Types of data and describing data with summary statistics

DATA SCIENCE FOR BIOLOGISTS

STEPHANIE J. SPIELMAN, PHD

Types of data

How you analyze and visualize data depends on the *type* of data you have

Quantitative data

- Continuous
- Discrete (includes count data)

Categorical data

- Nominal
- Ordinal
- Binary*

Human data type	R data type
Quantitative continuous	Numeric (double)
Quantitative discrete	Numeric (integer)
Categorical nominal	Character or factor
Categorical ordinal	Factor
Categorical binary	Factor or logical

Quantitative data

Continuous

- Any real-number value within some range

Discrete

- Values are in indivisible units, i.e. whole or counting numbers
- Includes **count data** (number of cups of coffee per day, number of amino acids…)

Categorical data

Nominal

- Hair color, eye color, sex genotypes (XX, XY, XXY, XYY, XO, …)

Ordinal – categories with a natural ordering

- Bad, fair, good, excellent
- A, B, C, D

Binary

- Yes/No
- True/False

Data types translated

Human data type	R data type
Quantitative continuous	Numeric (double)
Quantitative discrete	Numeric (integer)
Categorical nominal	Character or factor
Categorical ordinal	Factor
Categorical binary	Factor or logical

Measures of Location

Continuous

Mean

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

Median

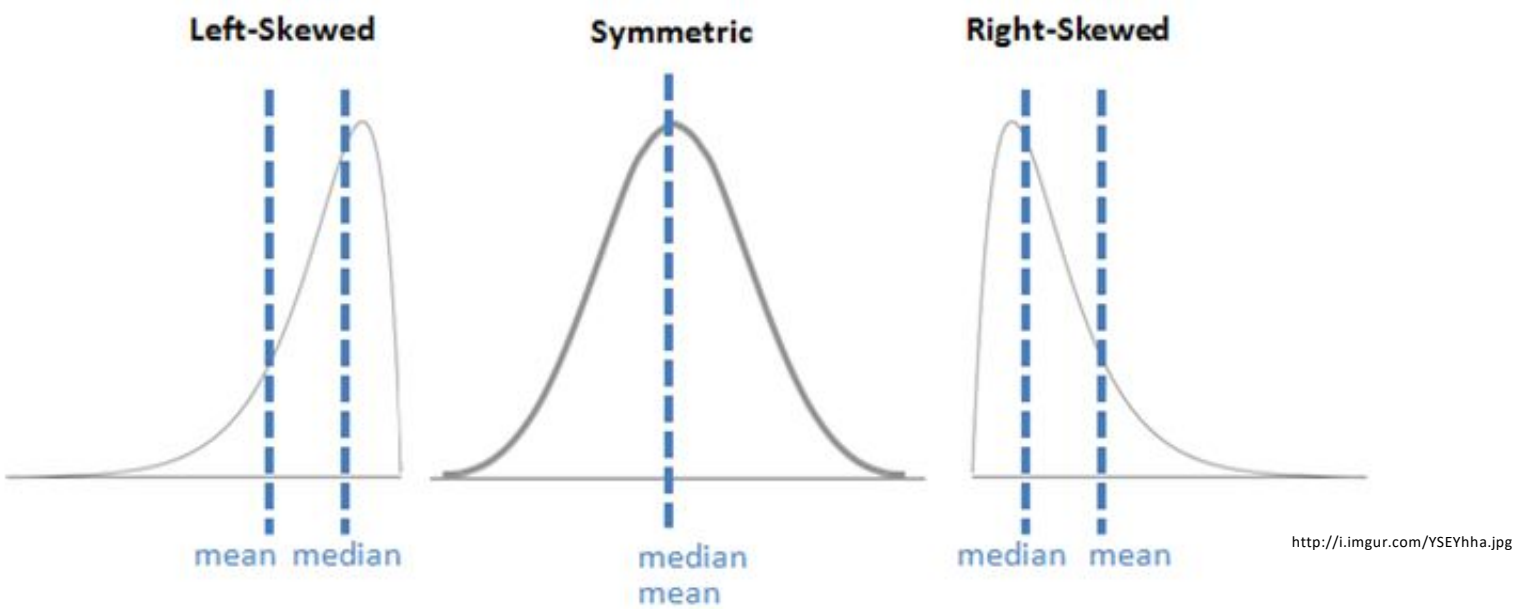
- For odd *n,* the $\left(\frac{n+1}{2}\right)th$ observation
- For even *n,* the average of the $\left(\frac{n}{2}\right)th$ and $\left(\frac{n}{2}+1\right)th$ observation

Discrete or categorical

Mode

- The most frequent appearing observation in the distribution
- 1, 2, 2, 2, 3, 4, 4, 5, 6 → **2**
- Large, large, large, small → **Large**

Measures of location in distributions



http://i.imgur.com/YSEYhha.jpg

Measures of spread

Range

Standard deviation and variance

Interquartile range

Range

Difference between largest and smallest value in a distribution

- 1, 2, 3, 7, 9 → **8**
- 1, 2, 3, 7, 9, 500 → **499**

Range is very sensitive to extreme observations and becomes very unwieldy very quickly.

Standard deviation and variance

Generally discussed in the context of **mean**

Deviance describes how each *n*th data point *deviates* from mean $\bar{Y}$:

- $Y_1 - \bar{Y},\ Y_2 - \bar{Y},\ Y_3 - \bar{Y}$**, …,** $Y_n - \bar{Y}$

Standard deviation of a sample

- $s = \frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

Variance

- s^2

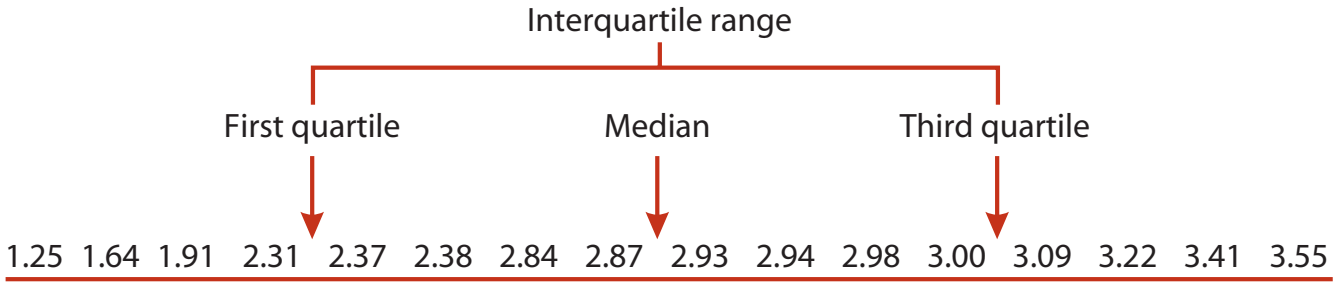
Interquartile range

Generally discussed in the context of **median**

Quartiles divide the data into **four** equal parts (“quar”!)

Interquartile range (IQR) is the difference between the third and first quartile

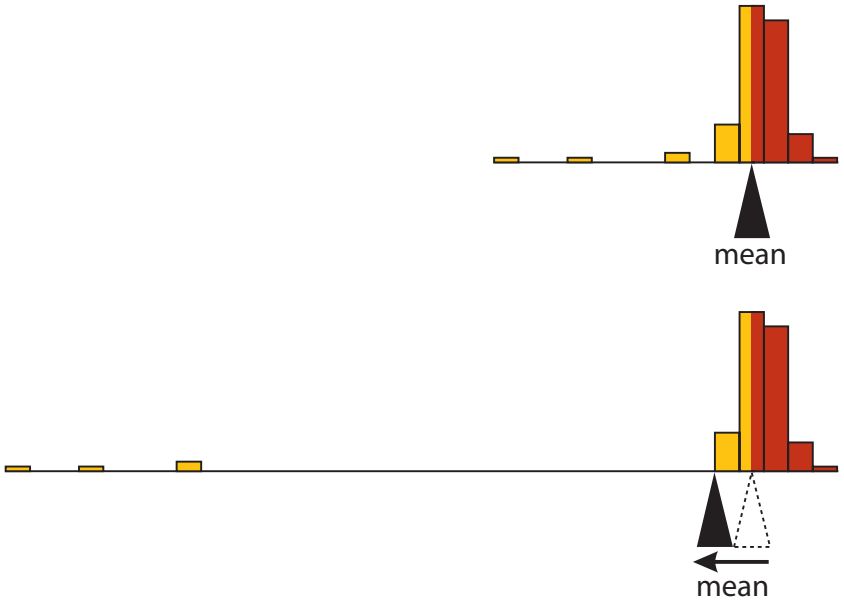
- How much of the data does the IQR encompass?



Five number summary: min, Q1, median, Q3, max

Mean or median?

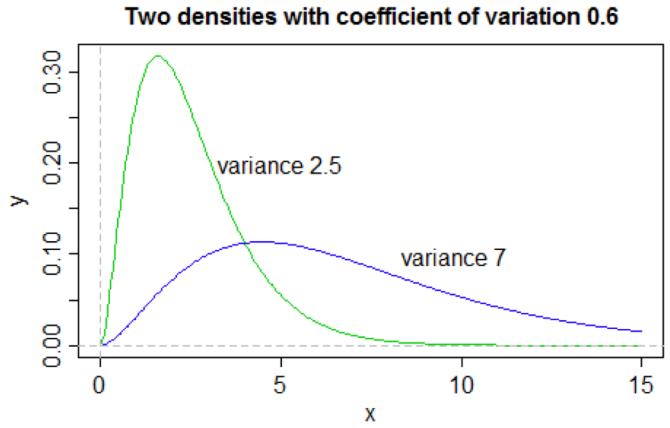
The median is much more robust to outliers compared to the mean.



Measures of variability

Coefficient of variation is the standard deviation of a sample expressed as a percentage of the sample mean (aka normalized)

- $\boldsymbol{COV = \frac{s}{\bar{Y}} \times 100\%}$
- Useful measure for comparing variability between two differently-scaled datasets

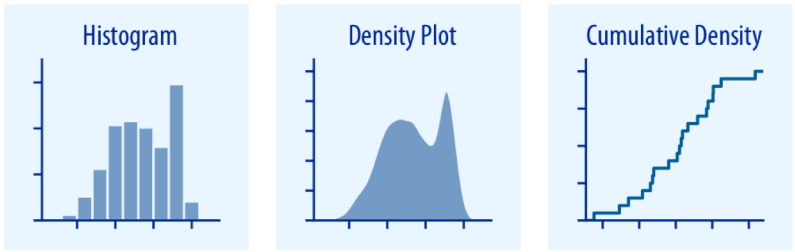


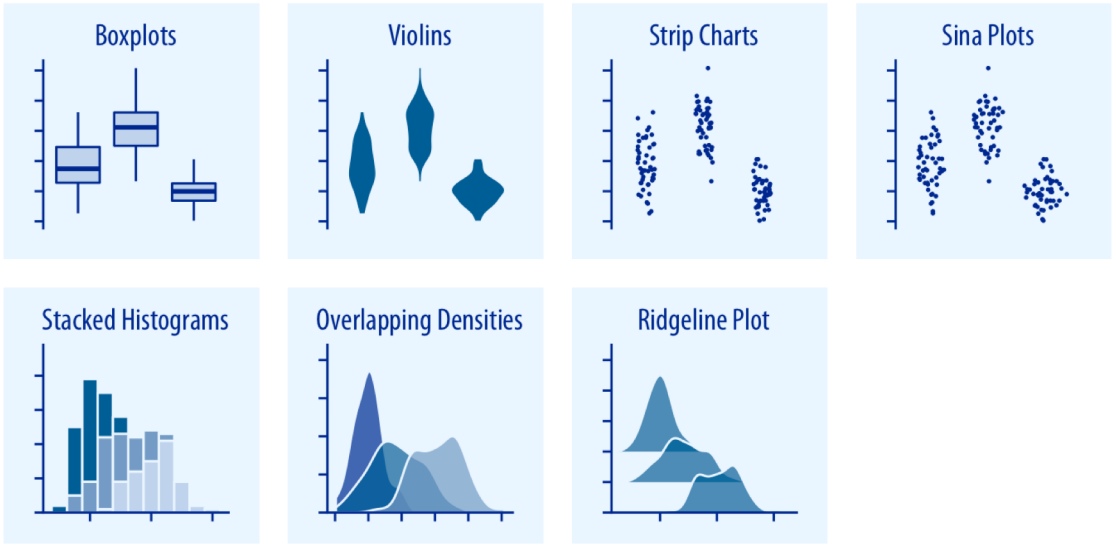
Visualizing data

Different types of plots are used to represent different types of data

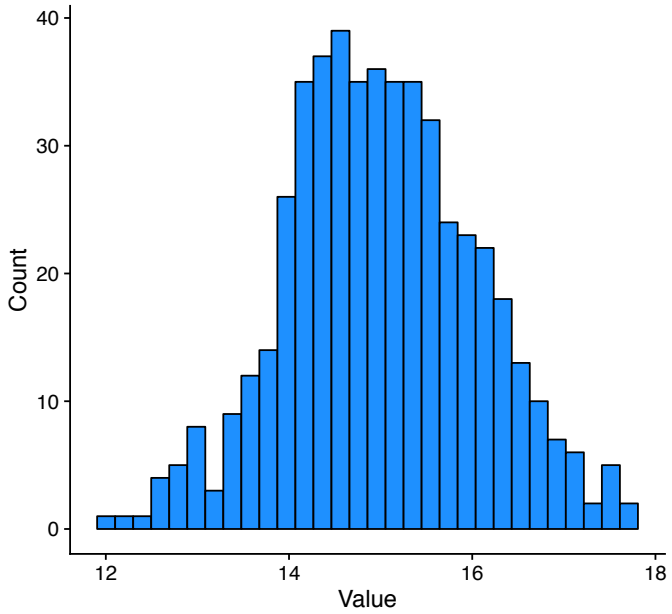
- Continuous data
  - Histogram
  - Density plot
  - Boxplot
  - Violin plot
- Discrete data
  - Bar plot
- Comparing two continuous variables
  - Scatterplot
- Trend over time
  - Line plot

“Directory of data visualizations” (ch5)

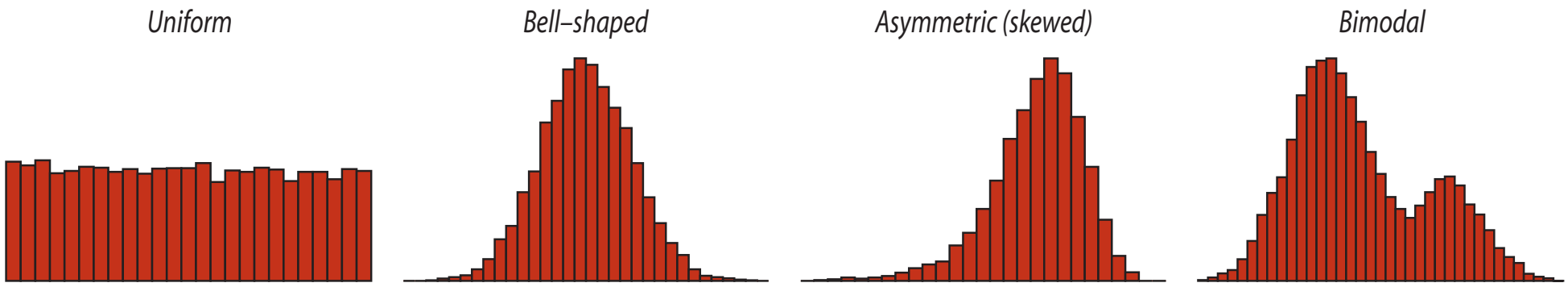




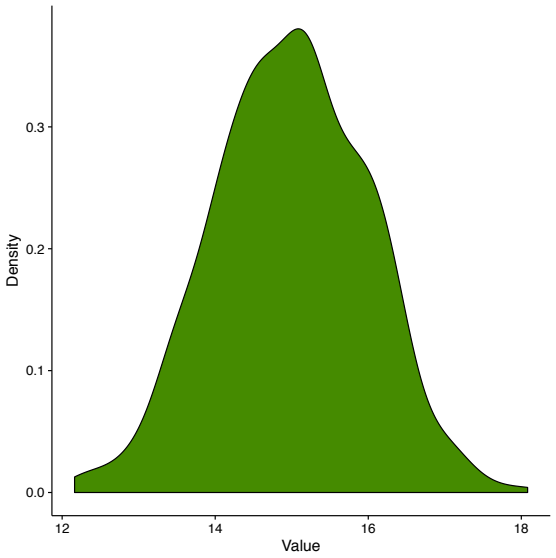
Histogram

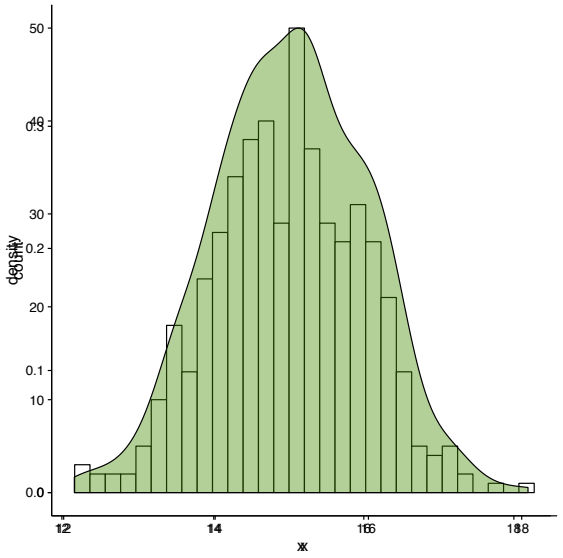


Using histograms to describe distributions



Density plots smoothen histograms

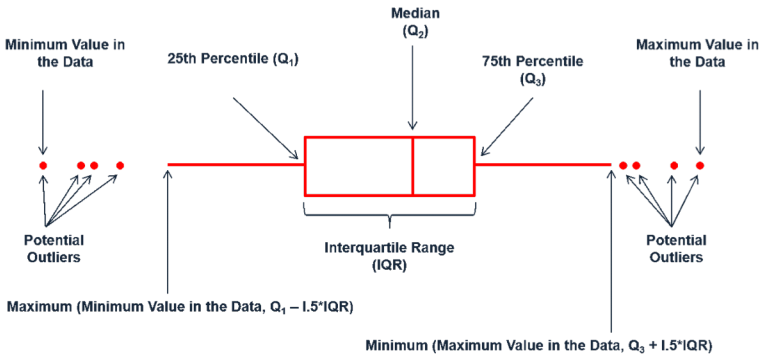


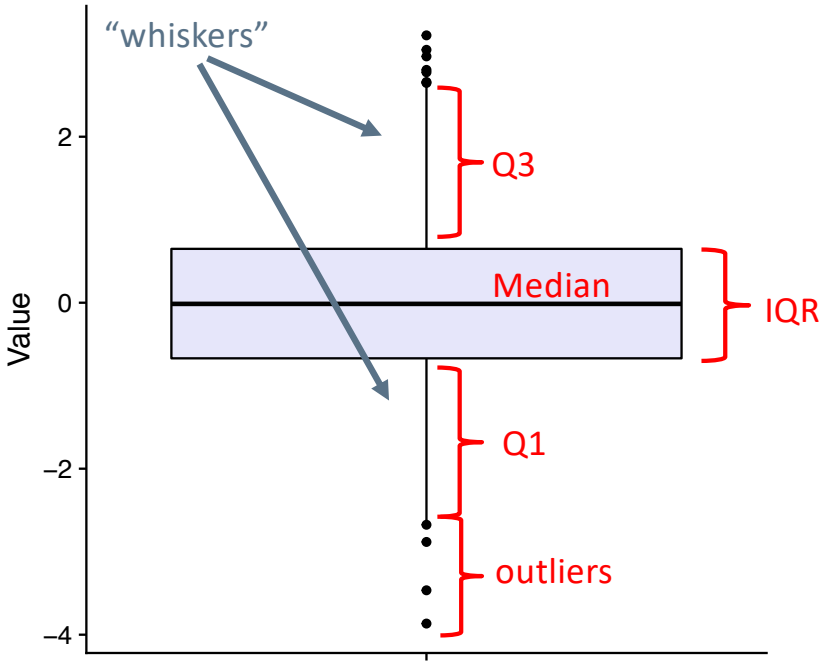


Boxplot

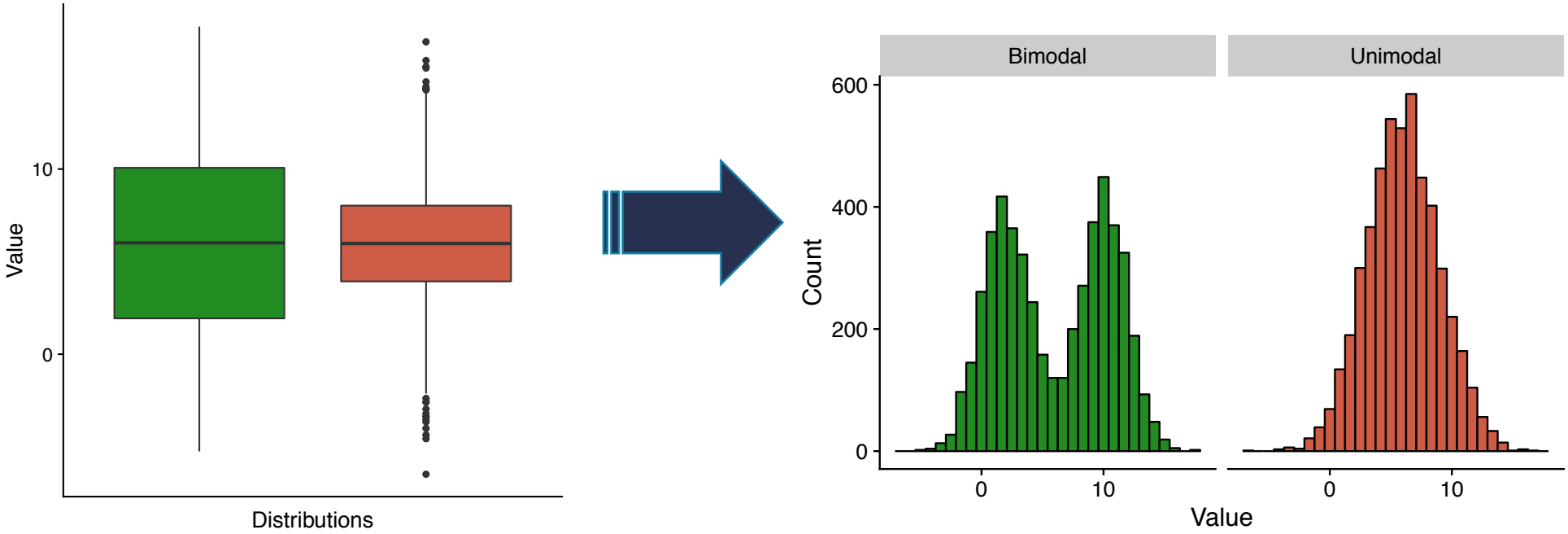
Graphical representation of a five-number summary

"Whiskers" calculated as data within +/- 1.5 IQR



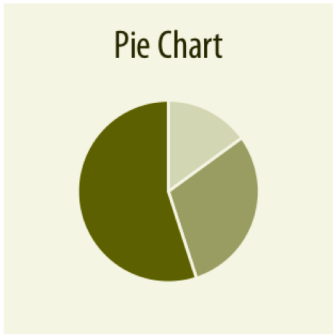


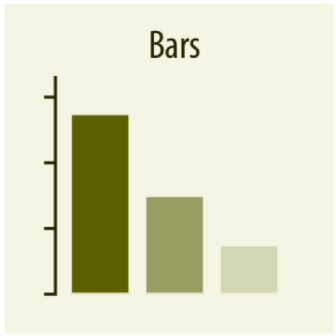
Boxplots: The plot thickens*

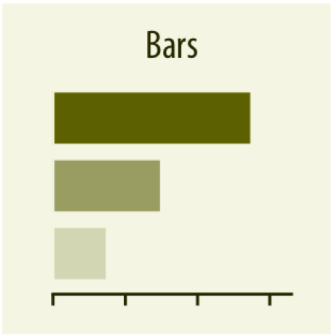


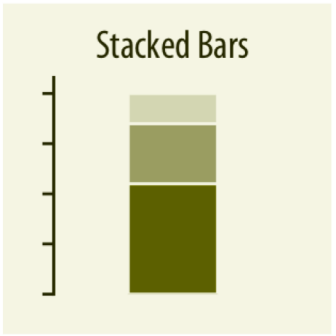
*Pun intended.

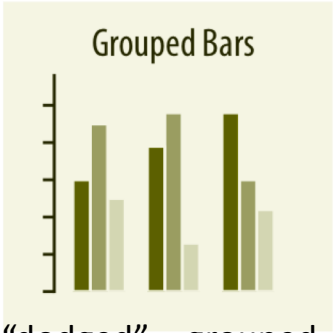
Visualizing amounts/proportions





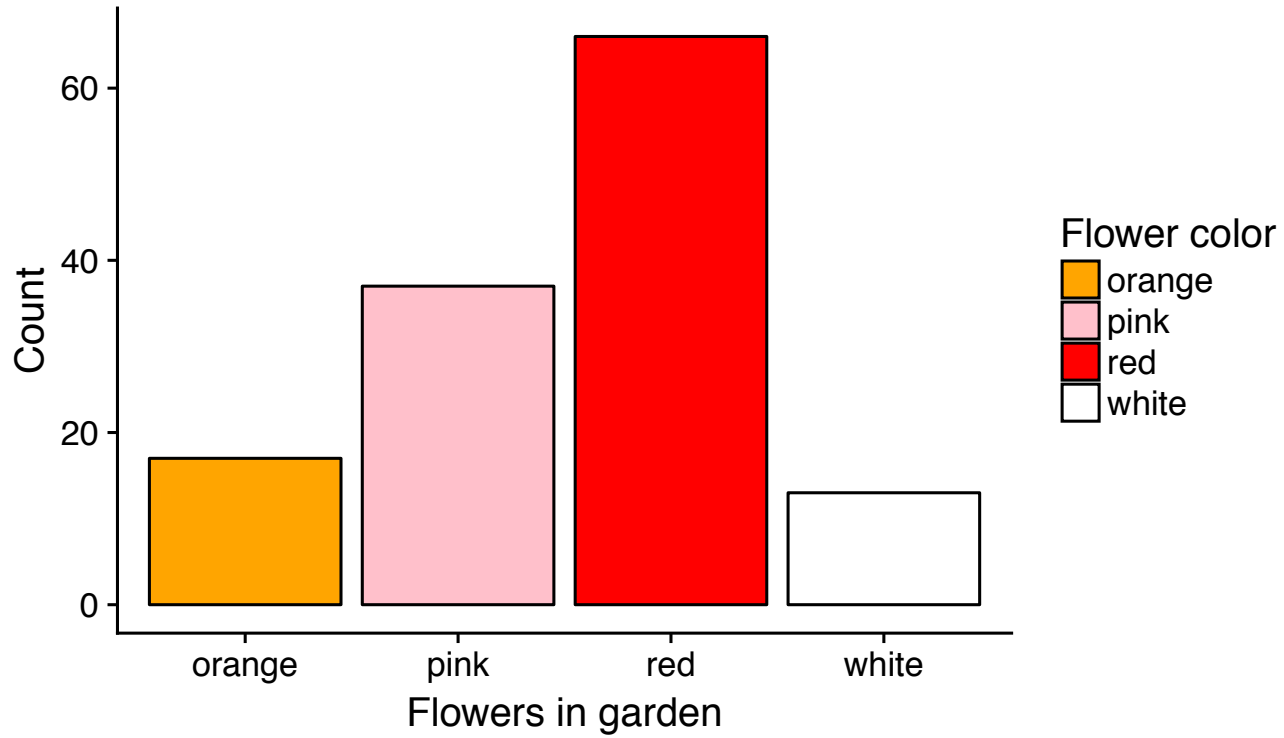




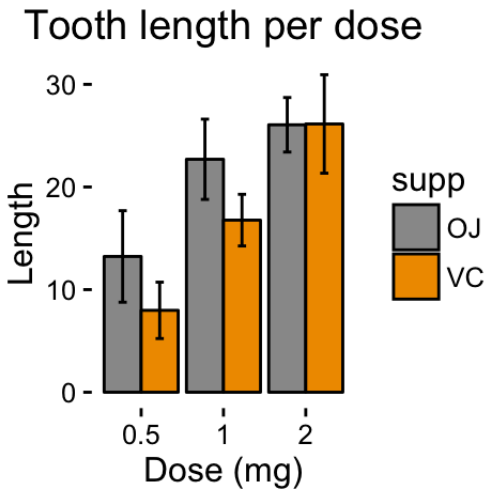


"dodged" = grouped

Barplot

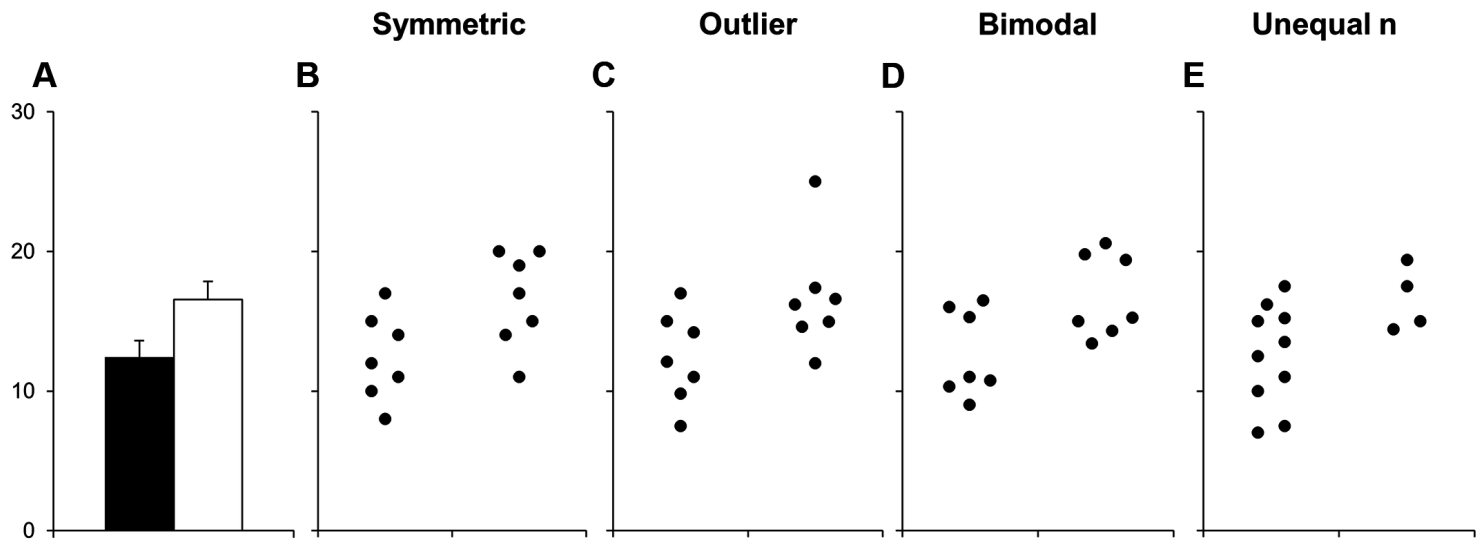

60
40
20
0
Count
orange
pink
red
white
Flowers in garden
Flower color
orange
pink
red
white

Barplots for quantitative data



The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as 'VC').

Why you should NEVER USE barplots for quantitative data

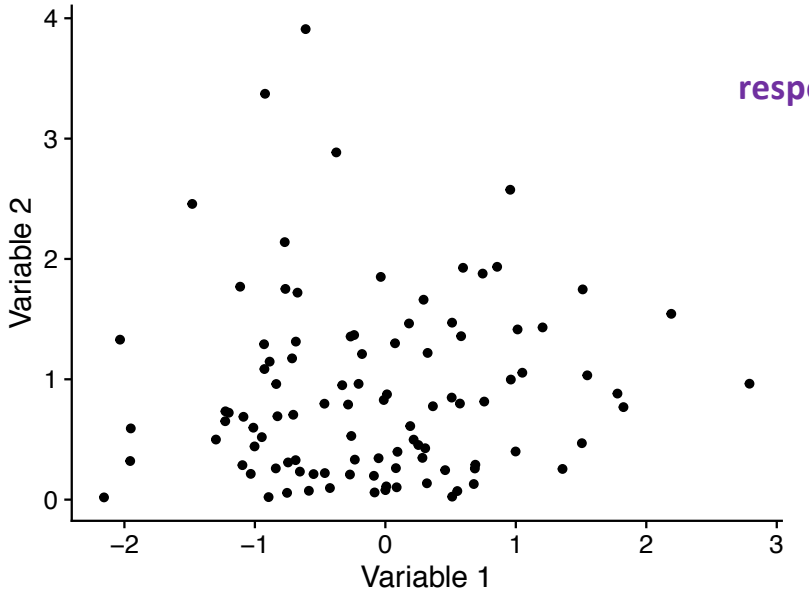


http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

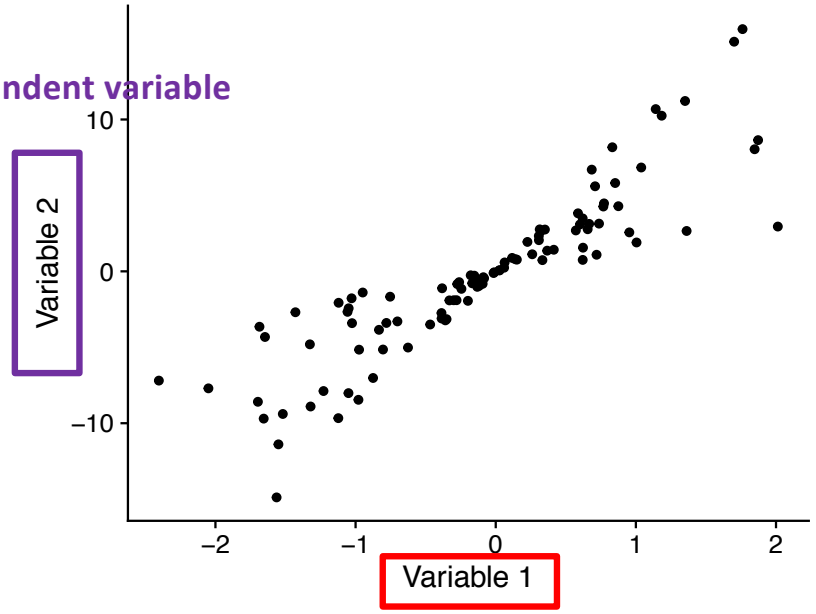
It’s just an awful way to visualize DATA

https://onlinelibrary.wiley.com/doi/full/10.1111/modl.12386

Scatterplot



response/dependent variable



explanatory/independent variable

Time series data