Descriptive and Summary Statistics

BIO5312 FALL2017

STEPHANIE J. SPIELMAN, PHD

Logistics

All course materials will be hosted here: http://sjspielman.org/bio5312_fall2017

Submit assignments via Canvas: https://templeu.instructure.com

Please bring your laptop to class!!!

Office SERC 643

- Weekly office hours Friday 1-3 ground floor of SERC ← vote?

Course goals

The primary goal is to **analyze, interpret, and visualize** data in the biological sciences

Achieved via statistical analysis and data science techniques in R

This is not a course in statistical theory.

Course topics

Descriptive and Summary Statistics

Data visualization

Fundamentals in probability, distributions

Statistical inference: hypothesis testing and confidence intervals

Linear modeling

Multiple testing

Binary classification

Clustering methods

Special topics in current biological data analysis

Course topics

Descriptive and Summary Statistics

Data visualization

Fundamentals in probability, distributions

Statistical inference: hypothesis testing and confidence intervals

Linear modeling

Multiple testing

Binary classification

Clustering methods

Special topics in current biological data analysis

But first, what are we doing here?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

We use statistics to **make inferences about phenomena using samples** and **quantify uncertainty of data**

Biostatistics is (surprisingly!) a branch of applied statistics geared towards to medical and biological problems

Populations and samples

Populations are the entire collection of individuals/units/etc. a researcher is interested in

- Generally we can never know the true composition of a population
- Populations are described with **parameters**

Samples are *subsets* of individuals/units from populations

- We use *hypothesis testing* to (try to) draw population-level conclusions from samples
- Samples are described with **estimates**

Parameters and estimates use different notations, as we will see

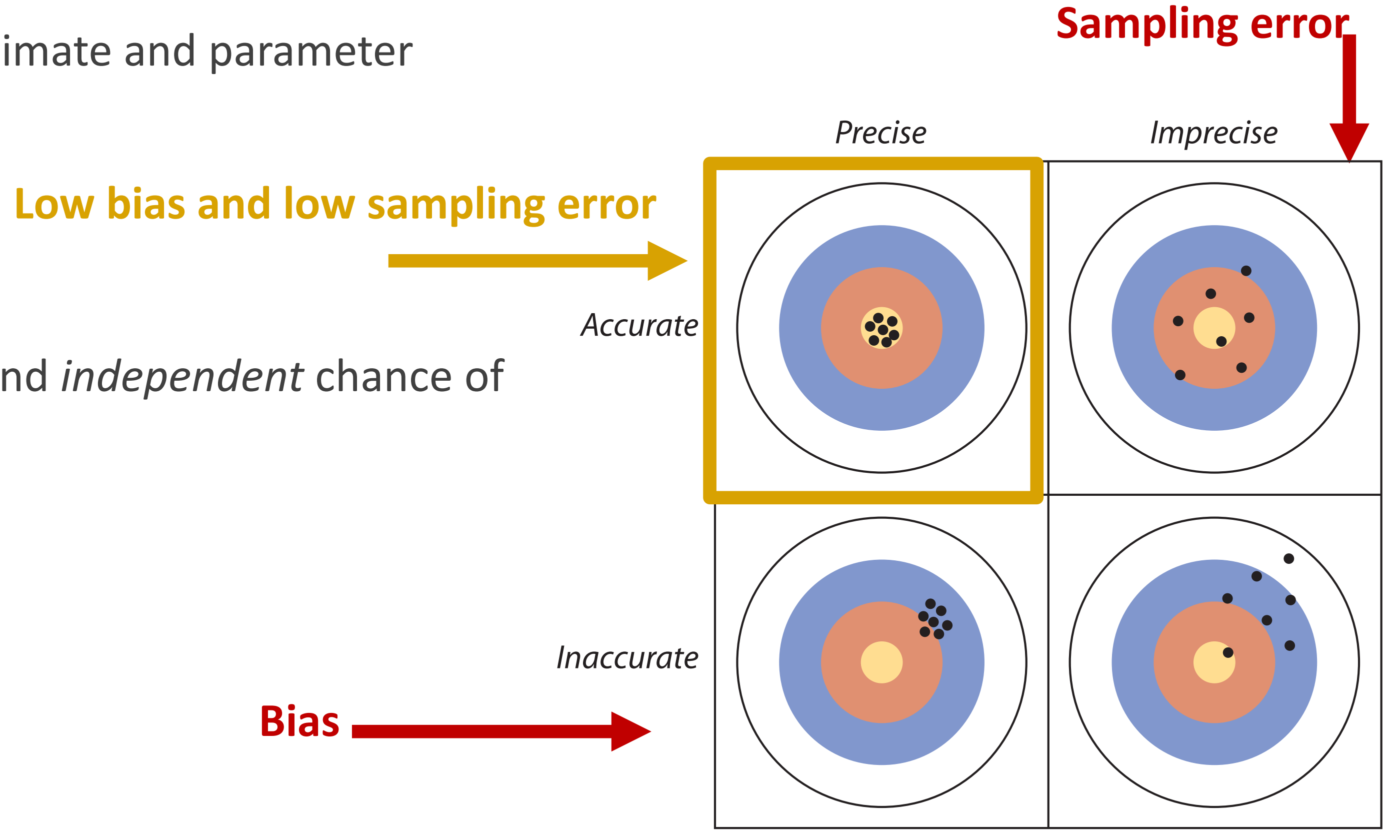
What makes a good sample?

In an ideal world, a sample is *unbiased* and features *low sampling error*

- Bias is a *systematic* discrepancy between estimate and parameter

Samples should be *randomly chosen*

- Each population unit should have an *equal* and *independent* chance of being chosen for a given sample



Sampling error

Low bias and low sampling error

Bias

Pop quiz: Is it random?

A researcher selects the first 58 student volunteers that sign up for a study

A computer program numbers all residents in a community, and then uses a random-number generator to select 26 residents

A researcher vigorously shakes a box containing equally sized balls and takes the first 3 that fall out of the box.

A researcher selects all study participants whose first name starts with an A, B, K, M, or O.

Pop quiz: Is it random?

~~A researcher selects the first 58 student volunteers that sign up for a study~~

A computer program numbers all residents in a community, and then uses a random-number generator to select 26 residents

A researcher vigorously shakes a box containing equally sized balls and takes the first 3 that fall out of the box.

~~A researcher selects all study participants whose first name starts with an A, B, K, M, or O.~~

Descriptive and Summary Statistics

Tools to concisely describe data, numerically and visually

Generally the first step in data exploration and statistical analysis

- Identify missing values, outliers, etc.
- Check assumptions required to fit models or perform statistical tests
- Identify trends that merit further study

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*

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 in a protein…)

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

Bonus: names of sex genotypes?

Measures of Location

Continuous

Mean

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

Median

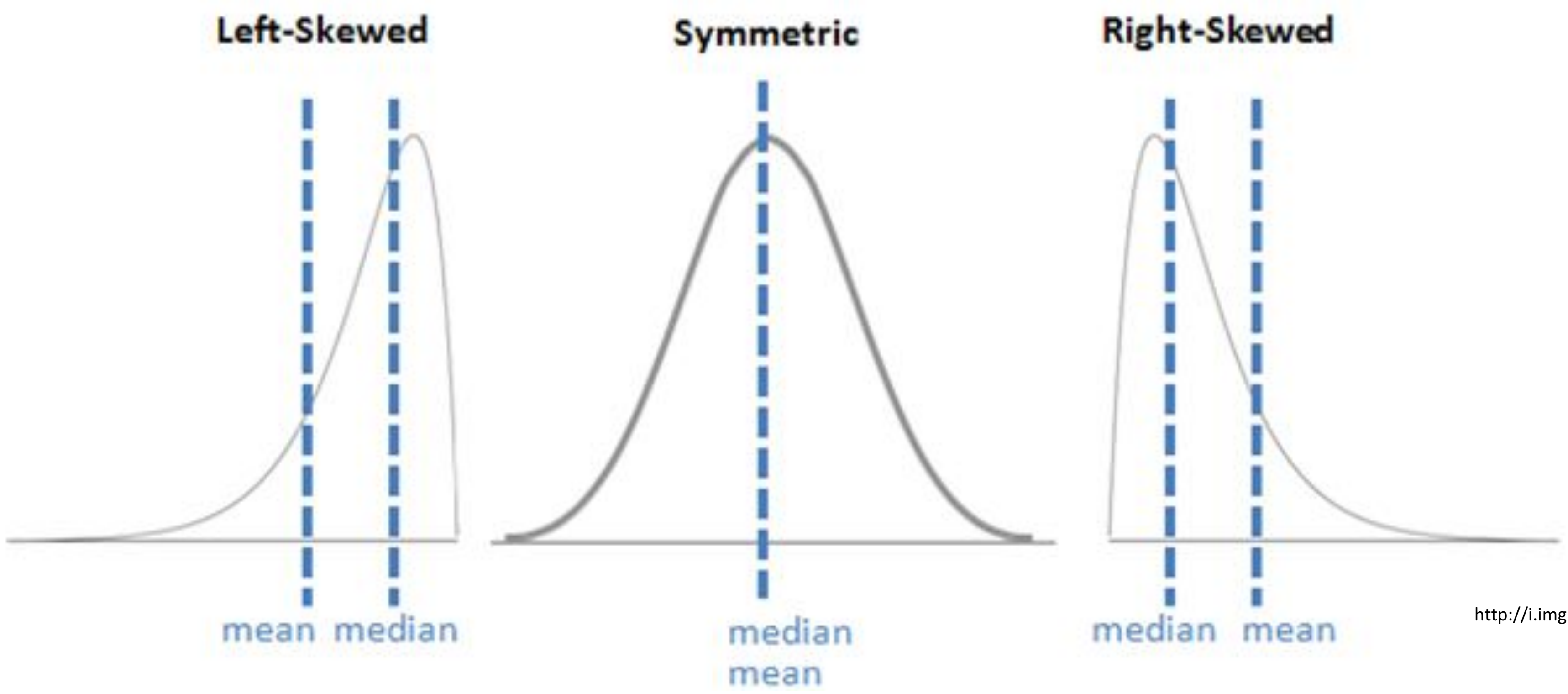
- For odd *n,* the $\left(\frac{n+1}{2}\right)$*th* observation
- For even *n,* the average of the $\left(\frac{n}{2}\right)$*th* and $\left(\frac{n}{2} + 1\right)$*th* observation

Discrete

Mode

- The most frequent appearing observation in the distribution (commonly used for discrete data)
- 1, 2, 2, 2, 3, 4, 4, 5, 6 → **2**

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 $\bar{Y}$:

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

Standard deviation of a sample

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

Variance

- s^2

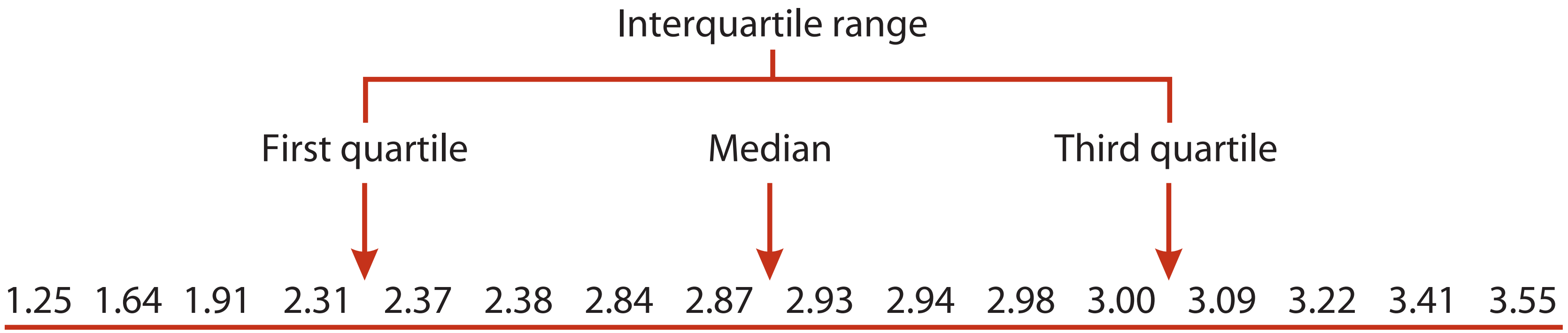
Interquartile range

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

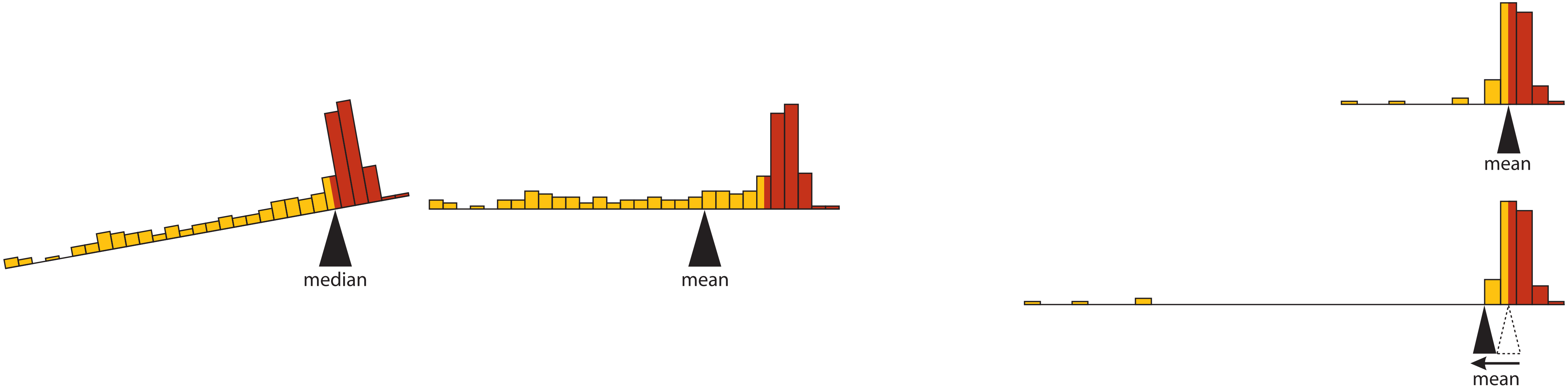
- How much of the data does the IQR encompass?



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

Mean or median?

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

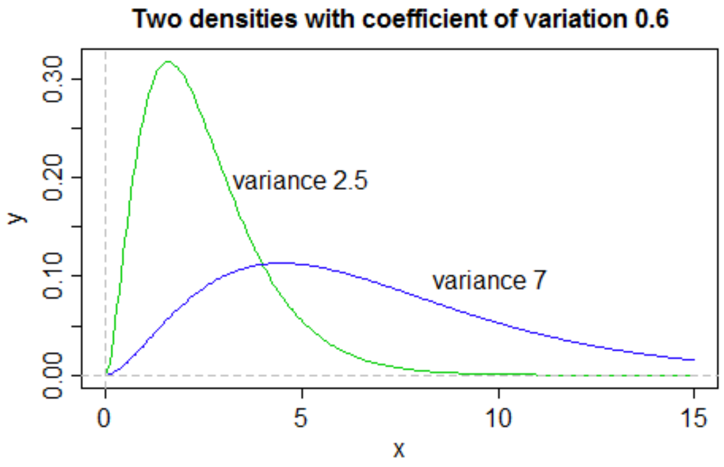


Which would you choose for a *symmetric* distribution and why?

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



Sample vs population notation

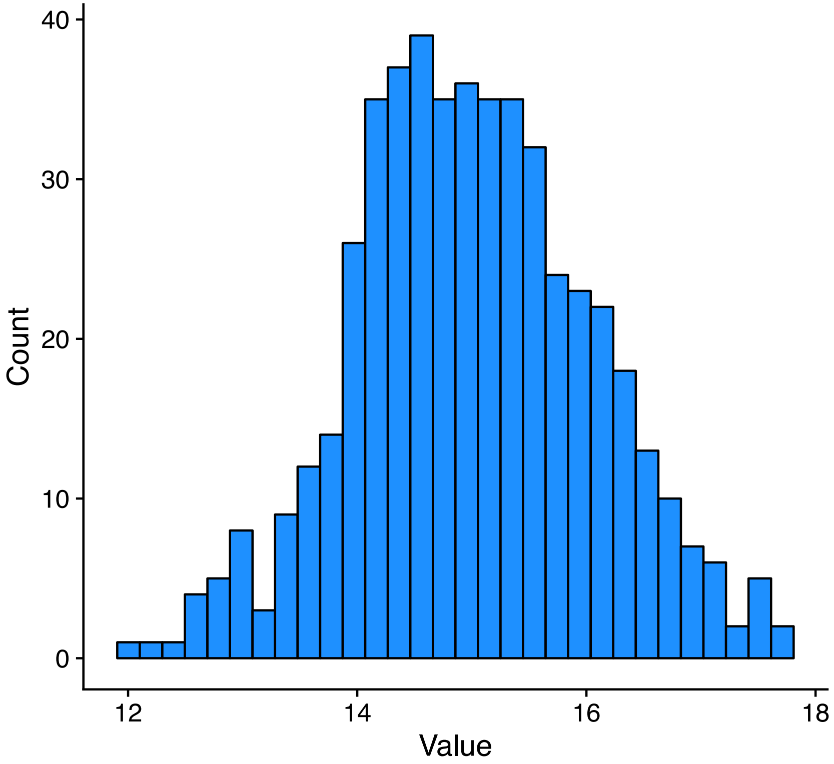
Measurement	Sample estimate	Population parameter
Mean	$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$	$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$
Standard deviation	$s = \frac{1}{n-1} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$	$\sigma = \frac{1}{n} \sqrt{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2}$
Variance	s^2	σ^2

Visualizing data

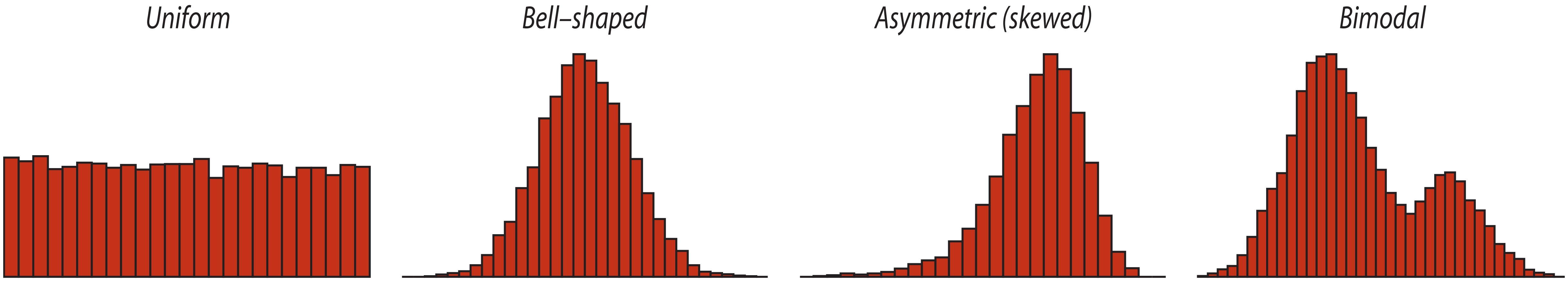
Different types of plots are used to represent different types of data

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

Histogram



Using histograms to describe distributions



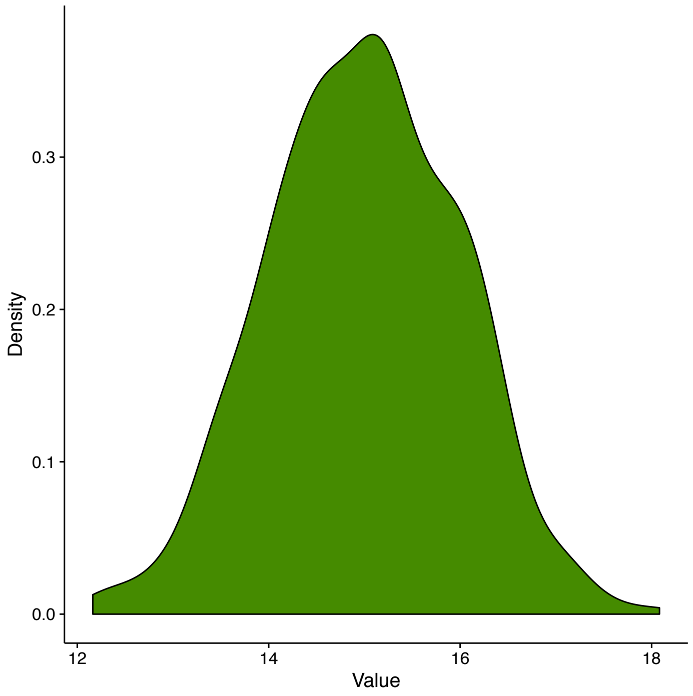
Uniform

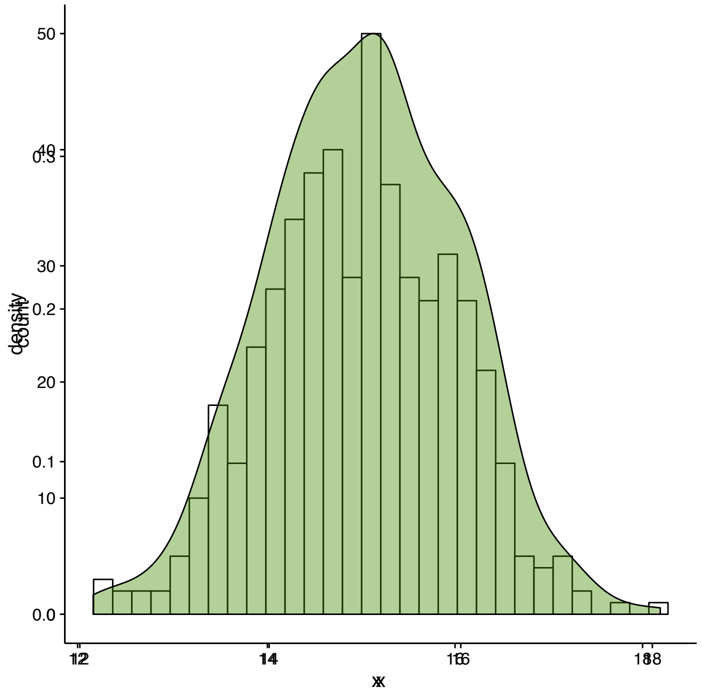
Bell–shaped

Asymmetric (skewed)

Bimodal

Density plots smoothen histograms

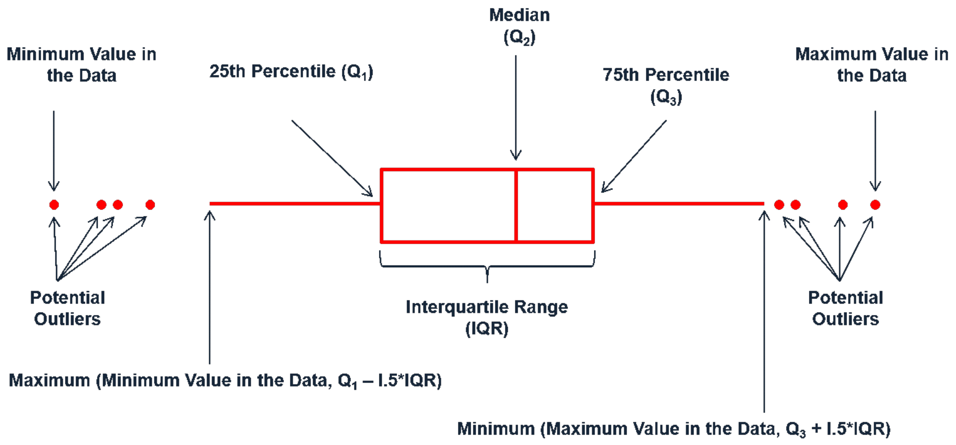


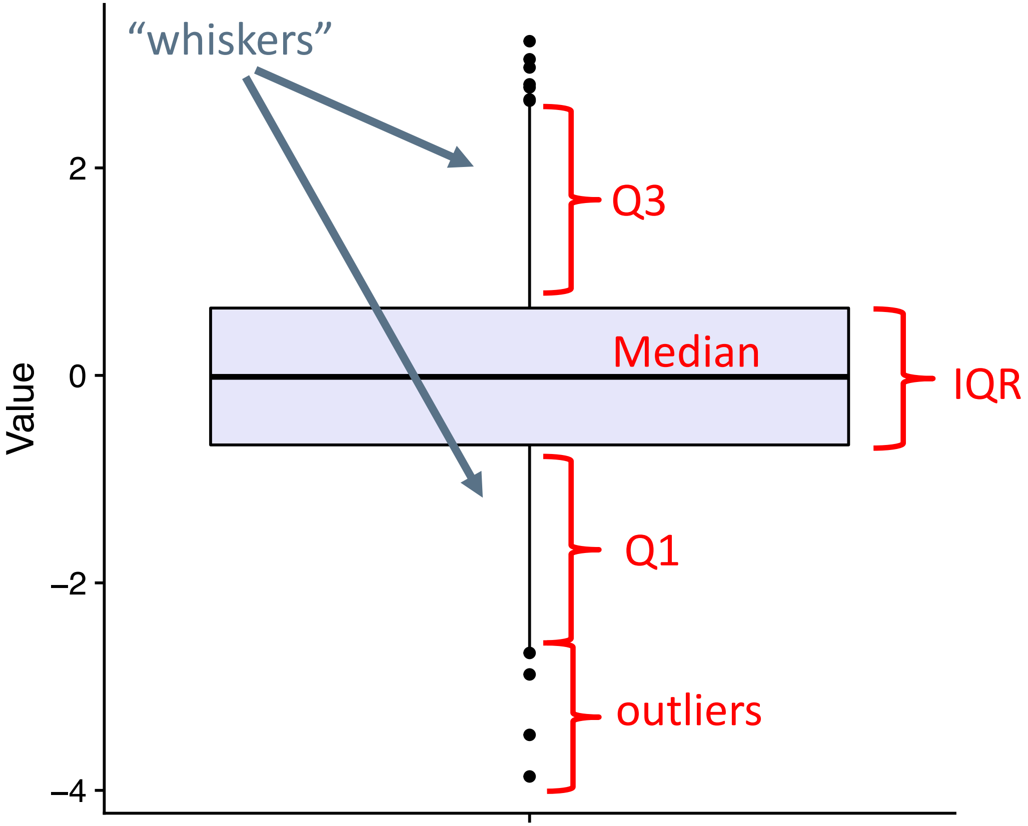


Boxplot

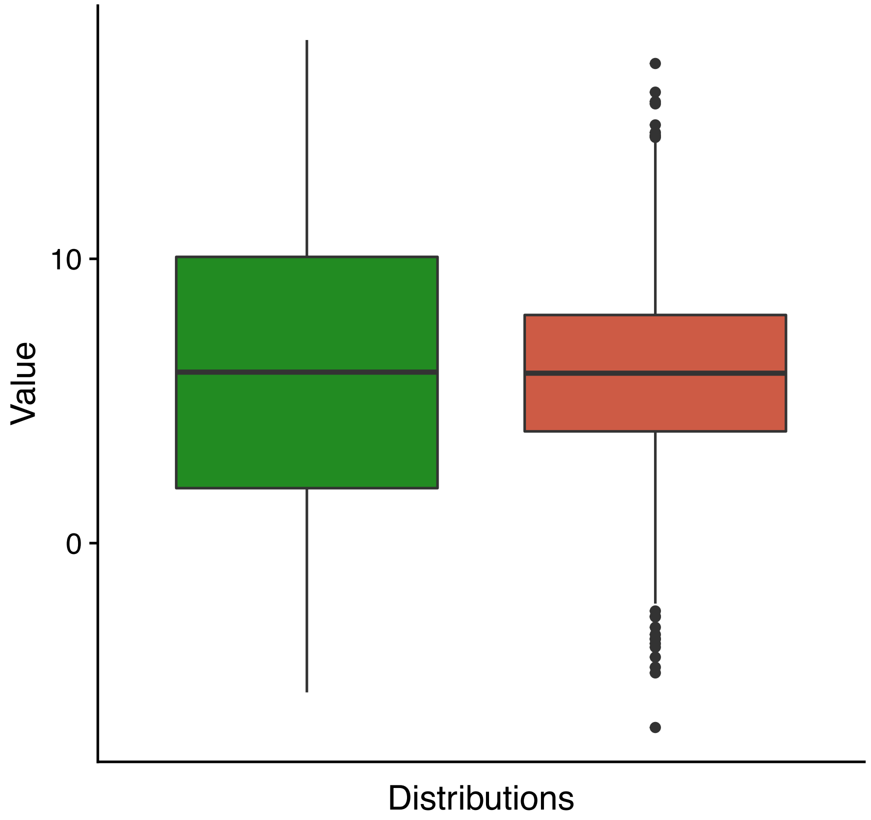
Graphical representation of a five-number summary

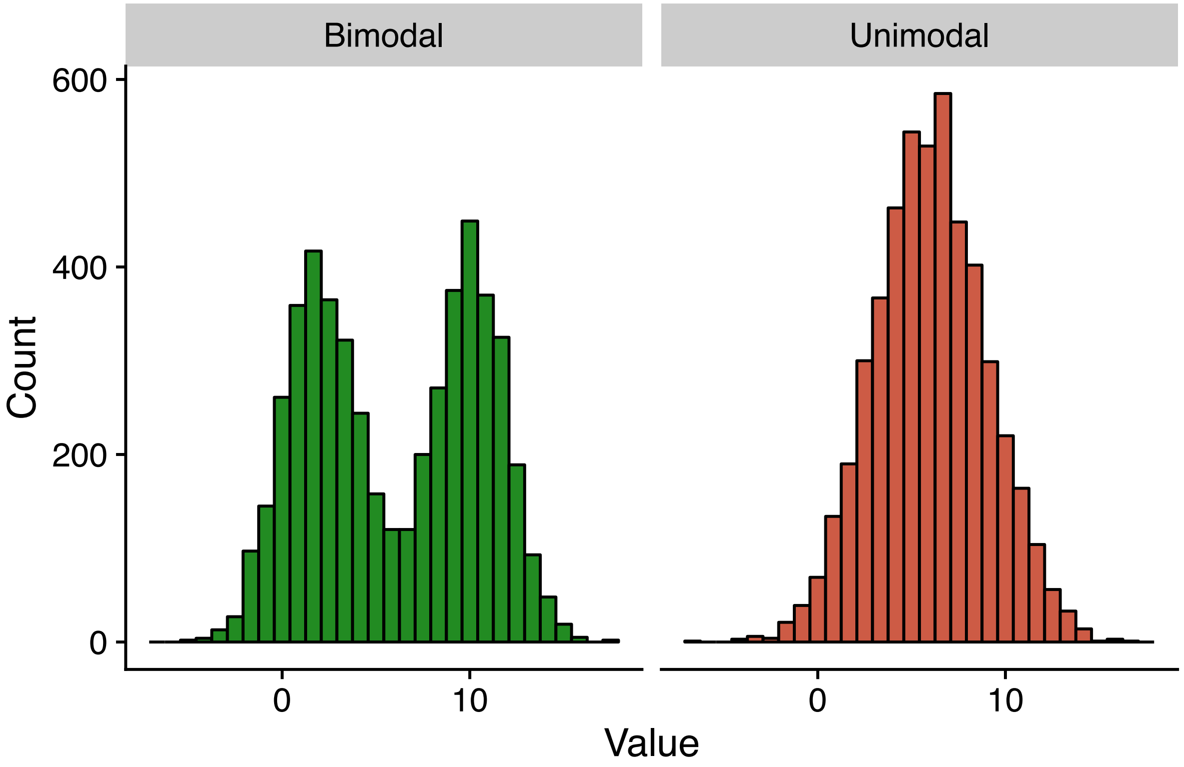
“Whiskers” calculated as data within +/- 1.5 IQR


Median (Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Maximum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Minimum (Maximum Value in the Data, Q_3 + 1.5*IQR)



Boxplots: The plot thickens*

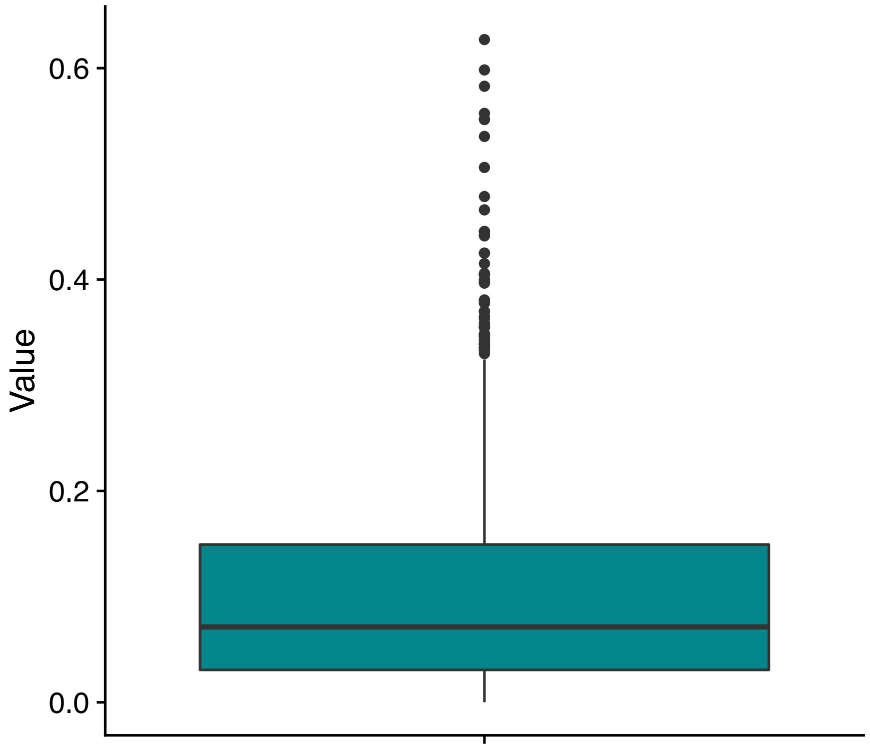




*Pun intended.

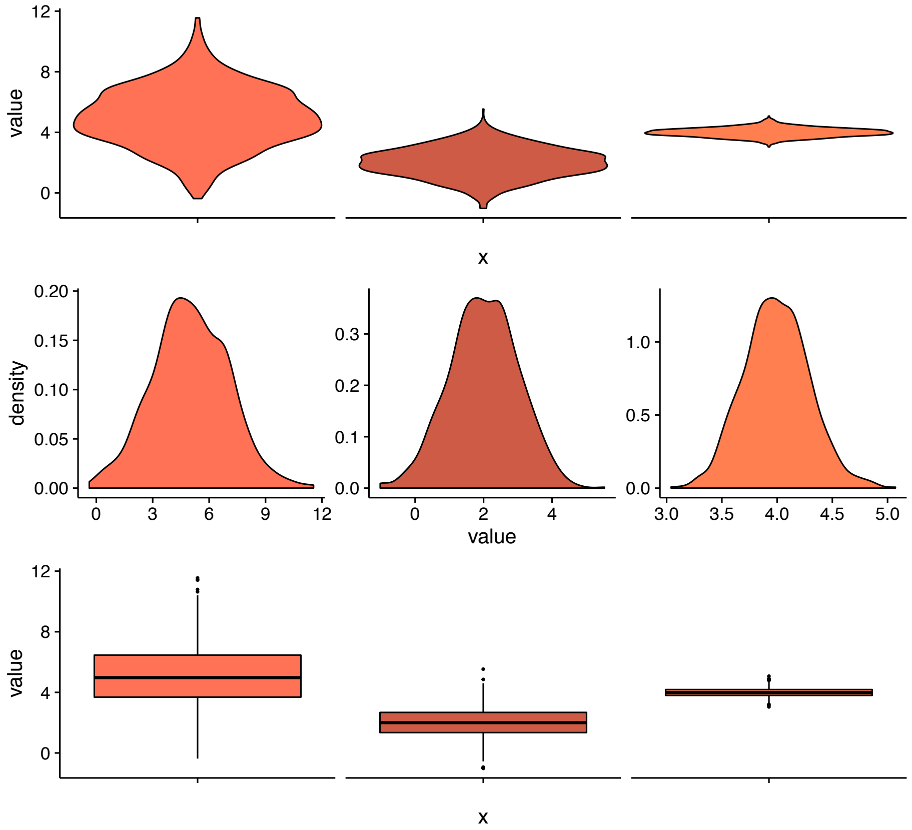
What can we say about this distribution based on its boxplot?

Symmetry? Asymmetric
Skewness? Left-skewed
Modality? Unclear



Violin plot: Density meets boxplot

N(5, 4) N(2, 1) N(4, 0.09)

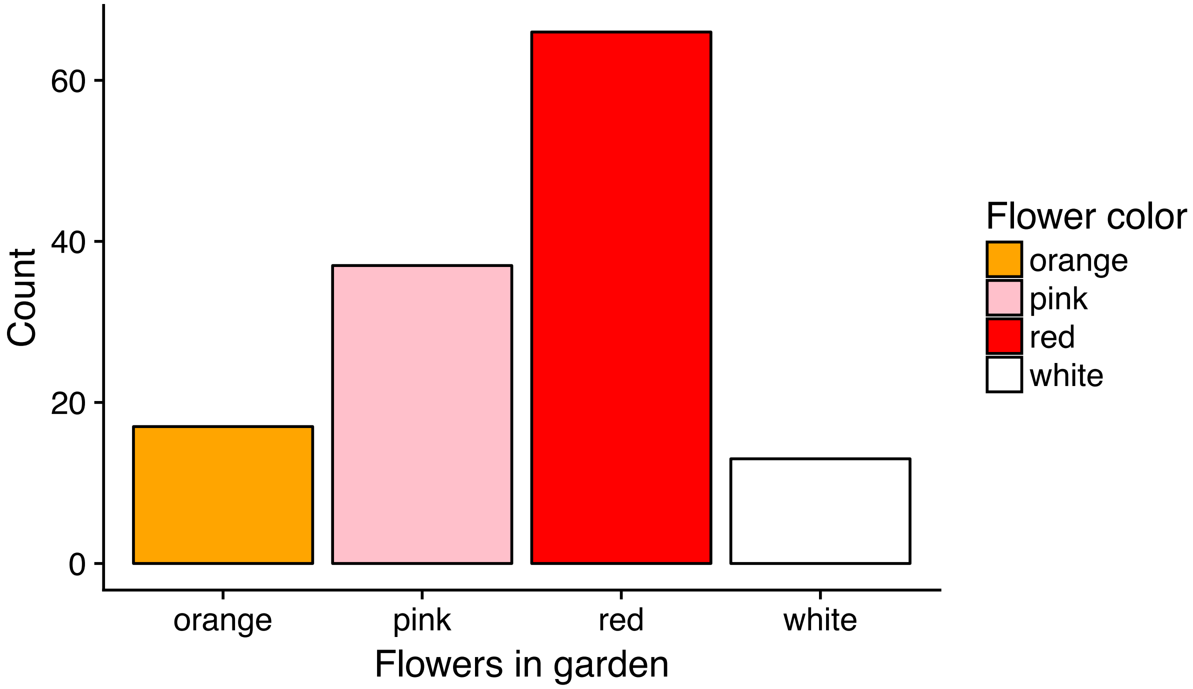


Violin plot

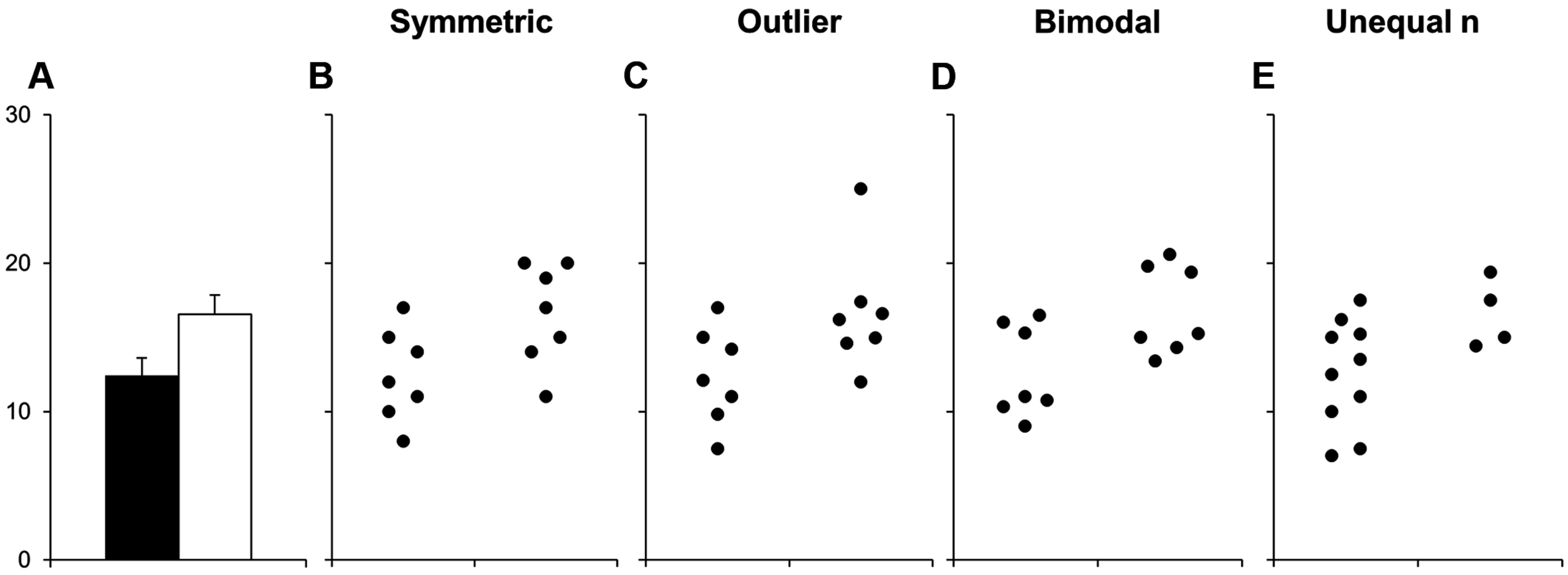
Density plot

Boxplot

Barplot

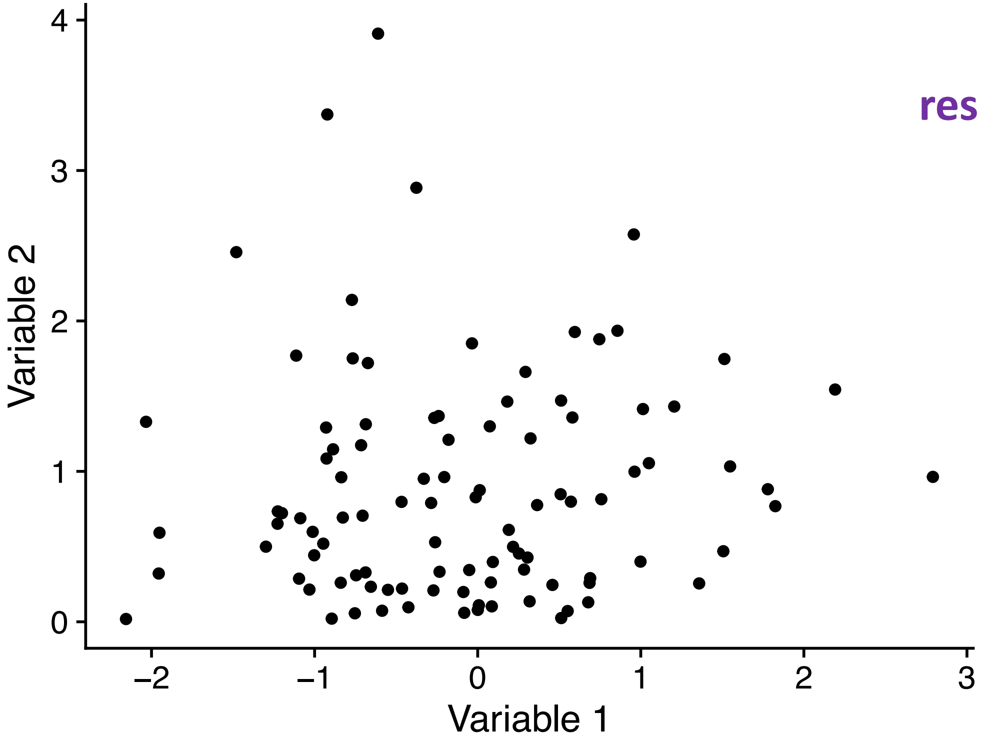

60
40
20
0
Count
orange
pink
red
white
Flowers in garden
Flower color
orange
pink
red
white

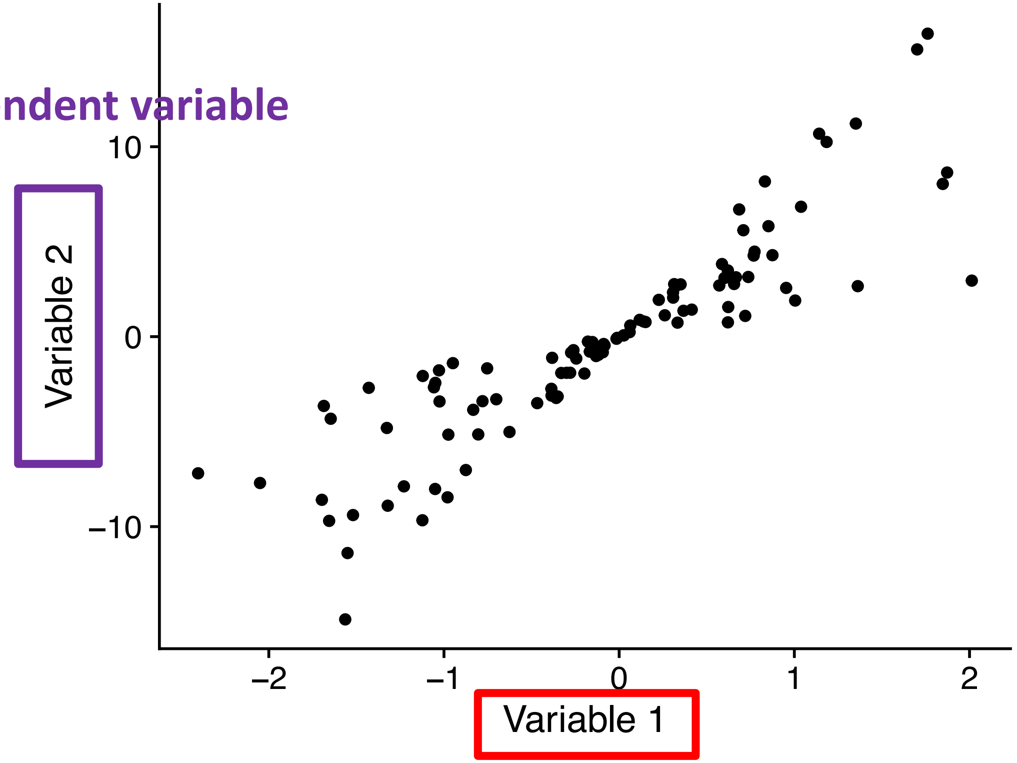
Cautionary tale in barplots



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

Scatterplot

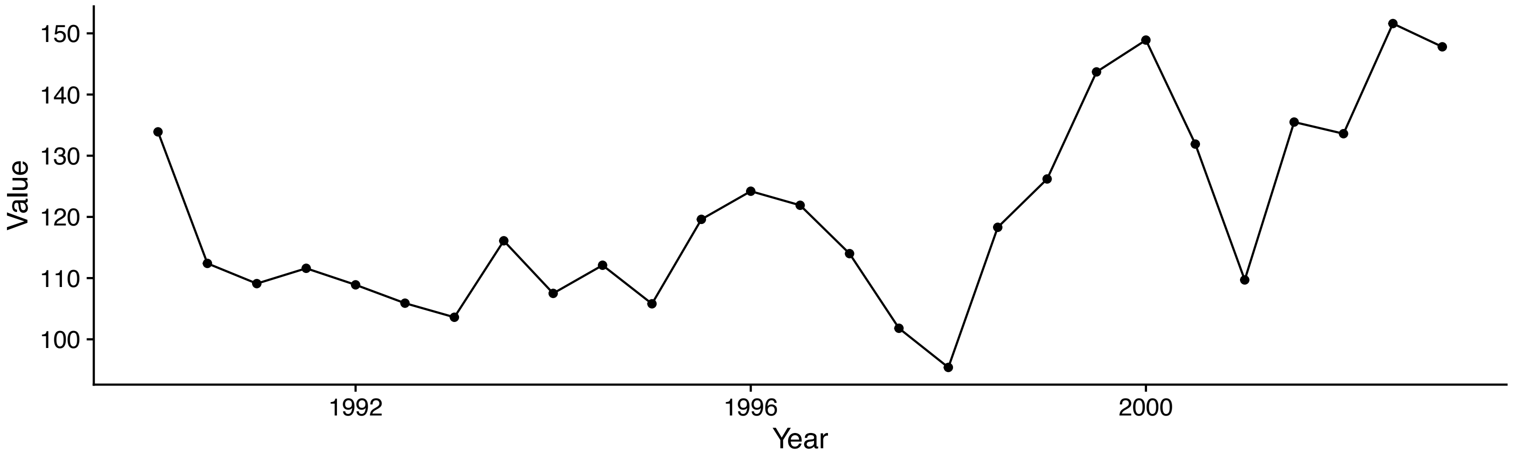


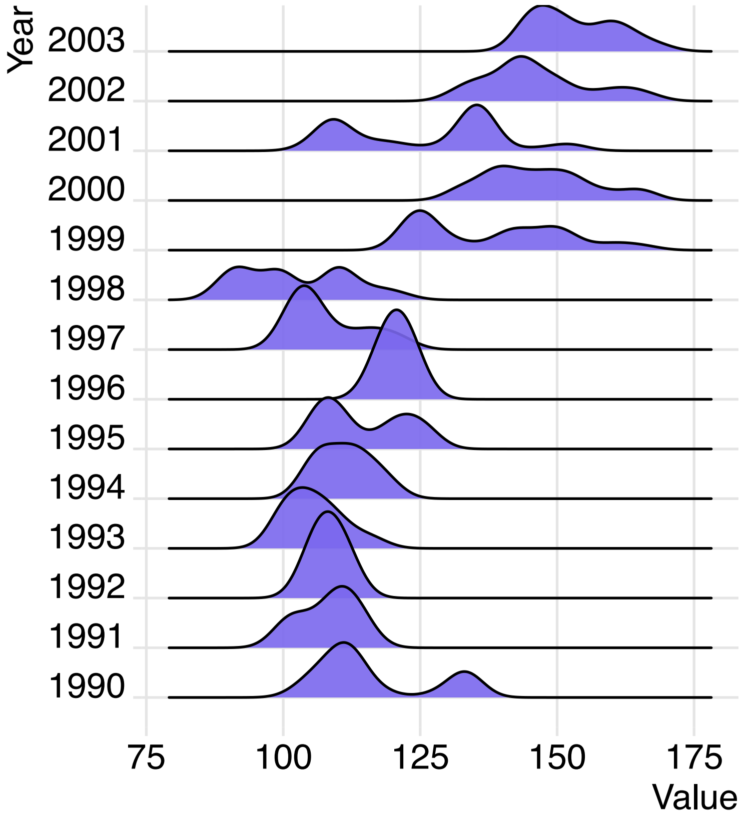


response/dependent variable

explanatory/independent variable

Time series data





BREAK