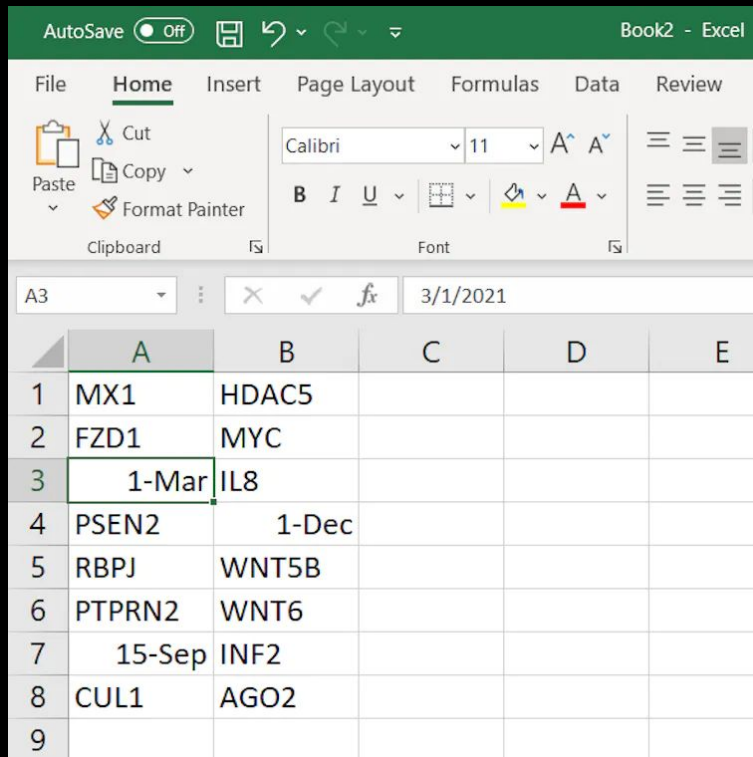
Project idea presentation (10 Oct, in class)

- Make your team by next Tue and update the sheet
- Check “Proposal” in [prof. yy's project page](#) for what you need to include in your presentation (Intro, Questions or objectives, Datasets and methods, References)
- We will have a presentation session without a report submission (but submit your presentation slides by 9 Oct)
- 6 minutes per team (elevator pitch!)

Quiz

- What do you find interesting in today's VotW?
- What are reasons that you don't want to manually edit data files?
- Explain data provenance. What is it? Why is it important?
- 1-d scatterplot (or "strip chart") has a big limitation, especially when there are many data points. What is it? List at least 5 different ways to mitigate the issue.
- What's good about the strip chart? When would you use it?

You've just found some typos in a dataset. What would you do?



The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The ribbon includes options for File, Home, Insert, Page Layout, Formulas, Data, and Review. The 'Clipboard' group shows Cut, Copy, and Paste options. The 'Font' group shows the font name 'Calibri', size '11', and various formatting options like Bold, Italic, Underline, and text color. The active cell is A3, which contains the text '1-Mar'. The formula bar shows the date '3/1/2021'. The worksheet contains a table with 9 rows and 5 columns (A-E). The data in the table is as follows:

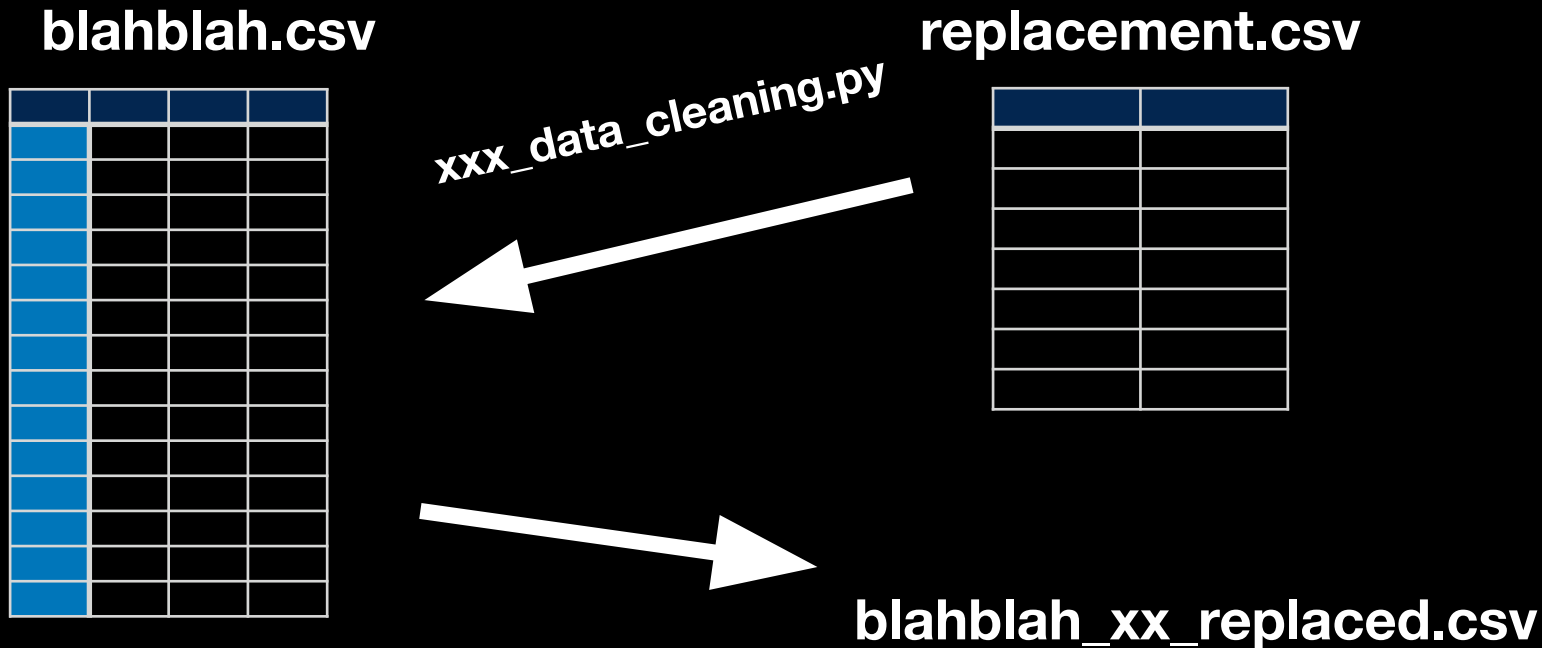
	A	B	C	D	E
1	MX1	HDAC5			
2	FZD1	MYC			
3	1-Mar	IL8			
4	PSEN2	1-Dec			
5	RBPJ	WNT5B			
6	PTPRN2	WNT6			
7	15-Sep	INF2			
8	CUL1	AGO2			
9					

Ok, let me quickly fix them by hand... 

What could be problems?

1. What if you introduce a different error?
2. What if you re-download the raw data?
3. What if you have to explain the process but you can't remember?
4. What if it breaks the pipeline and you can't remember exactly what you fixed?
5. What if someone else takes over your job?

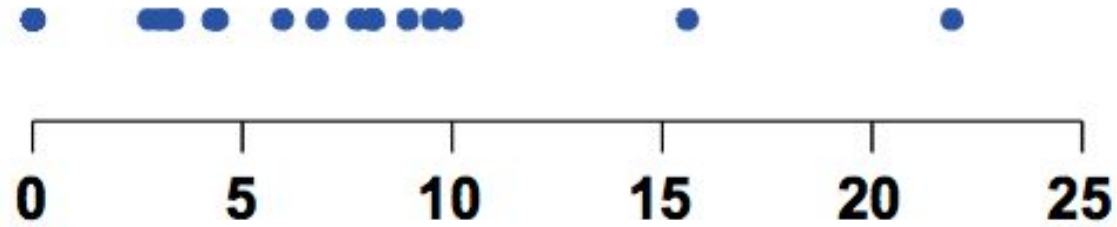
Never handle your data manually & Be explicit!



Data Provenance

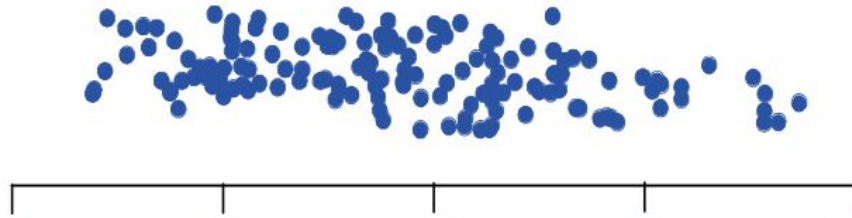
“a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.”

(Or the *feasibility* to trace back any dataset to its origin)



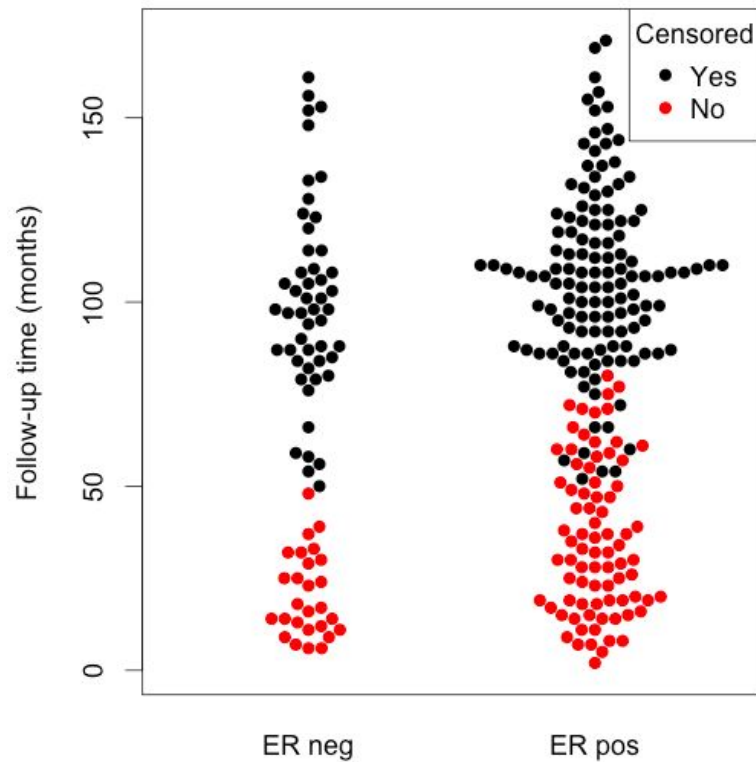
Occlusion problem

1-d “Jittered” Scatterplot

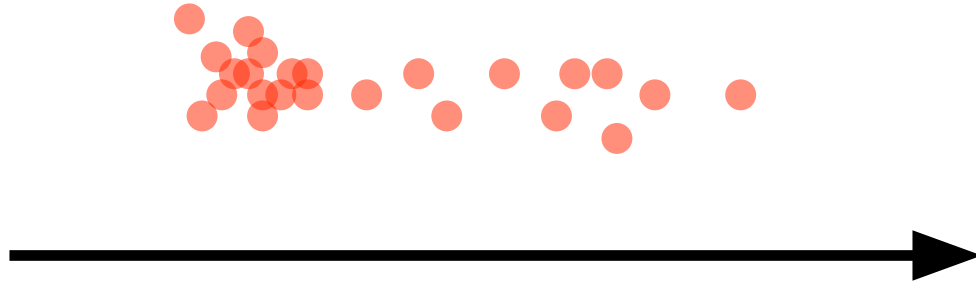


Problem?

"Beeswarm"

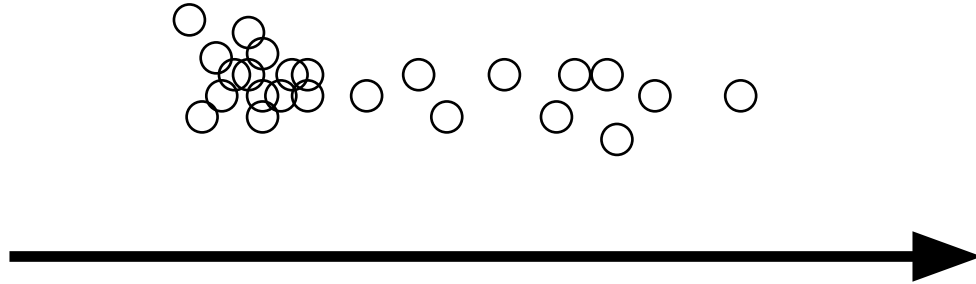


using “alpha”

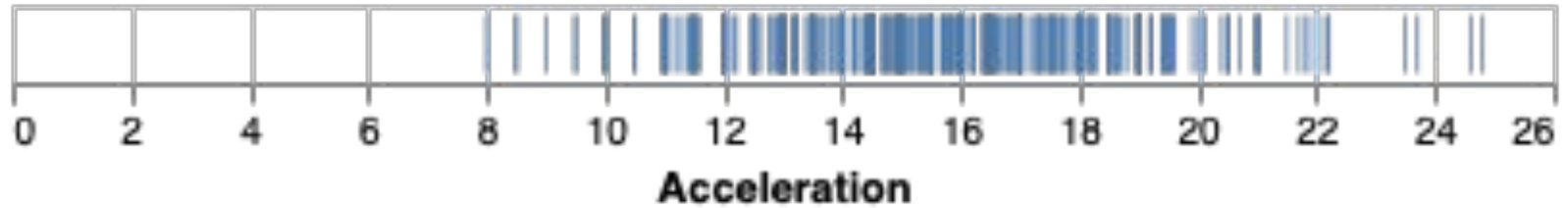


Problem?

Using empty symbols



Use lines



simplystatistics.org/posts/2019-02-21-open-letter-to-journal-editors-dynamite-plots-must-die

Open letter to journal editors: dynamite plots must die

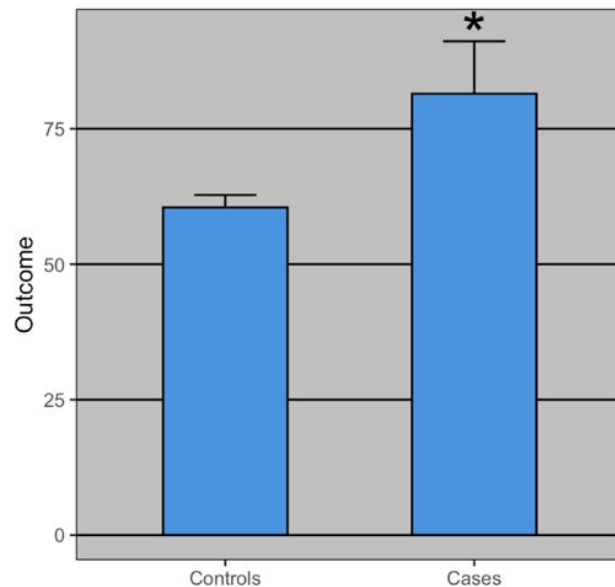
AUTHOR
Rafael Irizarry

PUBLISHED
Feb. 21, 2019

Statisticians have been pointing out the problem with dynamite plots, also known as bar and line graphs, for years. Karl Broman lists them as one of the [top ten worst graphs](#). The problem has even been documented in the peer reviewed literature. For example, [this British Journal of Pharmacology](#) paper titled *Show the data, don't conceal them* was published in 2011.

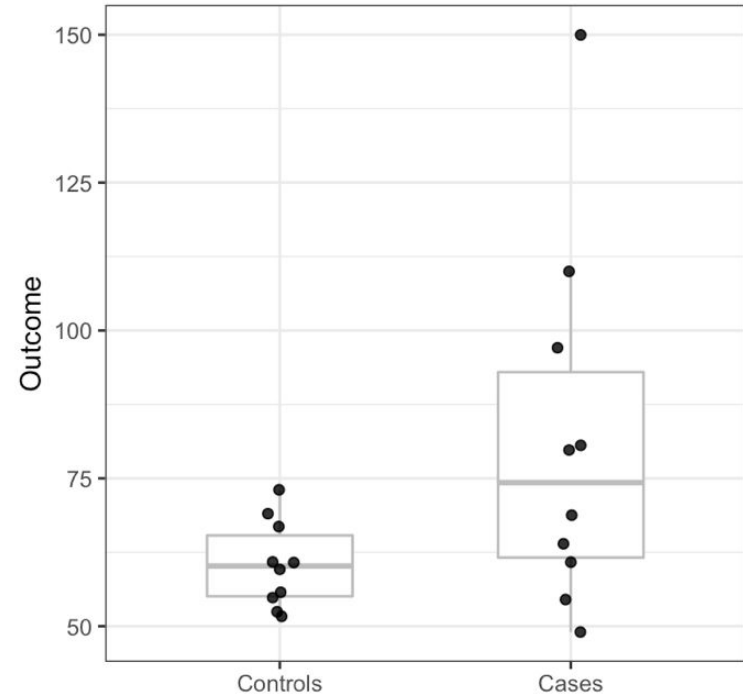
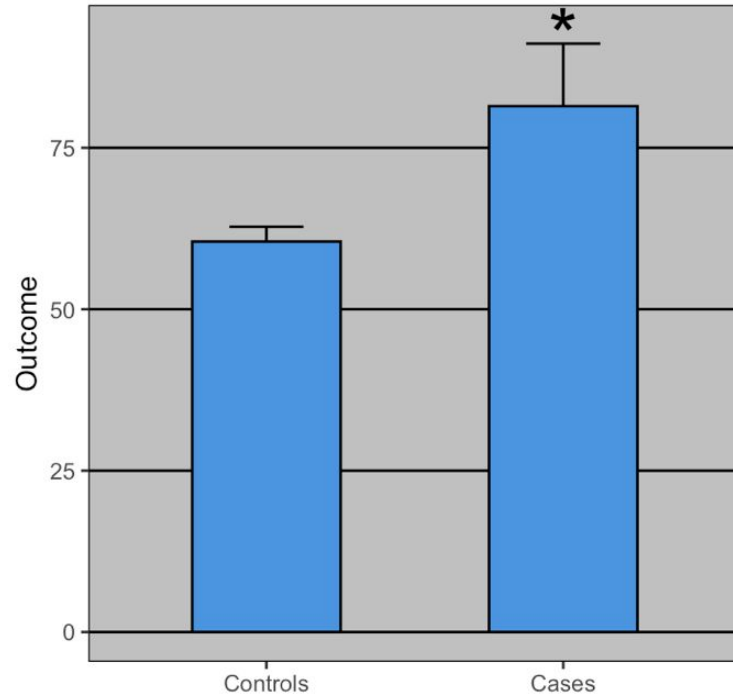
However, despite all these efforts, dynamite plots continue to be ubiquitous in the scientific literature. Just open the latest issue of Nature, Science or Cell and you will likely see a few. In fact, in this [PLOS Biology paper](#), Tracey Weissgerber and co-authors perform a systematic review of "top physiology journals" and find that "85.6% of papers included at least one bar graph". They go on to recommend "training investigators in data presentation, encouraging a more complete presentation of data, and changing journal editorial policies". In my view, the training will be accelerated if editors implement a policy that requires authors to show the data or, if the dataset is too large, show the distribution of the data with

	x	average	se
1	Controls	60	2.3
2	Cases	81	9.7

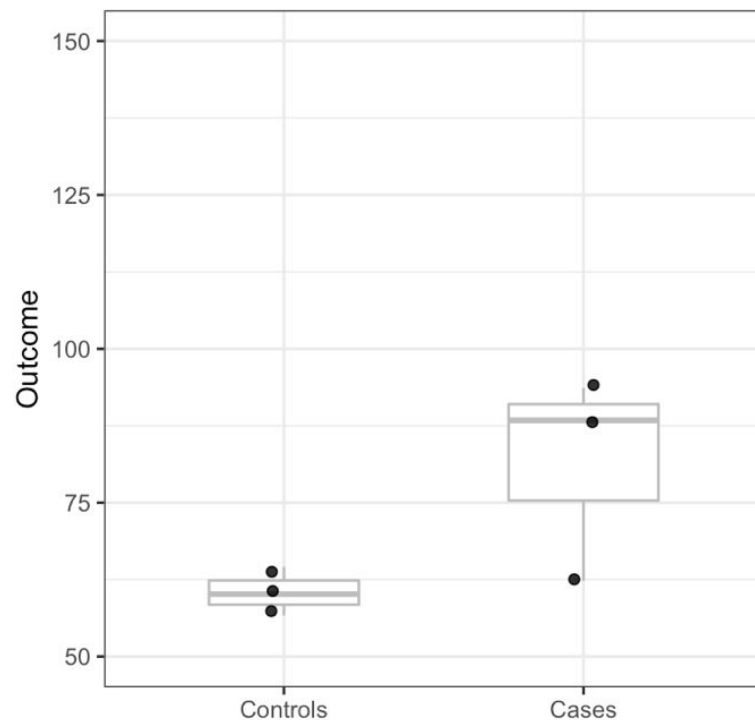
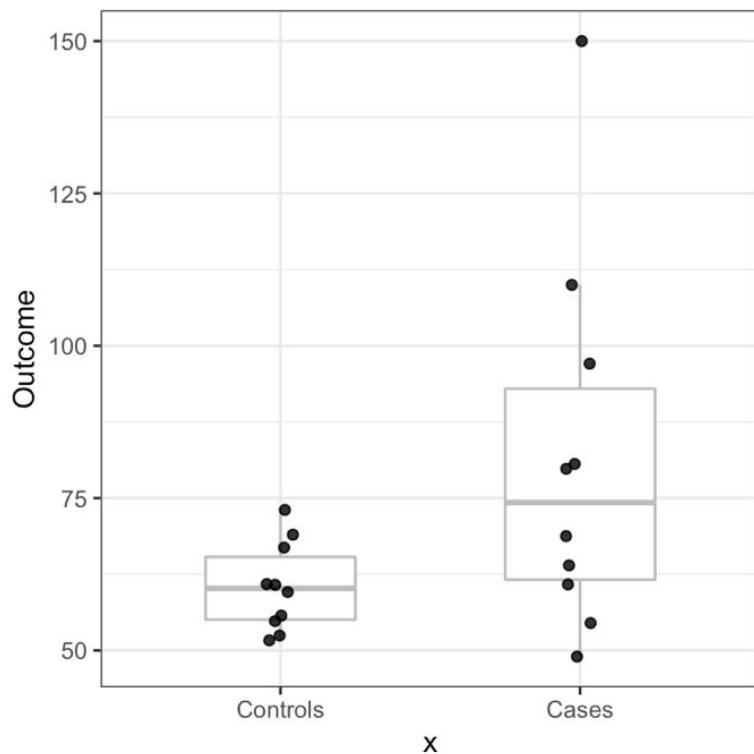


Second, the dynamite plot makes it appear as if there is a clear difference between the two groups.

Showing the data reveals more information. In our example, showing the data reveals that the lowest blood pressure is actually in the treatment group. It also reveals the presence of one somewhat extreme value of 150. This might represent a data entry mistake. Perhaps systolic pressure was recorded by accident? Note that without that data point, the difference is no longer significant at the 0.05 level.

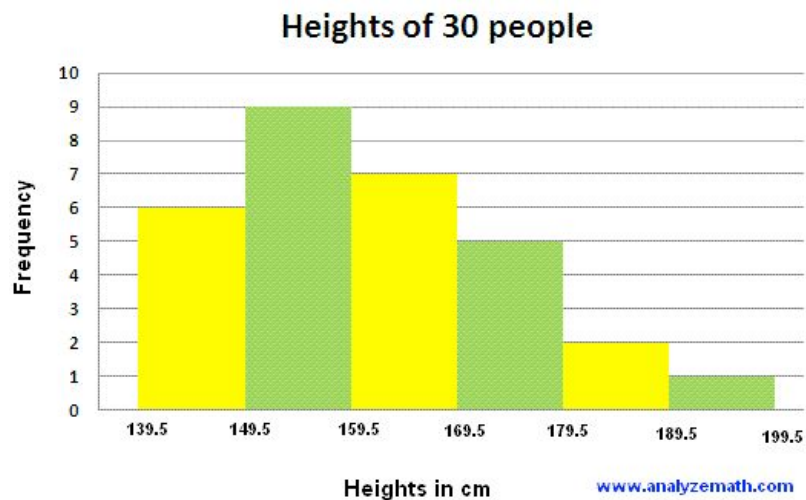
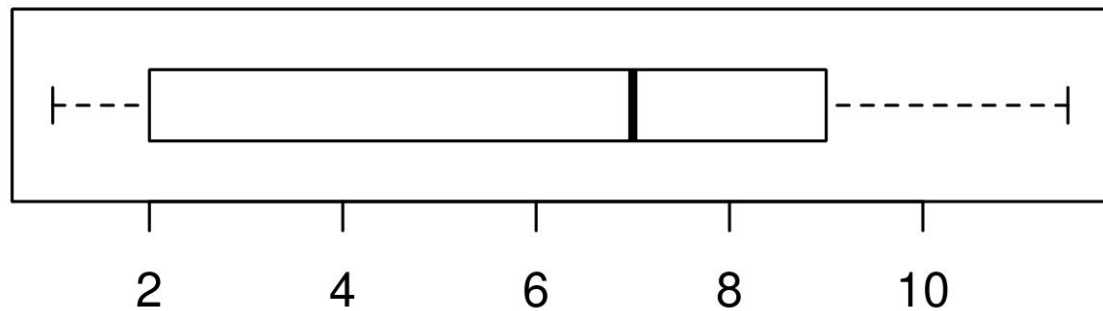


Note also that, as pointed out by Weissgerber, data that look quite different can result in exactly the same barplot. For instance, the two datasets below would produce the same barplot as the one shown above.

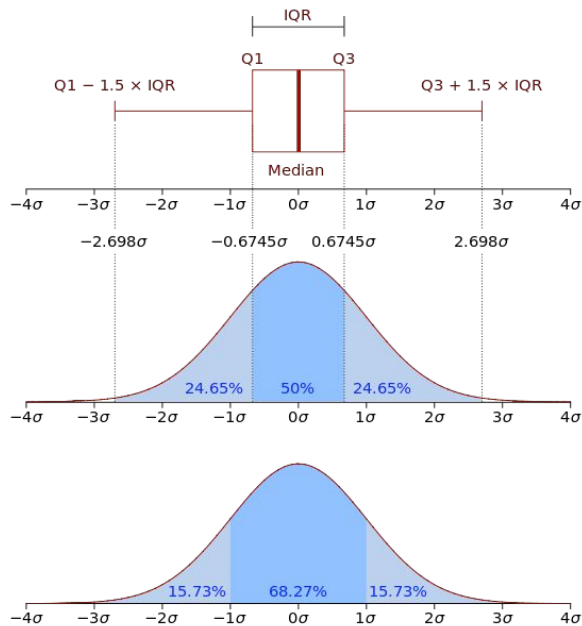


When possible, show your
data directly!

Summarization vs.
aggregation



Let's construct a box plot



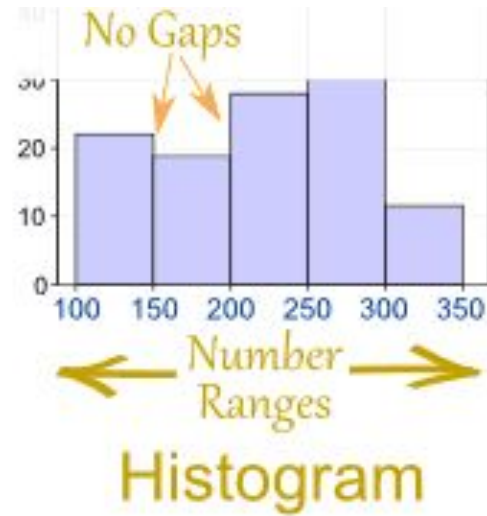
1. We want to know the **median** (the central value in the data).
2. What's the range that capture about the **half of the data points**?
3. How about **most of the data points**? (Many ways to estimate this!)
4. What do we do with the **rest of the data points**?

Histogram

Draw a histogram

Score	Frequency
0-199	5
200-399	29
400-599	56
600-799	17
800-999	3

Bar graph vs. histogram



The semantics of a bar chart

Show the quantity corresponding to each category

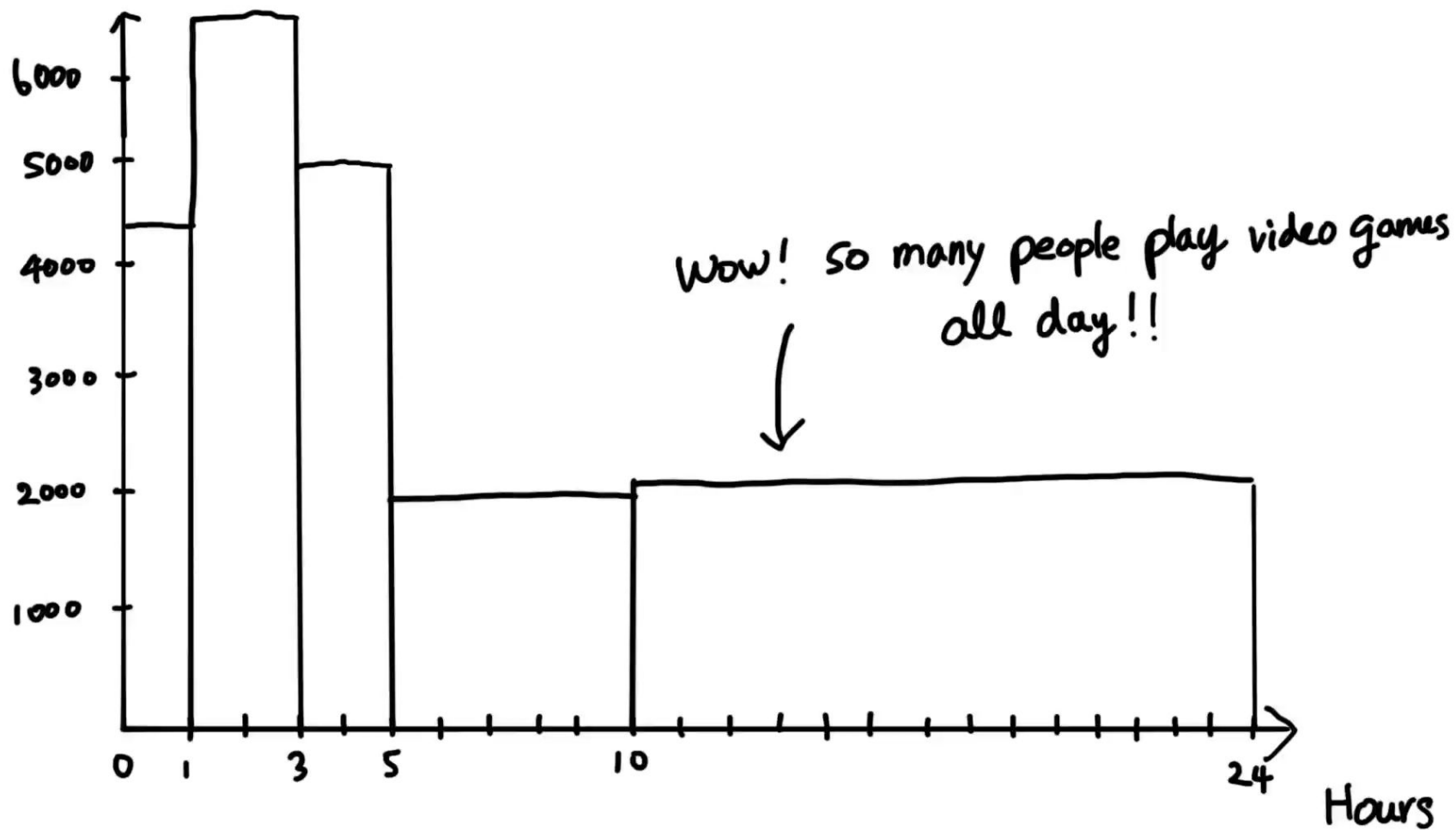
The semantics of a histogram

divide a (continuous) data range into bins
and show the data density in each bin

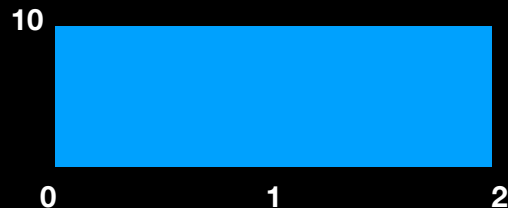
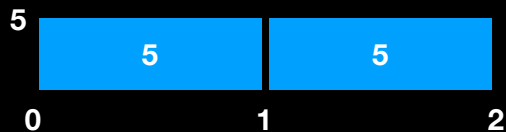
Draw a histogram



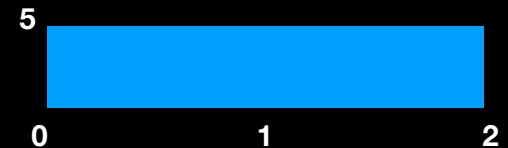
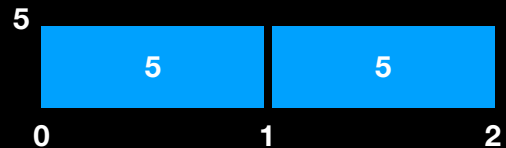
Hours	Frequency
0-1	4,300
1-3	6,900
3-5	4,900
5-10	2,000
10-24	2,100

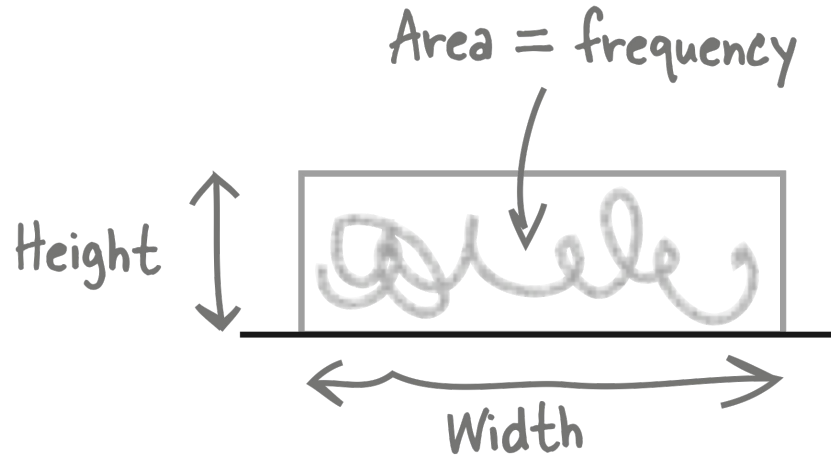


Frequency as height



Frequency as area

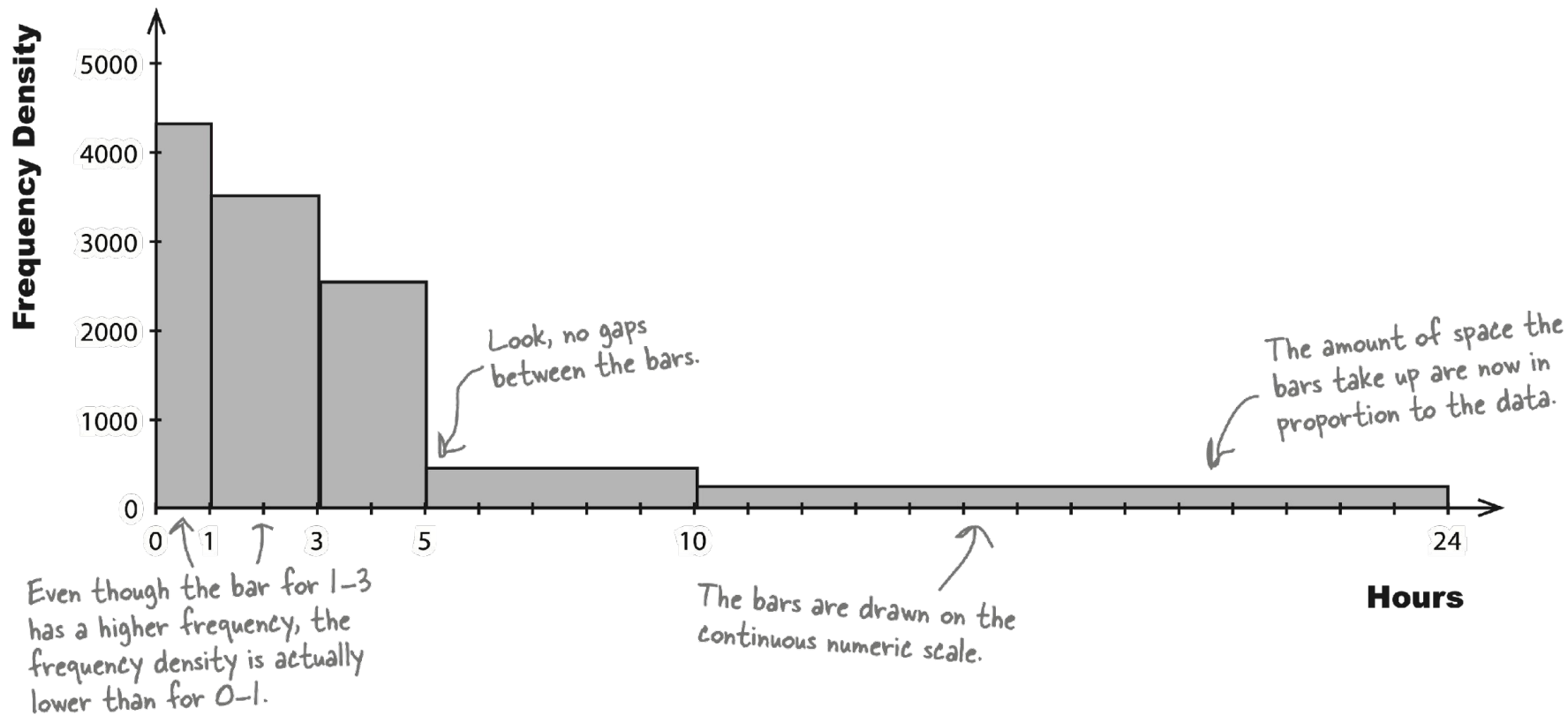




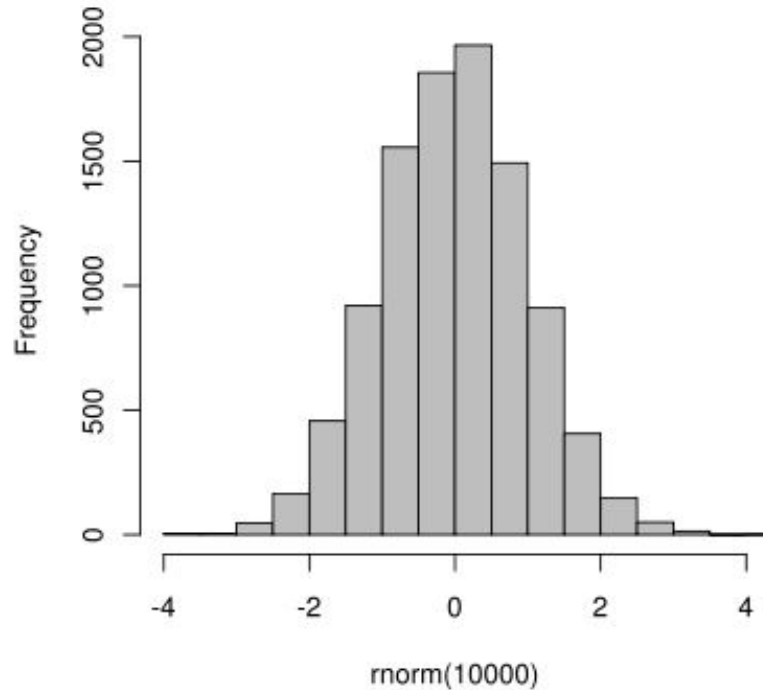
Hours	Frequency	Width	Height (Frequency Density)
0-1	4,300	1	$4,300 \div 1 = 4,300$
1-3	6,900	2	$6,900 \div 2 = 3,450$
3-5	4,900	2	$4,900 \div 2 = 2,450$
5-10	2,000	5	$2,000 \div 5 = 400$
10-24	2100	14	$2,100 \div 14 = 150$

Head First Statistics

Hours Spent Gaming per Day



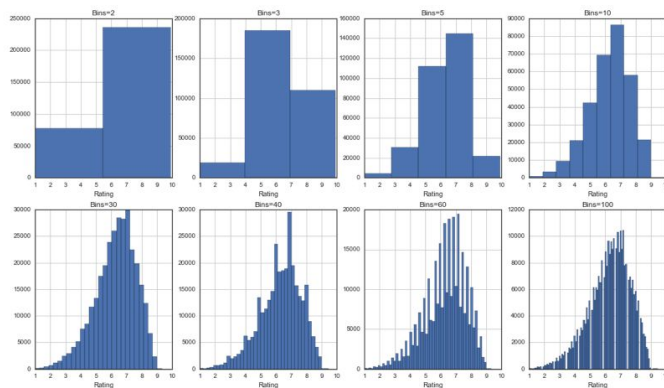
What kinds of choices do you need to make when you draw a histogram?

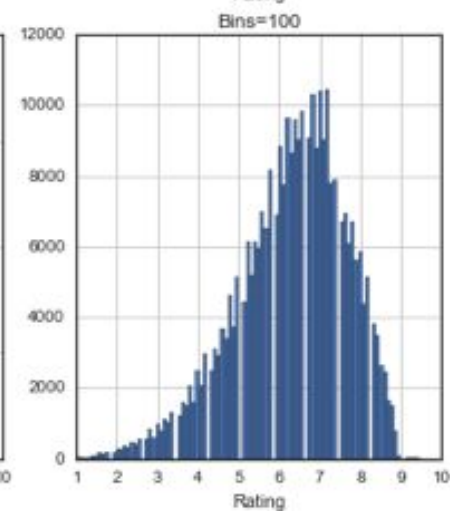
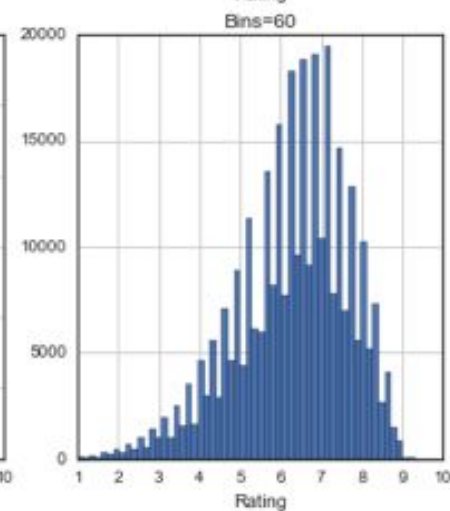
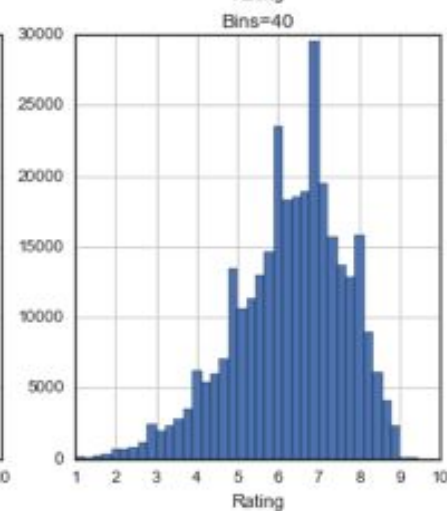
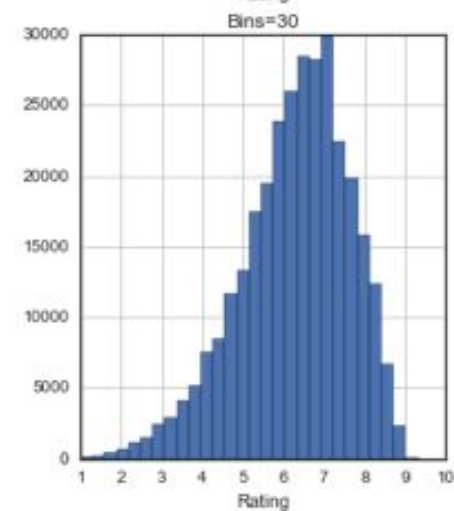
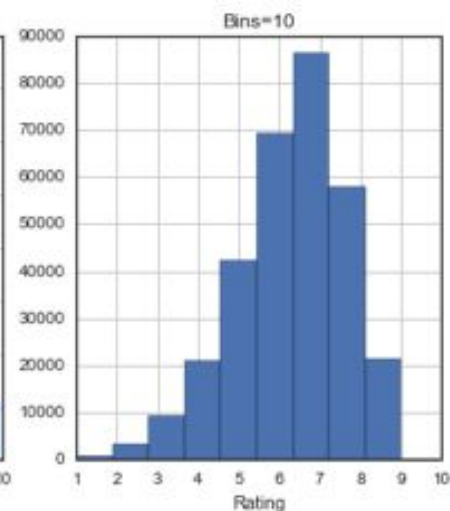
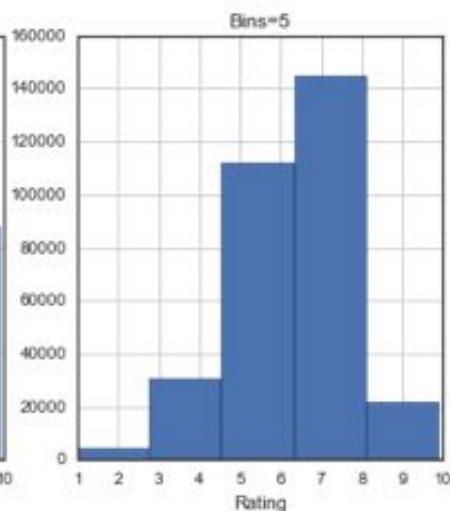
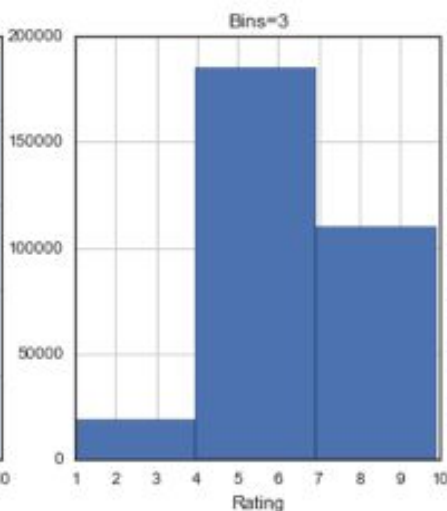
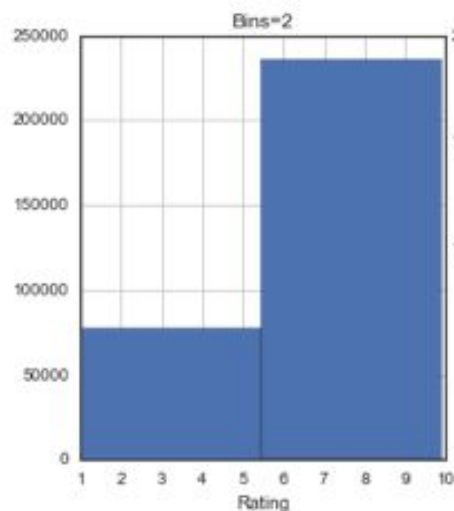


What are the problems of

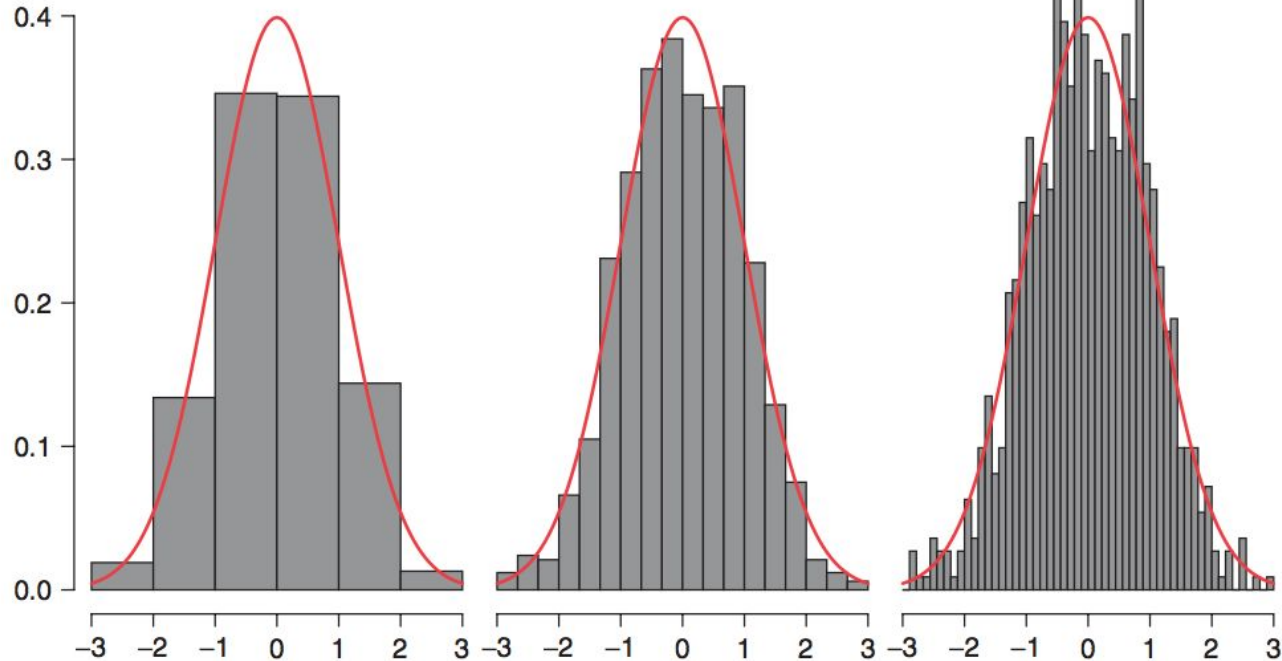
Too **wide** bins?

Too **narrow** bins?





Histogram can be considered as a density estimation method.

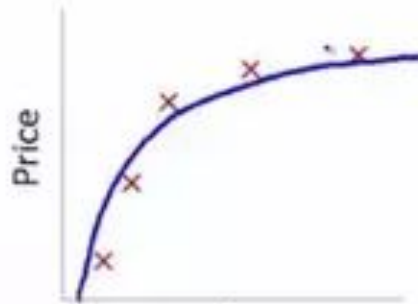


Bias-Variance tradeoff



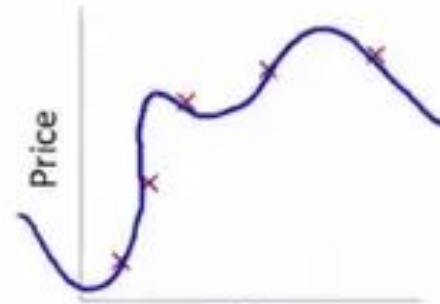
$$\theta_0 + \theta_1 x$$

High bias
(underfit)



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

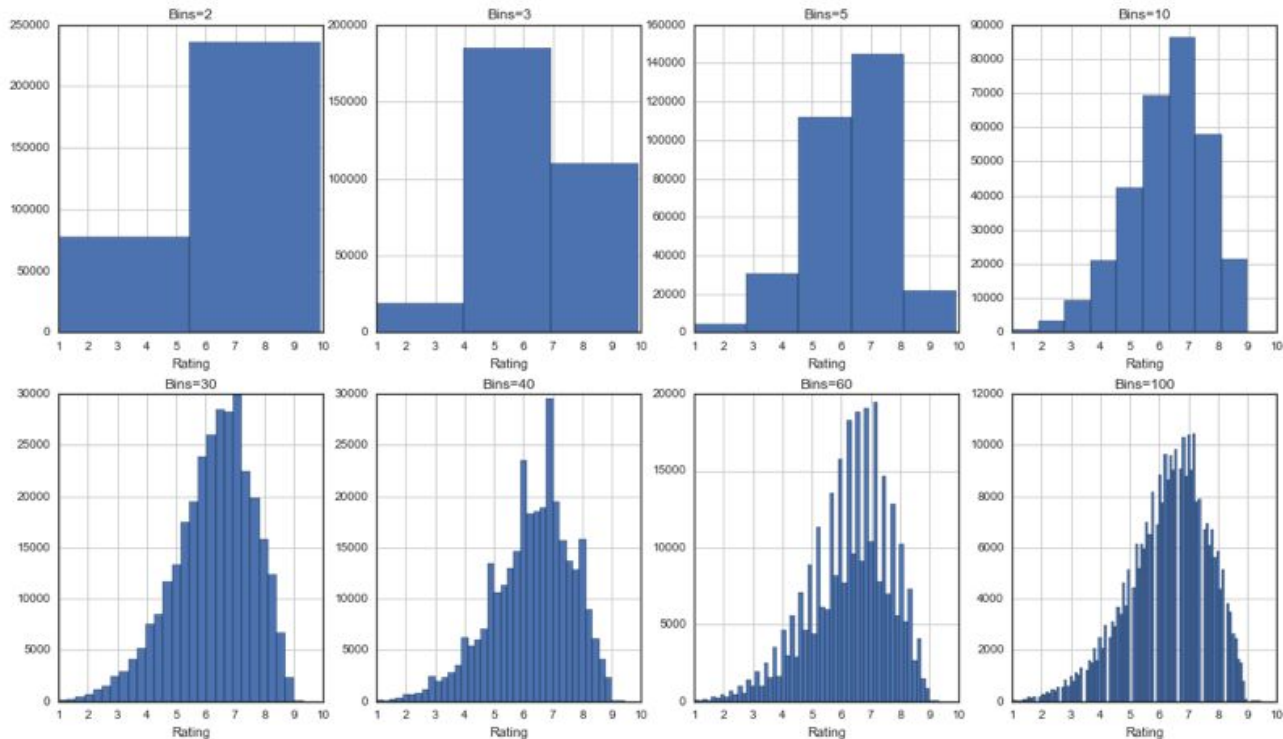
“Just right”



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

High variance
(overfit)

Too few bins \rightarrow large bias (inaccurate)
Too many bins \rightarrow large variance (overfit to the data/noise/artifacts)

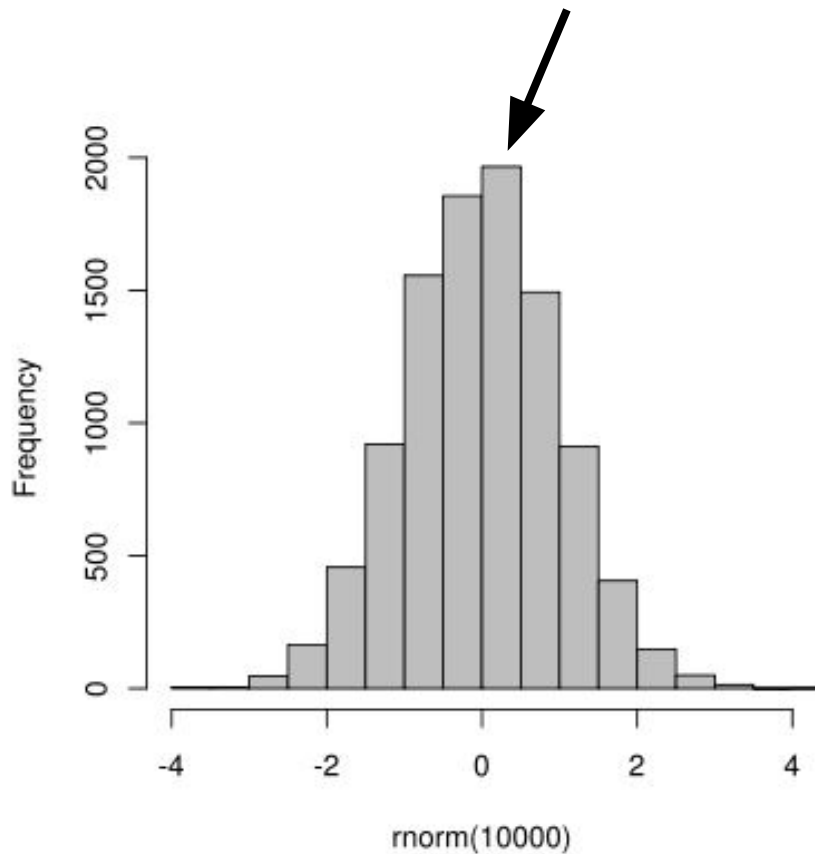


How to choose bins?

Equal bins?

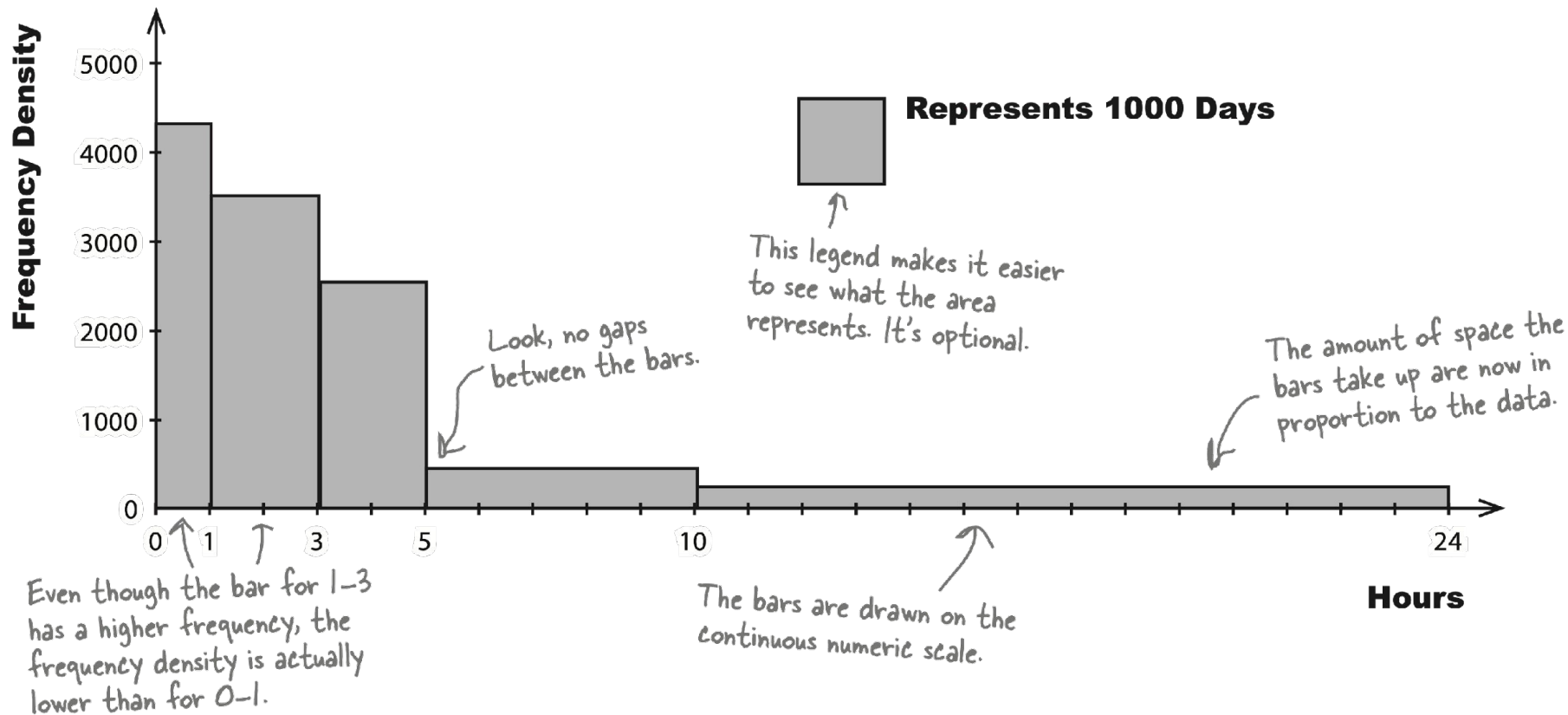
Variable bins?

More data here

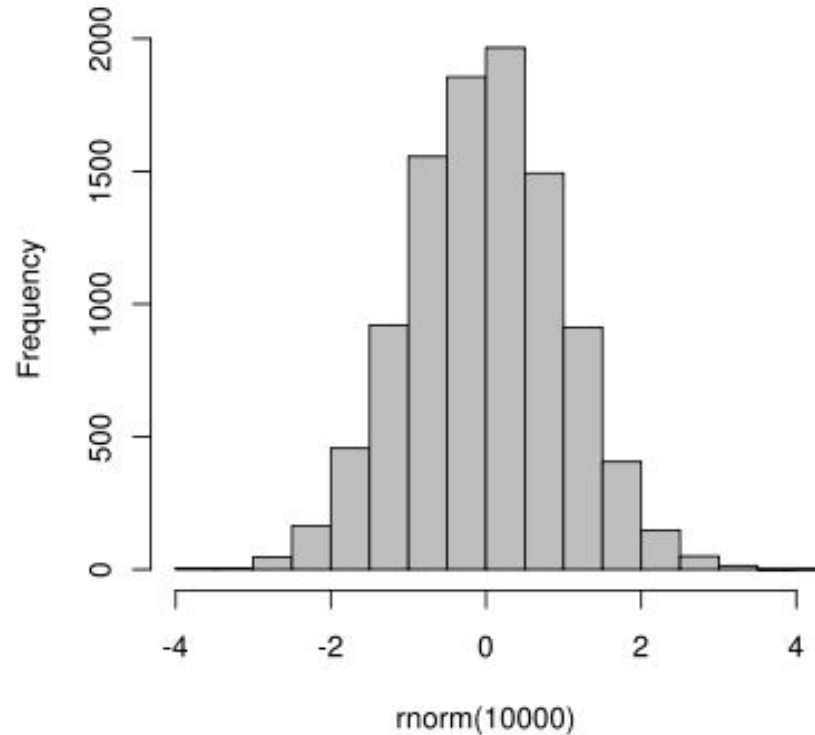


Variable bin →
Complicated

Hours Spent Gaming per Day



Simplicity and transparency



Histogram is simple, direct,
and transparent.

How to pick the number
of bins?

numpy.histogram

numpy.histogram (*a*, *bins*=10, *range*=None, *normed*=False, *weights*=None, *density*=None)

[\[source\]](#)

Compute the histogram of a set of data.

Parameters: *a* : *array_like*

Input data. The histogram is computed over the flattened array.

bins : *int* or *sequence of scalars or str, optional*

If *bins* is an int, it defines the number of equal-width bins in the given range (10, by default). If *bins* is a sequence, it defines the bin edges, including the rightmost edge, allowing for non-uniform bin widths.

New in version 1.11.0.

If *bins* is a string from the list below, **numpy.histogram** will use the method chosen to calculate the optimal bin width and consequently the number of bins (see *Notes* for more detail on the estimators) from the data that falls within the requested range. While the bin width will be optimal for the actual data in the range, the number of bins will be computed to fill the entire range, including the empty portions. For visualisation, using the 'auto' option is suggested. Weighted data is not supported for automated bin size selection.

'auto'

Maximum of the 'sturges' and 'fd' estimators. Provides good all around performance.

'fd' (Freedman Diaconis Estimator)

Robust (resilient to outliers) estimator that takes into account data variability and data size.

'doane'

An improved version of Sturges' estimator that works better with non-normal datasets.

'scott'

Less robust estimator that takes into account data variability and data size.

'rice'

Estimator does not take variability into account, only data size. Commonly overestimates number of bins required.

'sturges'

R's default method, only accounts for data size. Only optimal for gaussian data and underestimates number of bins for large non-gaussian datasets.

'sqrt'

Square root (of data size) estimator, used by Excel and other programs for its speed and simplicity.

Sturges' rule

THE CHOICE OF A CLASS INTERVAL

CASE I. COMPUTATIONS INVOLVING A SINGLE SERIES

In case a single statistical series of range R with N items is involved in a computation, the optimal class interval may be estimated from the formula

$$C = \frac{R}{1 + 3.322 \log N}$$

This formula gives the class interval for the computation of the averages, measures of dispersion, skewness, etc., of frequency distributions. It is based on the principle that the proper distribution into classes is given, for all numbers which are powers of 2, by a series of binomial coefficients. For example, 16 items would be divided normally into 5 classes, with class frequencies 1, 4, 6, 4, 1. Thus if a statistical series had 16 items with values ranging from 20 to 70, or a range of 50 points, it should be divided into 5 classes of 10 points each, that is, the class interval would be 10. Similarly, 64 is the sixth power of 2, so a statistical series containing 64 items should be divided into 6 plus 1, or 7 classes. If such a series had a range of 35 points the class interval would be 5.

The most convenient class intervals are 1, 2, 5, 10, 20, etc., so that in practice the formula for the theoretical class interval may be used as a means of choosing among these convenient ones. In general the next smaller convenient class interval should be chosen, that is, the one next below the theoretically optimal interval. If the formula gives 9, 10 may be chosen, but if the formula indicates 7 or 8, the one actually used should generally be the next lower convenient class interval, 5.

Scott's rule

David W. Scott*

The optimal construction of a histogram as a reference distribution for data with a discrepancy the mean inter-bin width and formulation of this form and their relative performance.

DEFINITION OF SCOTT'S RULE

Scott's rule is a formula giving a choice of bin width for an equally spaced histogram of continuously measured data, namely,

$$\text{bin width} = 3.5 \hat{\sigma}_x n^{-1/3} \equiv \hat{h}.$$

Here h is the bin width, n is the sample size, and $\hat{\sigma}_x$ is the usual estimate of the standard deviation

$$\hat{\sigma}_x = \sqrt{s_x^2}, \quad \text{where} \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2)$$

(Sometimes the coefficient in Eq. (1) is taken to be 3.49).

THE HISTOGRAM AS DENSITY ESTIMATOR

The general histogram is derived from counts of the

$$\text{bin width} = 3.5 \hat{\sigma}_x n^{-1/3} \equiv \hat{h}. \quad (1)$$

Here h is the bin width, n is the sample size, and $\hat{\sigma}_x$ is the usual estimate of the standard deviation

$$\hat{\sigma}_x = \sqrt{s_x^2}, \quad \text{where} \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2)$$

$$f(x) = \frac{v_k}{n h_k}, \quad x \in B_k \quad (\text{general case}). \quad (3)$$

Usually, the bins are selected to have the same width, that is, $h_k = h$ for all k . Hence,

$$\hat{f}_h(x) = \frac{v_k}{n h}, \quad x \in B_k \quad (\text{equal-width case}). \quad (6)$$

Observe that a plot of the unnormalized frequencies $\{v_k\}$ can be quite misleading in the general case (5), unlike in case (6) for an equally spaced histogram case, where the two curves are related by a scalar.

The bin origin t_0 is often chosen to be 0 or the smallest observed data value. Thus, a choice of the two parameters (h, t_0) completely specifies a histogram. If

Freedman-Diaconis (FD) rule

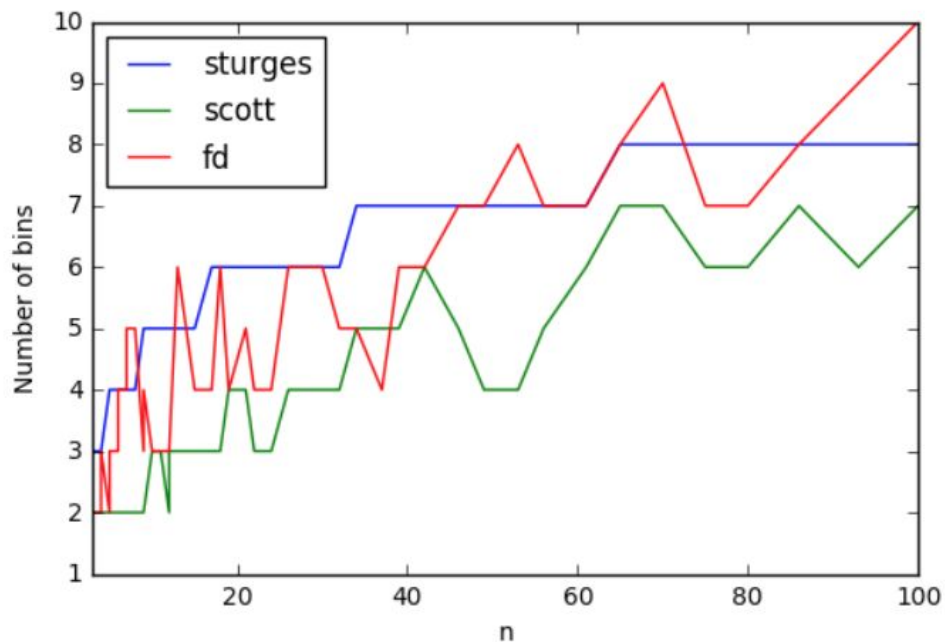
$$\text{Bin size} = 2 \frac{\text{IQR}(x)}{n^{1/3}}$$

```
In [16]: data = {'n': np.logspace(0.5, 2, num=50).astype(int)}

for method in ['sturges', 'scott', 'fd']:
    data[method] = np.array([nbins(np.random.normal(size=x),
                                method) for x in data['n']])

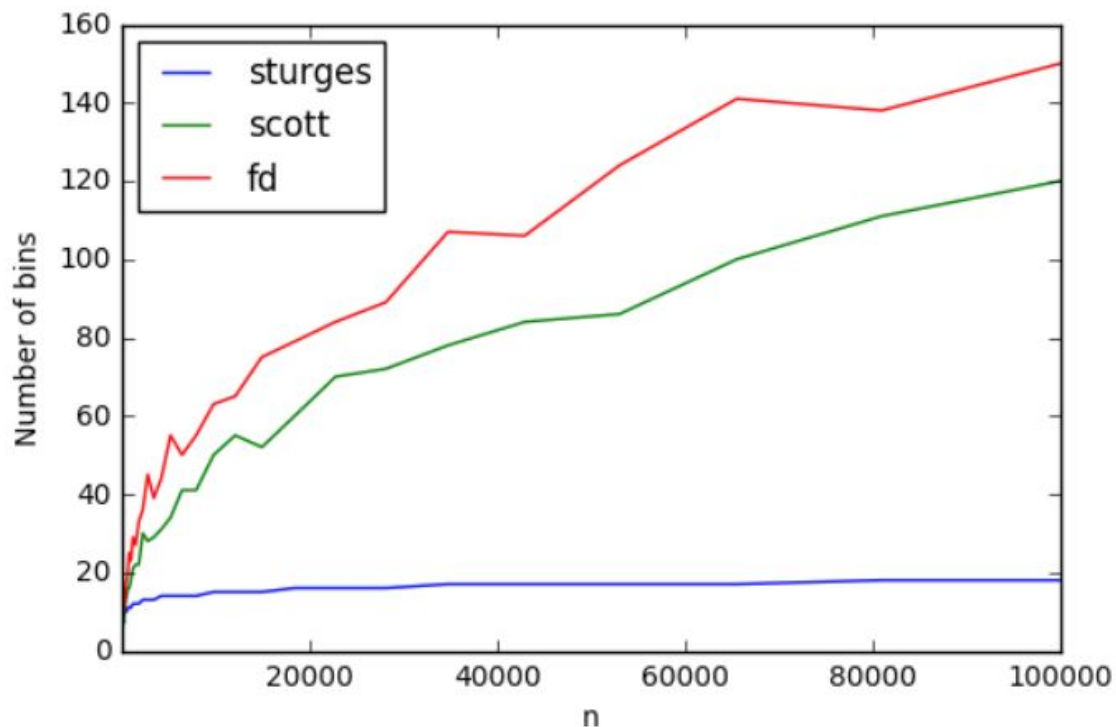
df = pd.DataFrame(data, columns=['n', 'sturges', 'scott', 'fd'])
ax = df.plot(x='n')
ax.set_ylabel("Number of bins")
```

Out[16]: <matplotlib.text.Text at 0x10fe6b208>



```
In [15]: ax = df.plot(x='n')  
ax.set_ylabel("Number of bins")
```

```
Out[15]: <matplotlib.text.Text at 0x10d1bc588>
```



Default number of bins

R: Sturges

numpy/matplotlib: $\max(\text{Sturges}, \text{FD}) \rightarrow 10$

There are formulae, but they
are **fragile**, so don't
rely on them.

Histogram is cheap and
easy.

Try multiple histograms

Boxplot? histogram?

Which should I use?

