Descriptive Statistics

The Importance of Visualization

Michael Luu, MPH | Marie Lauzon, MS
Biostatistics & Bioinformatics Research Center | Cedars Sinai Medical Center
September 11, 2024

Slides can be accessed with the following link:

https://mluu921.github.io/cshs-fall-lecture-descriptive-statistics

Also available as a PDF on Onedrive

Why do we need to visualize our data?

Data

A B C D

X	Y
55.4	97.2
51.5	96.0
46.2	94.5
42.8	91.4
40.8	88.3

Let’s begin by taking descriptive measures

DATASET	N	MEAN_X	SD_X	MEAN_Y	SD_Y
A	142	54.3	16.8	47.8	26.9
B	142	54.3	16.8	47.8	26.9
C	142	54.3	16.8	47.8	26.9
D	142	54.3	16.8	47.8	26.9

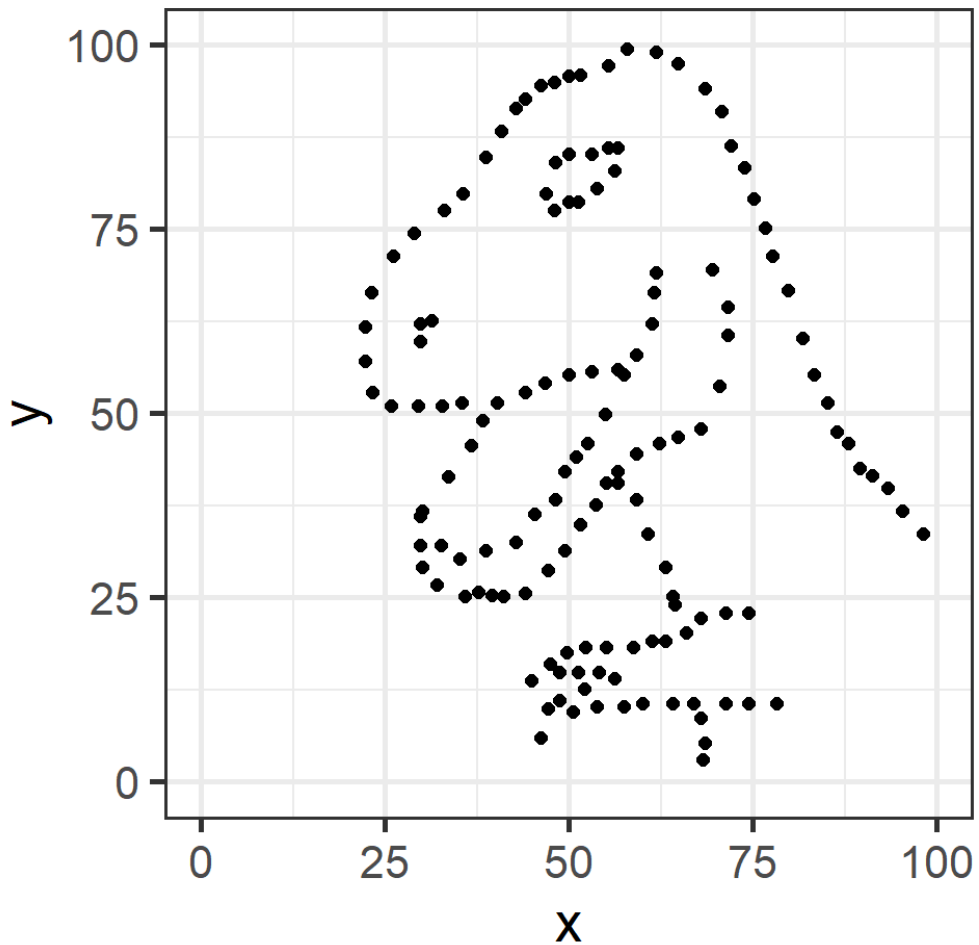
It appears the counts (n), mean (x), mean (y), and sd (x) and sd (y) are identical for ALL four datasets!

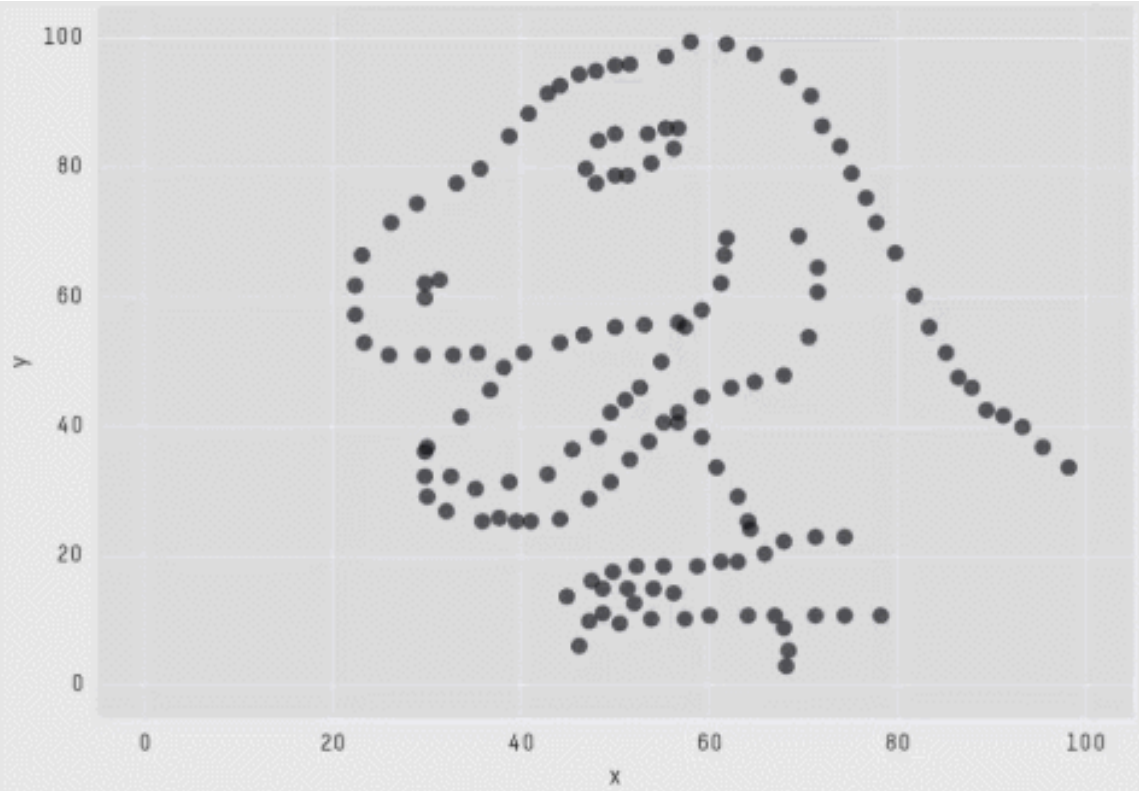
Can we conclude the datasets are similiar or identical?

Not quite yet!

Let’s visualize the relationship of x and y

A B C D





X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

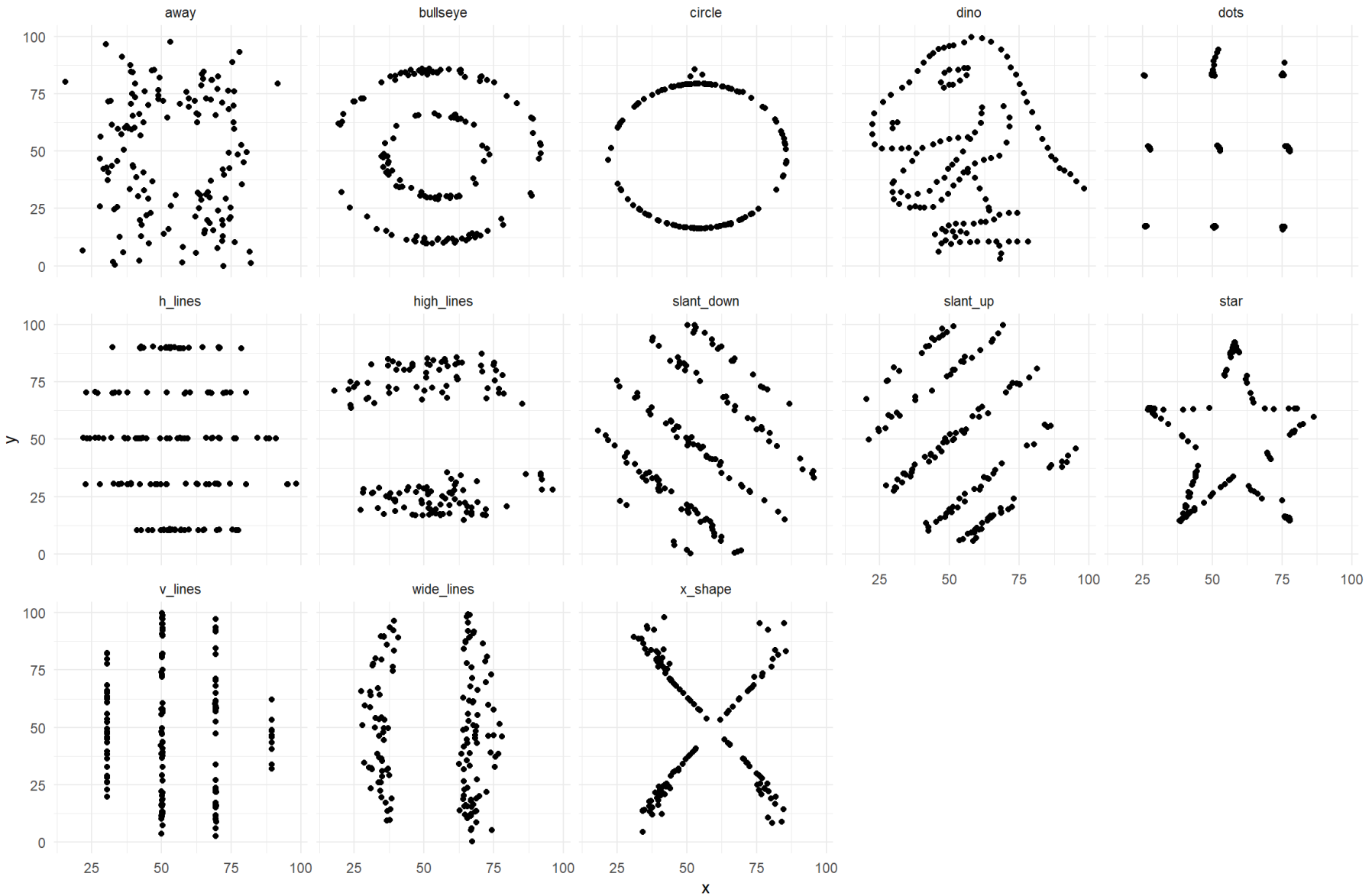
Although simple quantitative summaries are similar …

They can appear drastically different when visualized!

Datasaurus Dozen

- The original “Datasaurus” or “dino” was created by **Alberto Cairo** in the following blog post[1]
- He was then later made famous by the paper published by **Justin Matejka** and **George Fitzmaurize**, titled ‘Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing’[2], where they simulated 12 additional datasets in addition to the original “Datasaurus” with nearly identical simple statistics

Datasaurus Dozen

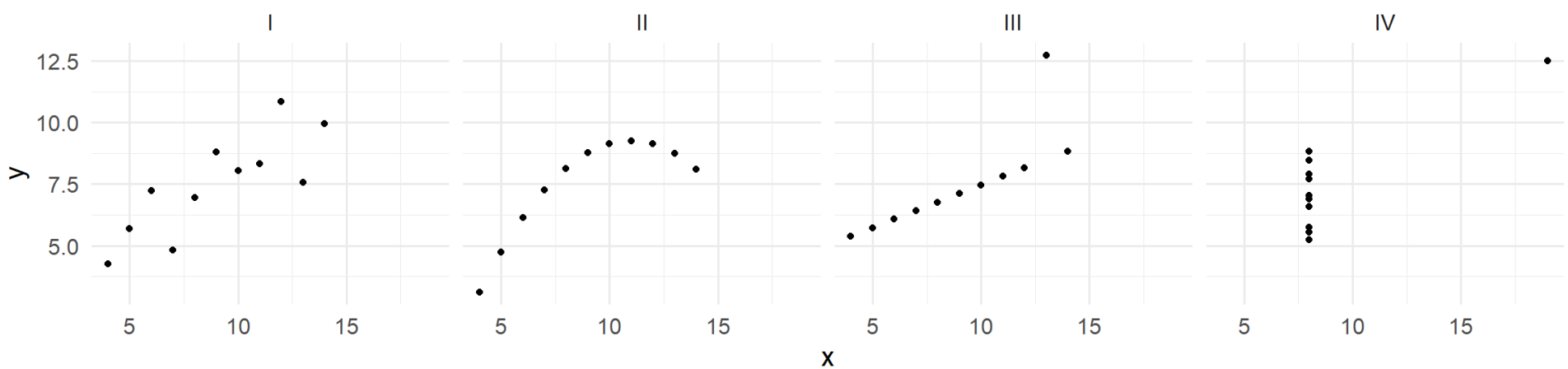


Anscombe Quartet

- The datasaurus dozen is a modern take on the classical **"Anscombe's Quartet"**[1]
- Comprised of four datasets that have nearly identical simple summary measures, yet have very different distributions and appear vastly different when plotted

Anscombe Quartet

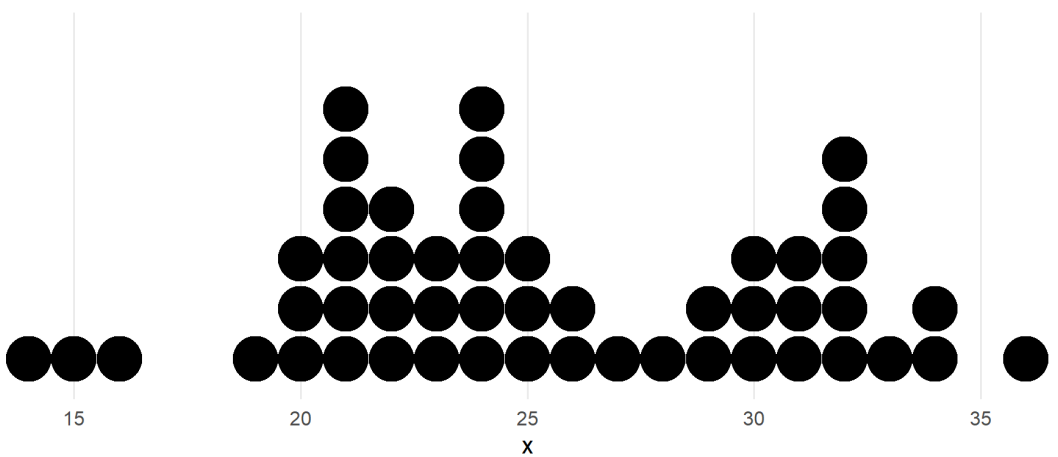
dataset	n	mean_x	sd_x	meay_y	sd_y
I	11.00	9.00	3.32	7.50	2.03
II	11.00	9.00	3.32	7.50	2.03
III	11.00	9.00	3.32	7.50	2.03
IV	11.00	9.00	3.32	7.50	2.03



Types of Graphical Visualizations

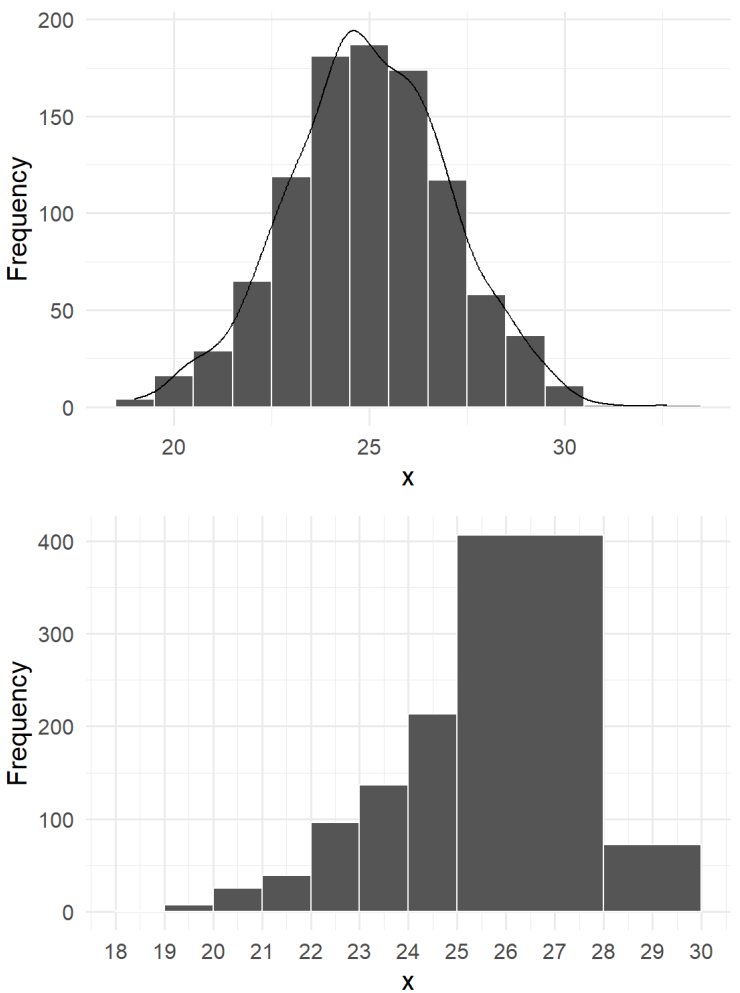
Dot plot

- Useful for small to moderate sized data
- Allows us to visualize the spread and distribution of one continuous discrete variables
  - e.g. length of stay
- The X axis is the variable of interest and each dot represents a single observation
- Easy to identify the mode
- Highlights clusters, gaps, and outliers
- Intuitive and easy to understand



Histogram

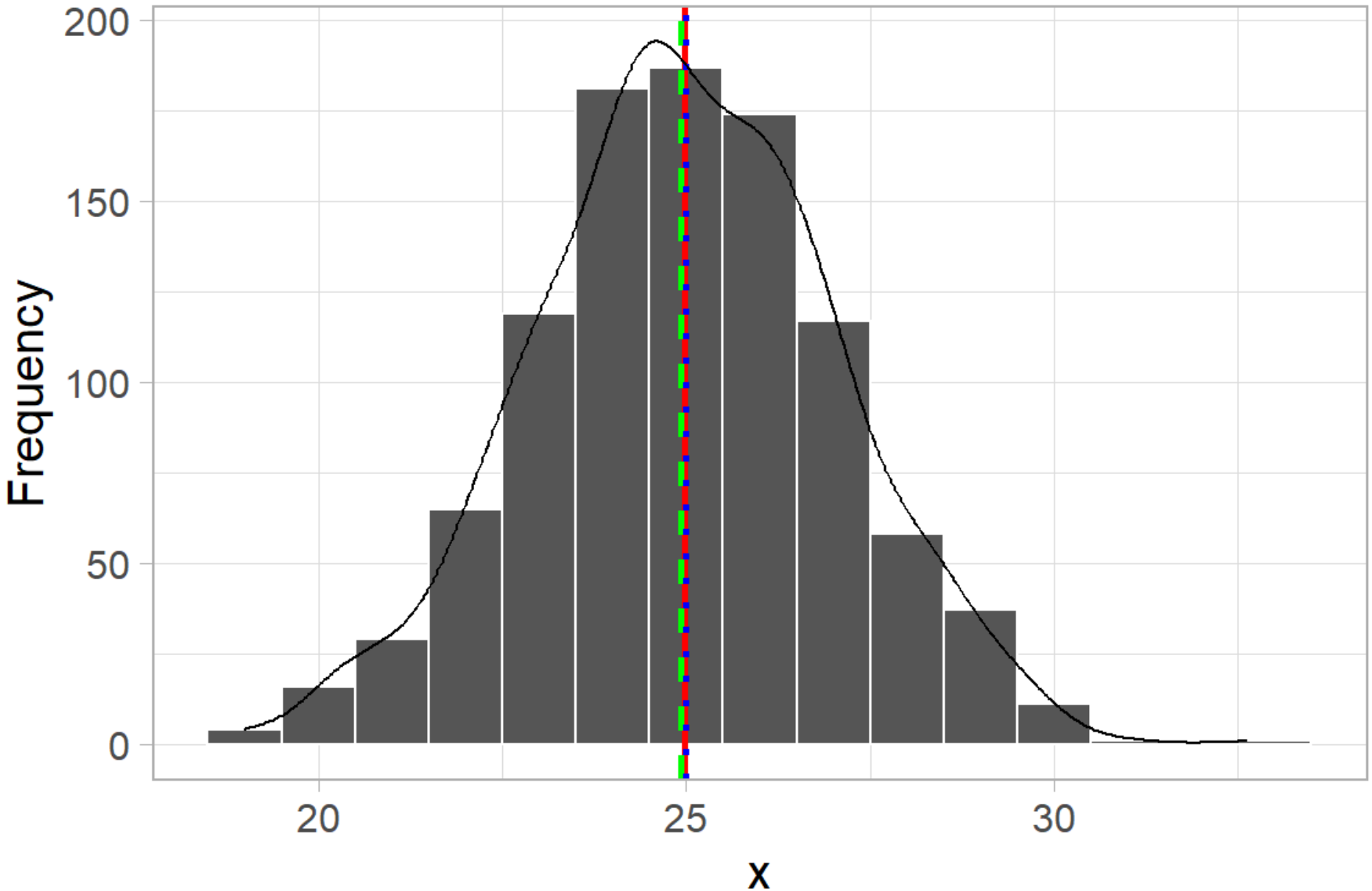
- Useful for all sized data (small and large)
- Allows us to visualize the spread and distribution of continuous variables
- Each bar represents a ‘bin’ or a defined interval of values
- Although not as common, the width of the bins does NOT have to be equal!
- The y axis or the height of the bar represents the count of the number of values that fall into each bin
- The y axis is also commonly normalized to ‘relative’ frequencies to show the proportion of cases or density that falls into each bin.



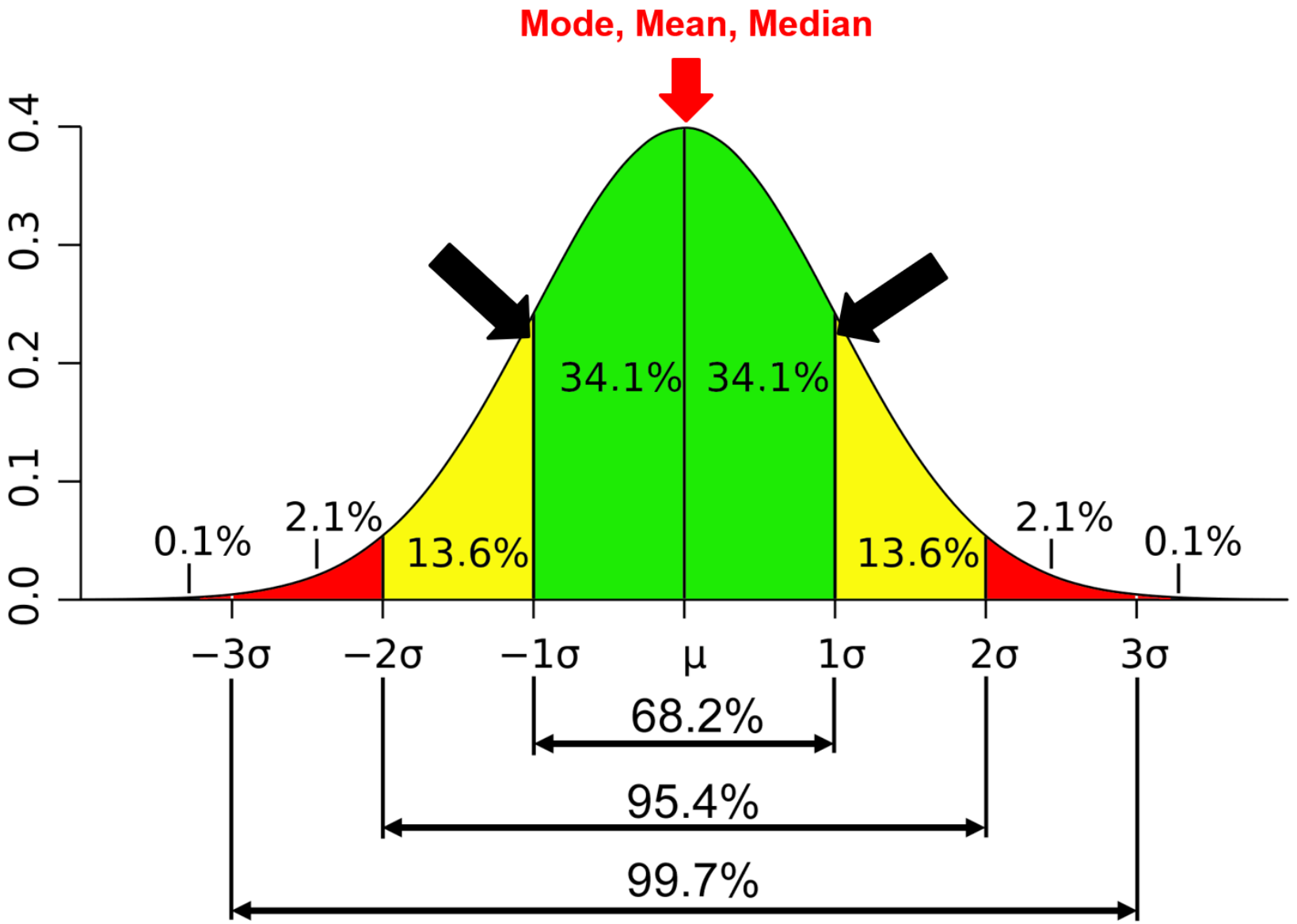
Distribution

> “A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.”[1]

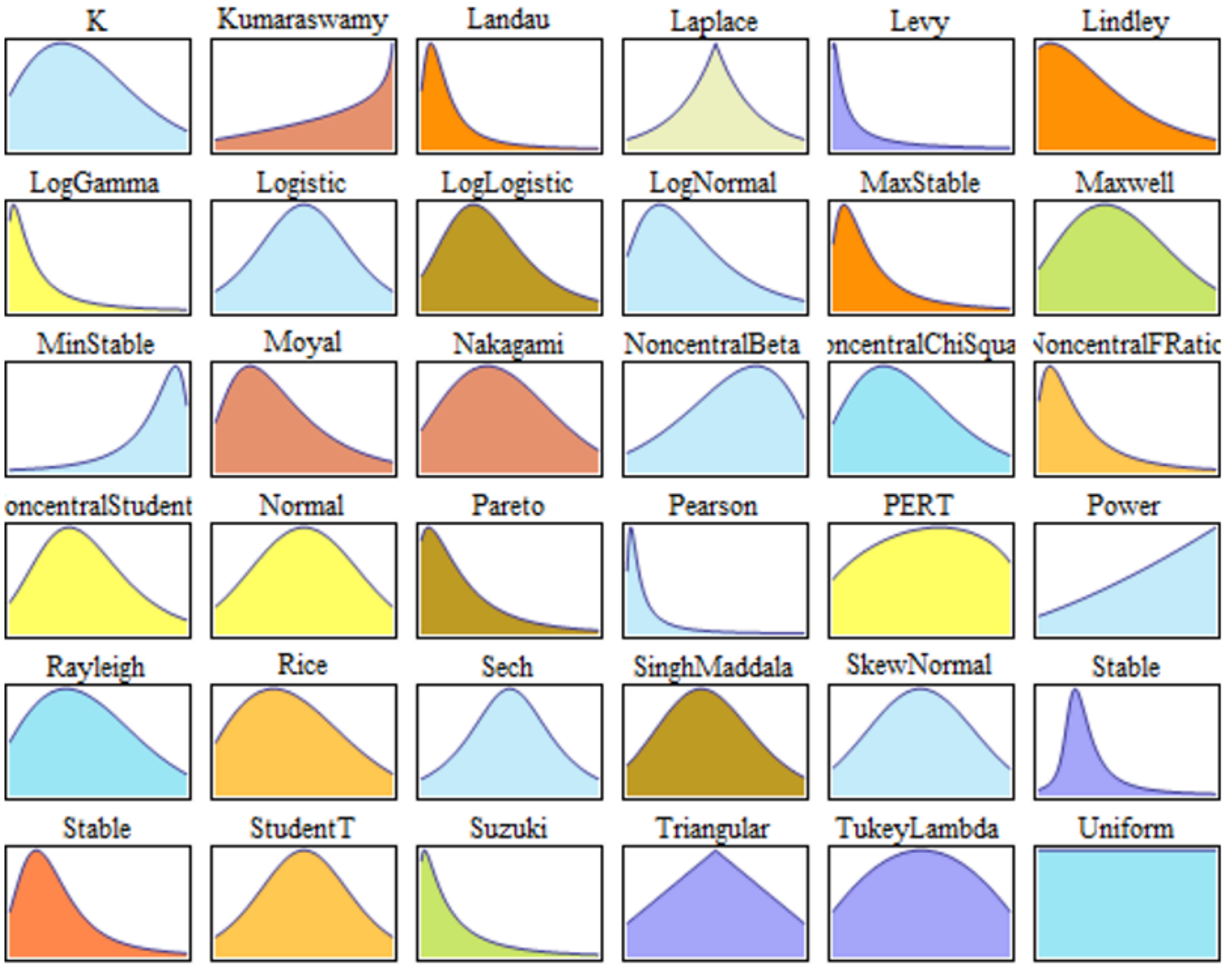
Distribution



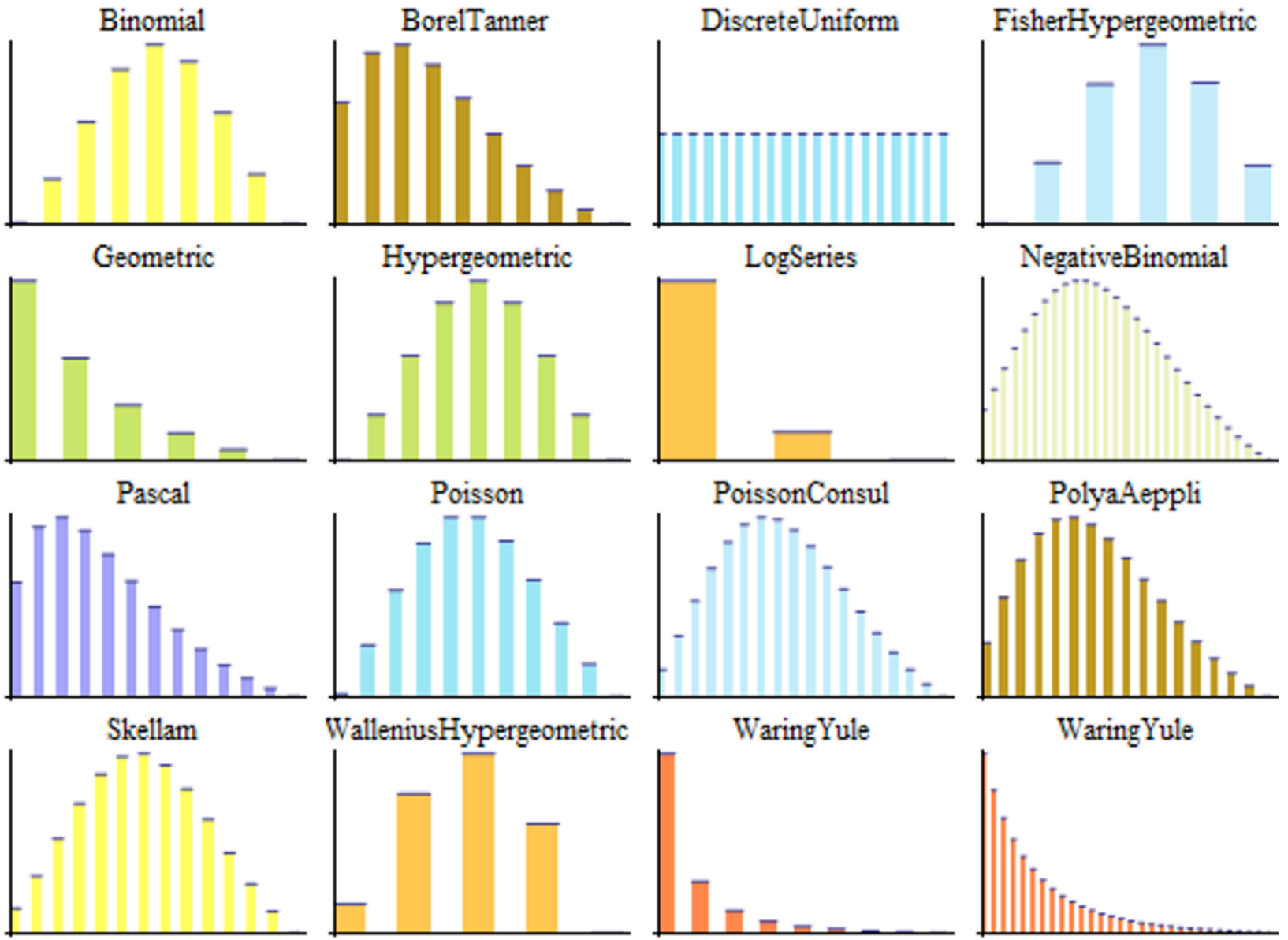
Normal Distribution



Univariate Continuous Distributions

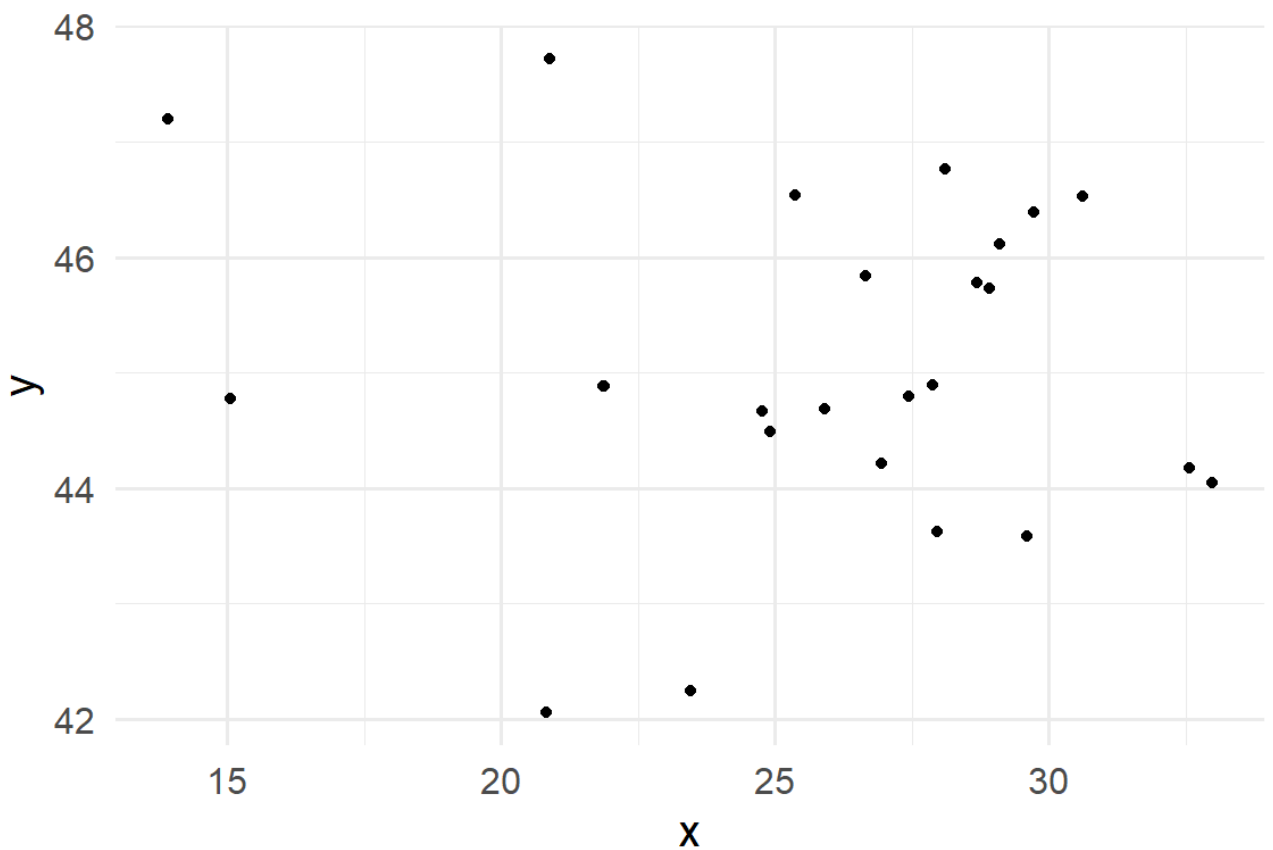


Univariate Discrete Distributions



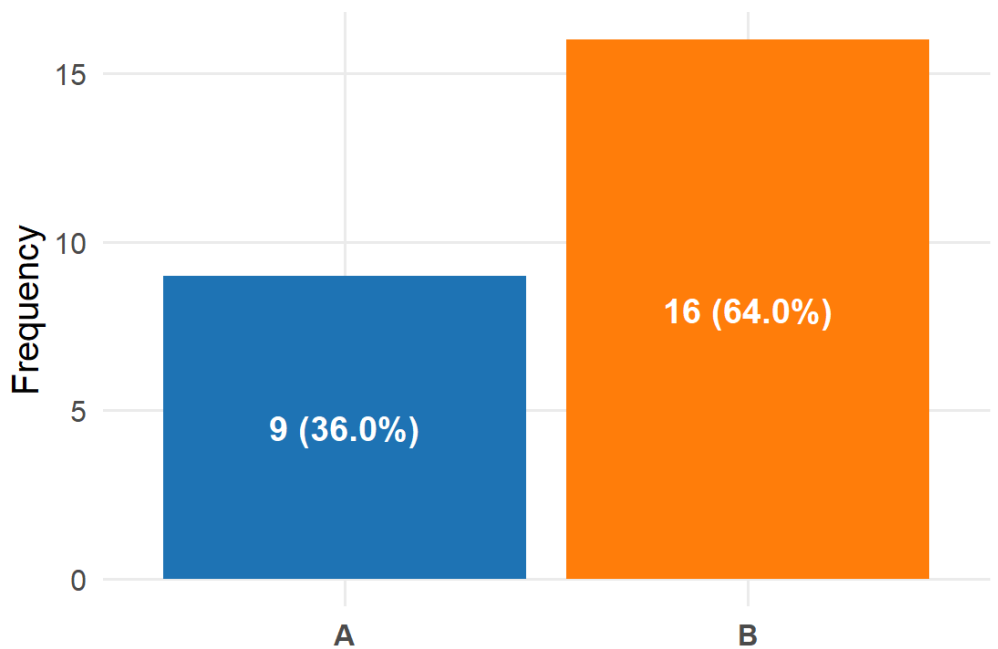
Scatter plot

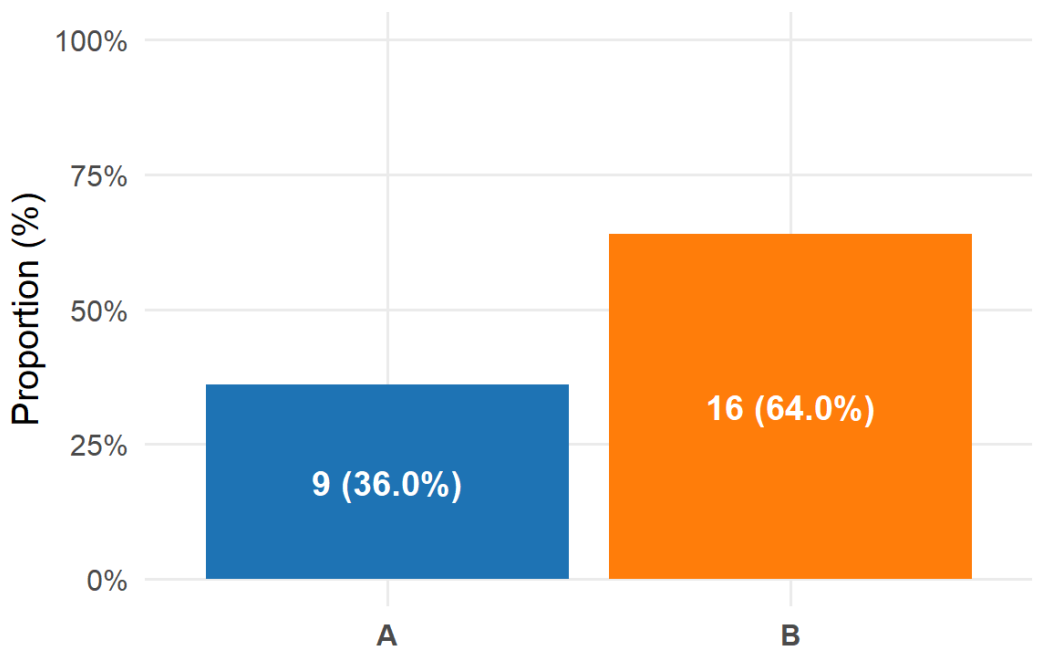
- Used to visualize the relationship between two continuous variables
- Useful for detecting patterns that are obscured from quantitative summaries like what we observed in Anscombe's quartet and the Datasaurus dozen.



Bar plot

- Useful for visualizing **categorical** data
- Commonly used to present counts and proportion of each level
- Allows us to quickly observe the difference in magnitude of each level based on the height of each bar



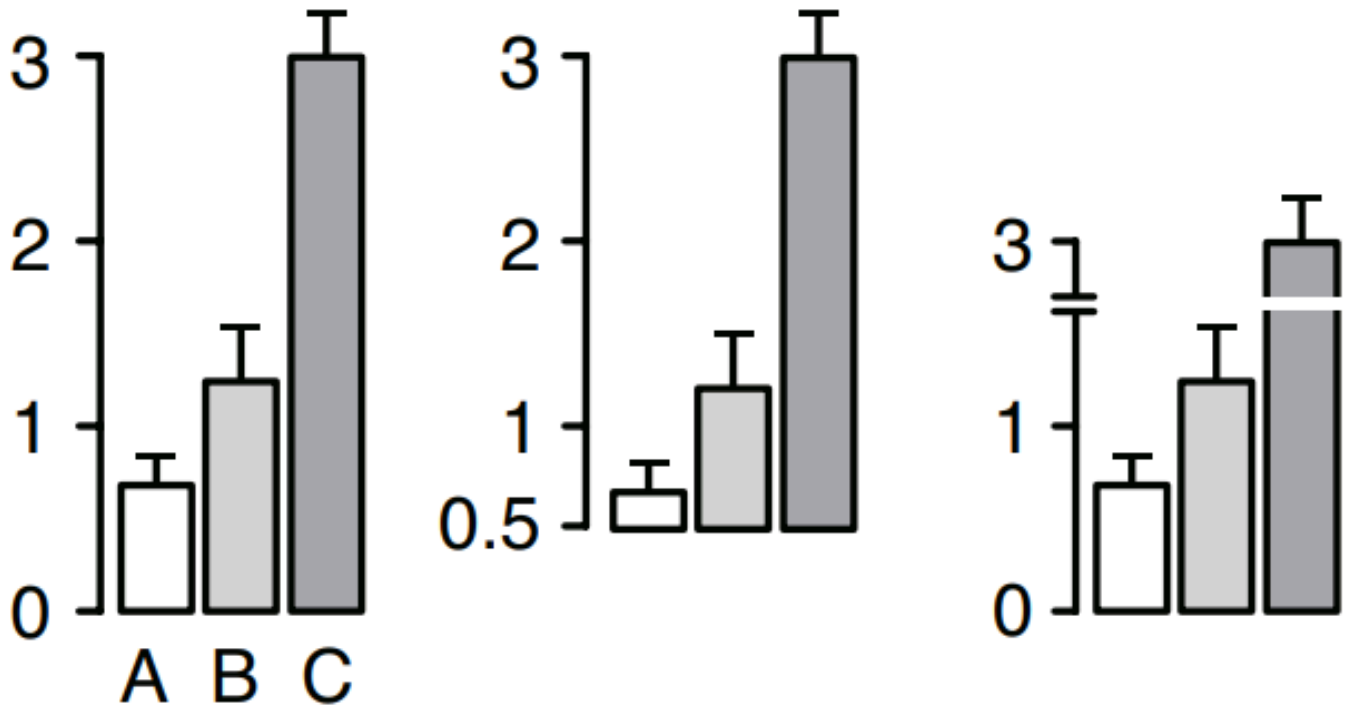


However…

Bar plots are commonly misued!

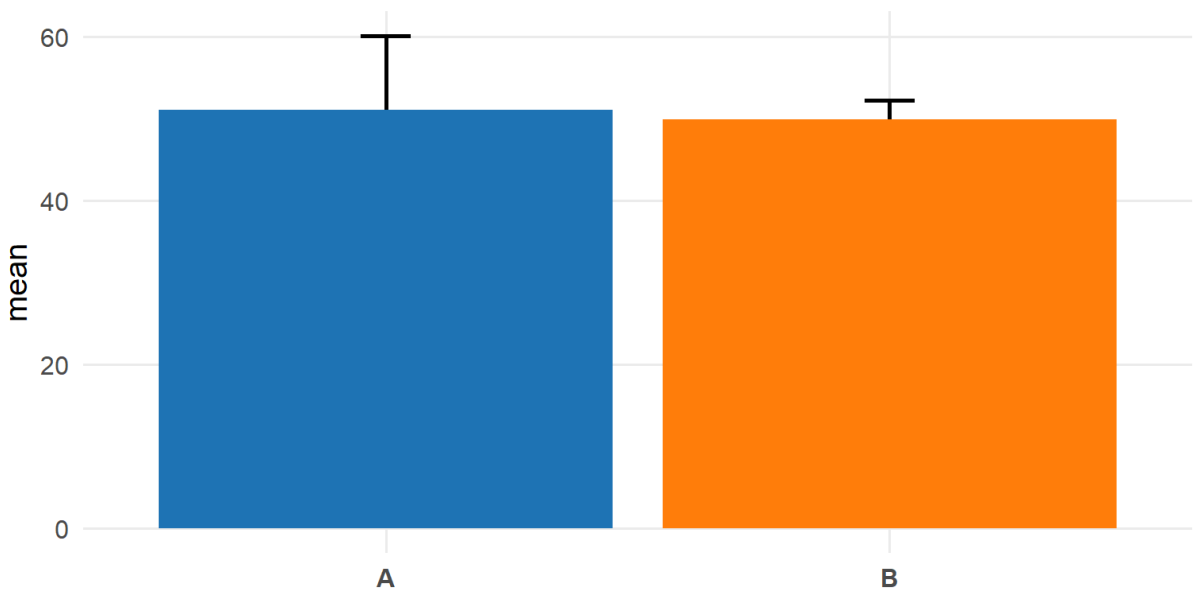
How NOT to Bar Plot

Not recommended



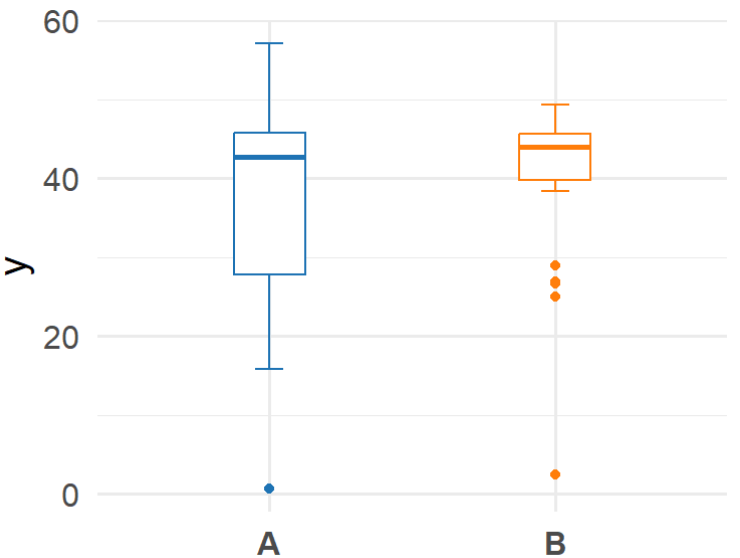
How NOT to Bar Plot

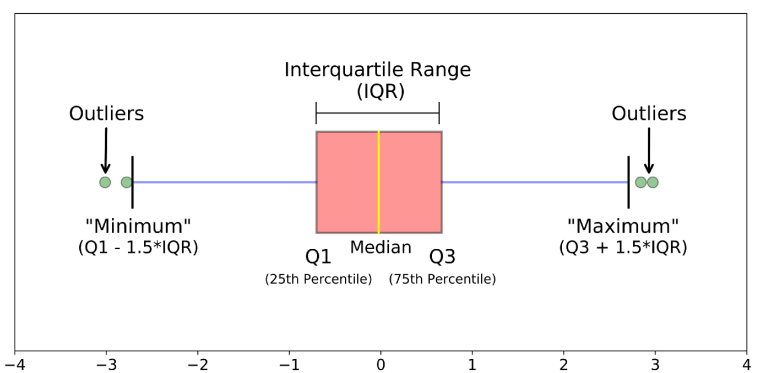
- Although frequently found and prevalent in the literature, this is NOT to be used to describe mean and dispersion (continuous data)
- Only shows one arm of the error bar, making overlap comparisons difficult
- Promotes misconception of the mean being related to its height rather the position of the top of the bar
- Obscures the distribution and spread of the data



Box plot

- Useful for describing continuous variables following a uni-modal distribution
  - e.g. a single peak
- The box is representative of common quantitative measures
  - Top of box is the 75th quantile
  - Middle dash inside box is the 50th quantile
  - Bottom of box is the 25th quantile
  - Width of the box is the interquartile range (IQR)
- The ‘whiskers’ are artificial ‘fences’ that helps identify potential outliers in the data
  - Defined as Q1 - 1.5*IQR and Q3 + 1.5*IQR

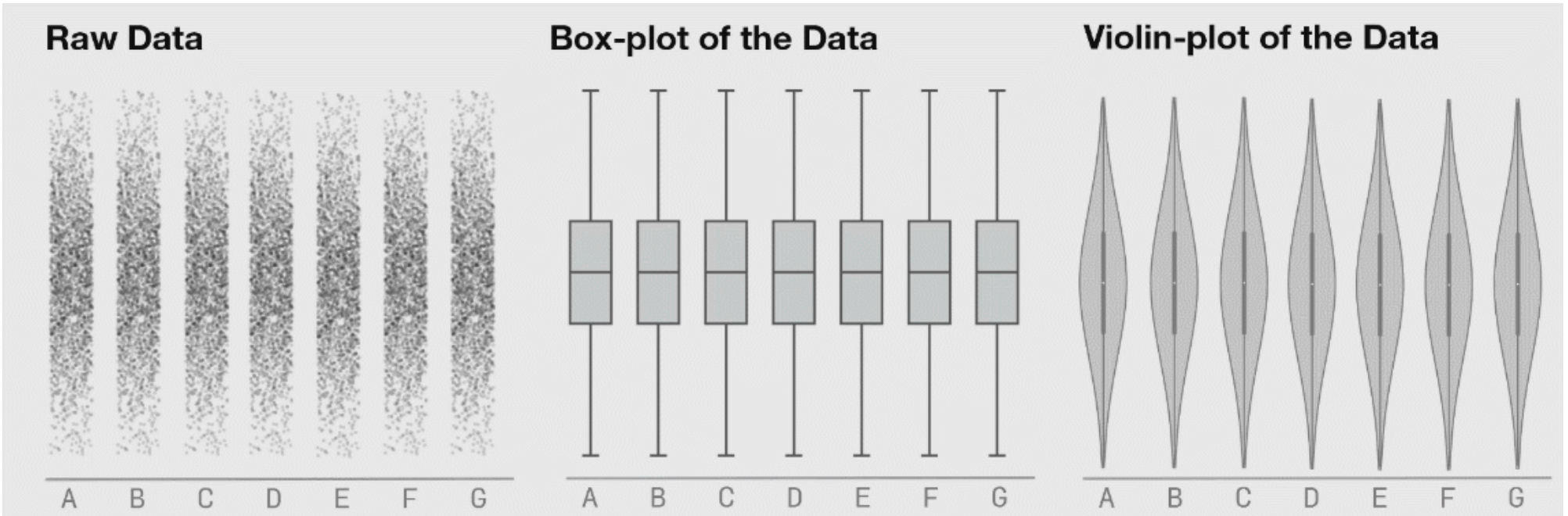




What are some of the problems with a box plot?

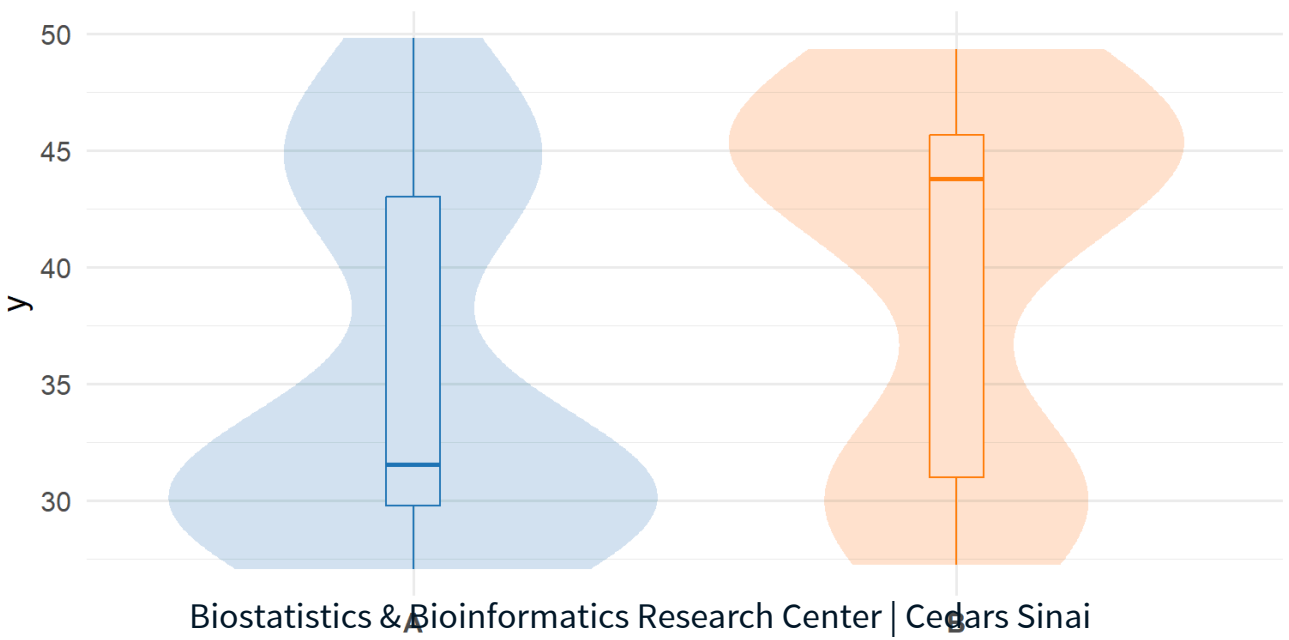
They are based on quantitative summaries!

Box plot

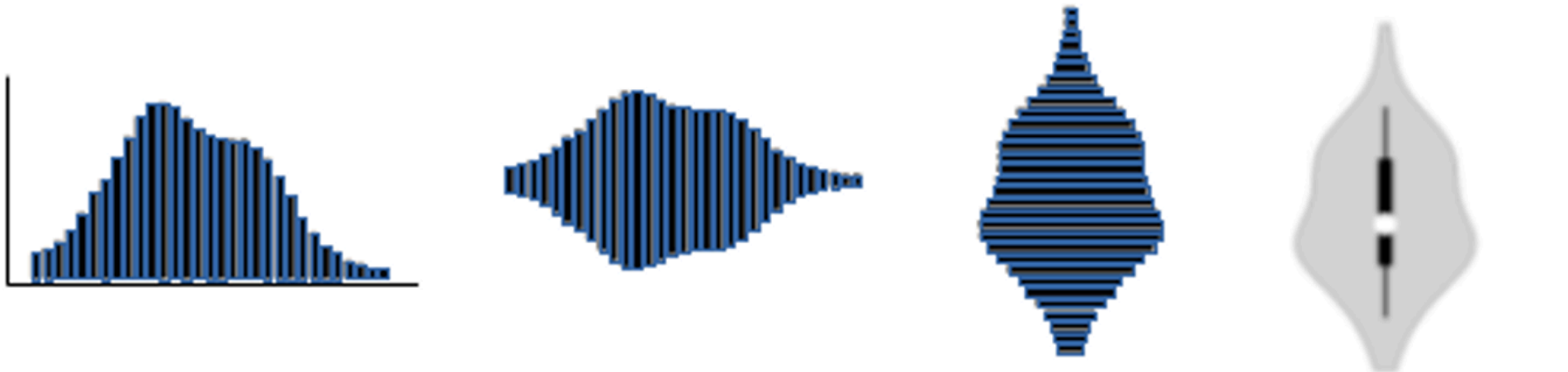

Raw Data
Box-plot of the Data
Violin-plot of the Data
A B C D E F G
A B C D E F G
A B C D E F G

Violin plot

- Violin plots are box plots, with an overlay of the density distribution (histogram) of the data
- More informative than a simple box plot
- Visualizes the full distribution of the data
- Especially useful for bimodal or multimodal distribution
  - e.g. distribution of data with multiple peaks



How are violin plots made?



1. Create histogram
2. Center the bars
3. Rotate
4. Replace shape

Summary

- **One continuous variable**
  - Dot plot
  - Histogram
  - Box plot
  - Violin plot
- **One or more categorical variable**
  - Bar plot
- **Two continuous variable**
  - Scatter plot
- **One continuous by categorical variable**
  - Dot plot
  - Box plot
  - Violin plot

Descriptive summaries are useful, however …

Don’t forget to visualize your data!

Questions