Summarizing and Presenting Data

Summary statistics

Location / Center	• mean (average) • median • mode • geometric mean • harmonic mean
Scale	• standard deviation (SD) • inter-quartile range (IQR) • range
Other	• quantile • quartile • quintile

Summary statistics

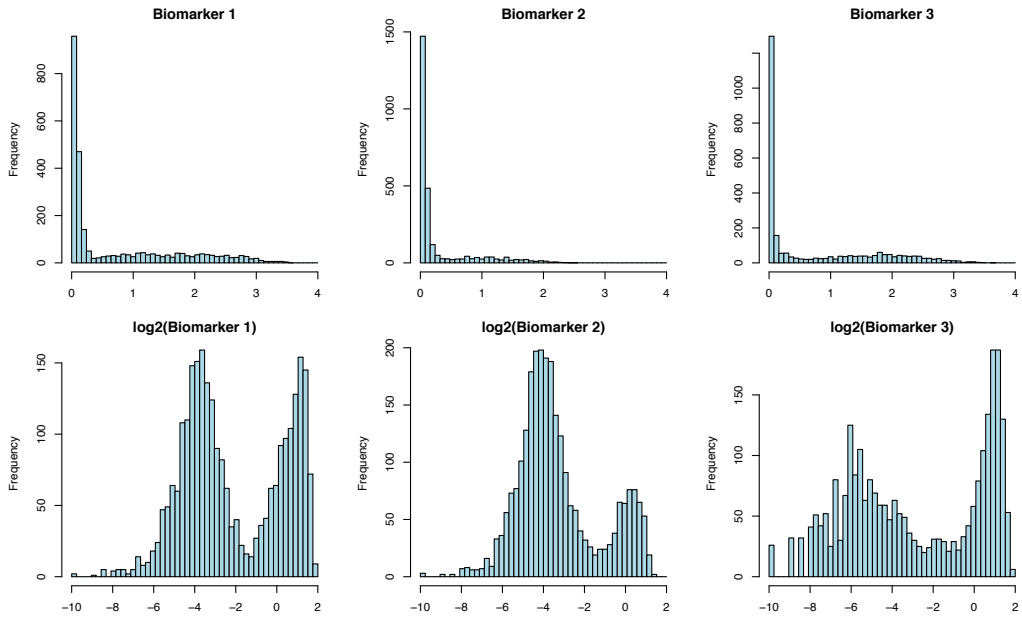
$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = (x_1 + x_2 + \ldots + x_n)/n$$

$$\text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log x_i\right\}$$

$$\text{harmonic mean} = 1/\left\{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)\right\}$$

$\longrightarrow$ Note: these are all sample means.

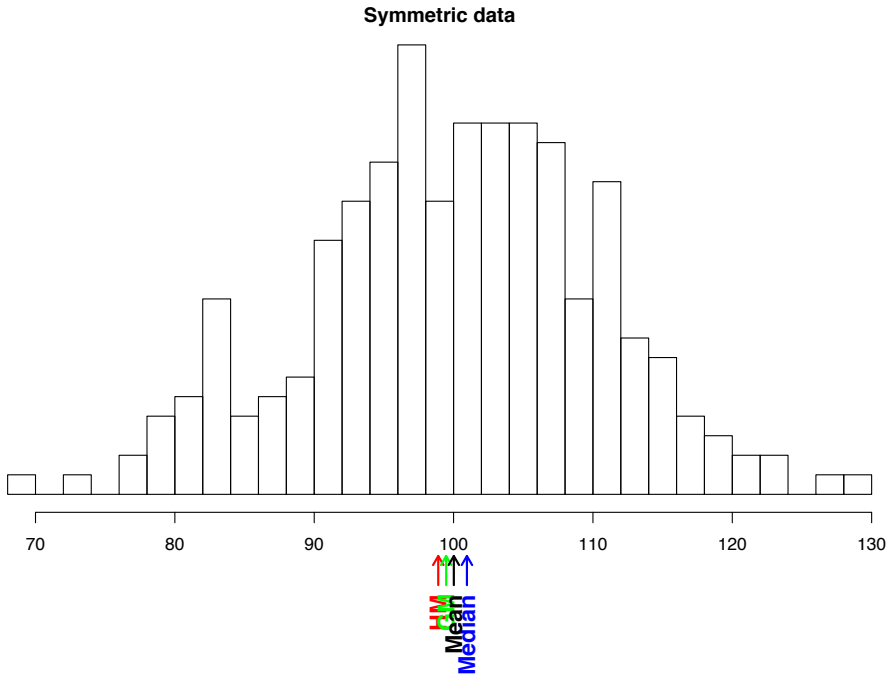
Measures of location / center



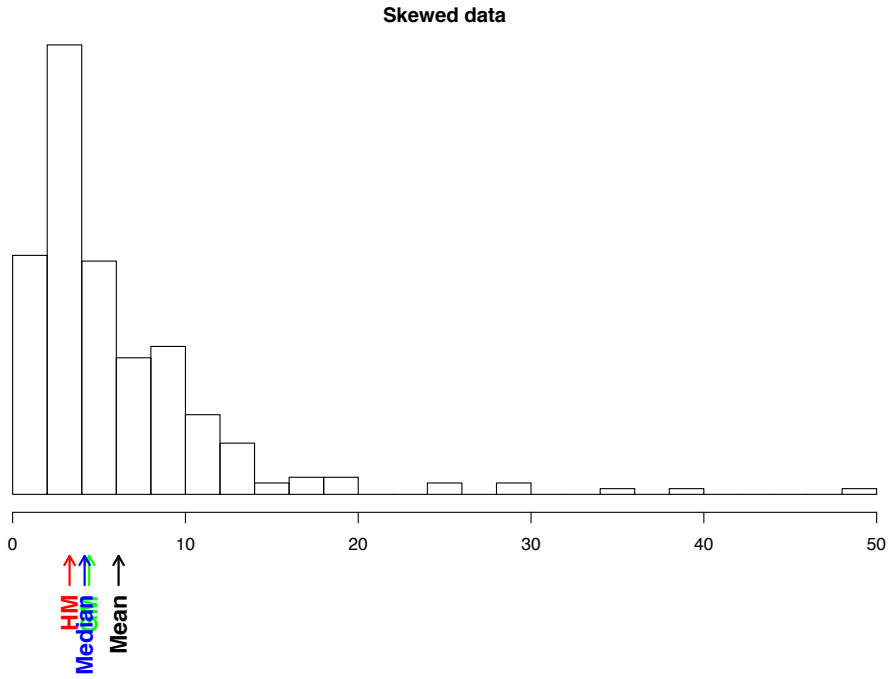
Measures of location / center

- Forget about the mode.
- The mean is sensitive to outliers.
- The median is resistant to outliers.
- The geometric mean is used when a logarithmic transformation is appropriate (for example, when the distribution has a long right tail).
- The harmonic mean may be used when a reciprocal transformation is appropriate (very seldom).

Measures of location / center



Measures of location / center



A key point

The different possible measures of the ”center” of the distribution are all allowable.

You should consider the following though:

- ⟶ Which is the best measure of the ”typical” value in your particular setting?
- ⟶ Be sure to make clear which ”average” you use.

Standard deviation (SD)

Sample variance = $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$

Sample SD = $\sqrt{s^2} = s$

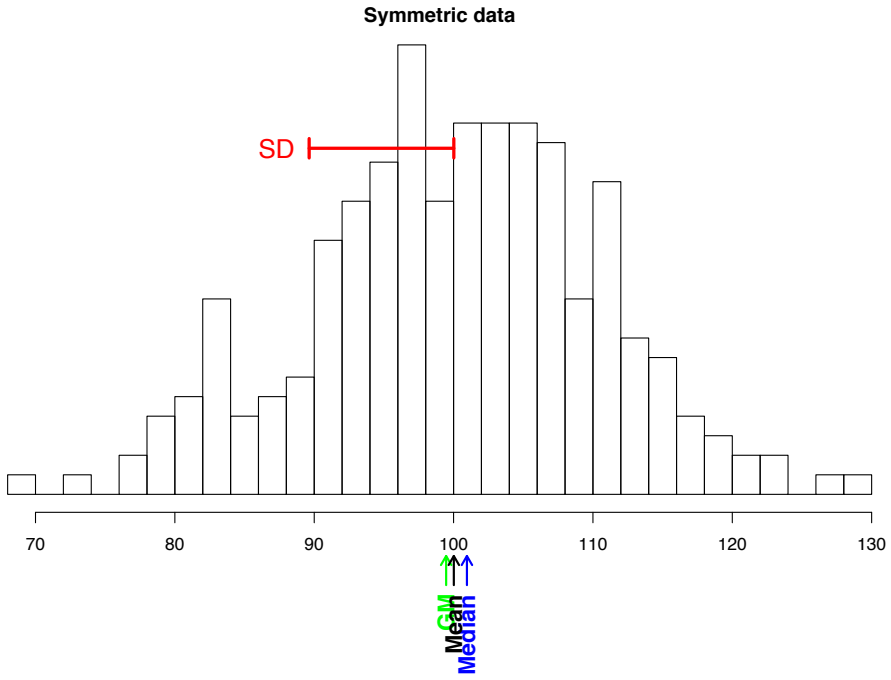
= RMS (distance from average)

= "typical" distance from the average

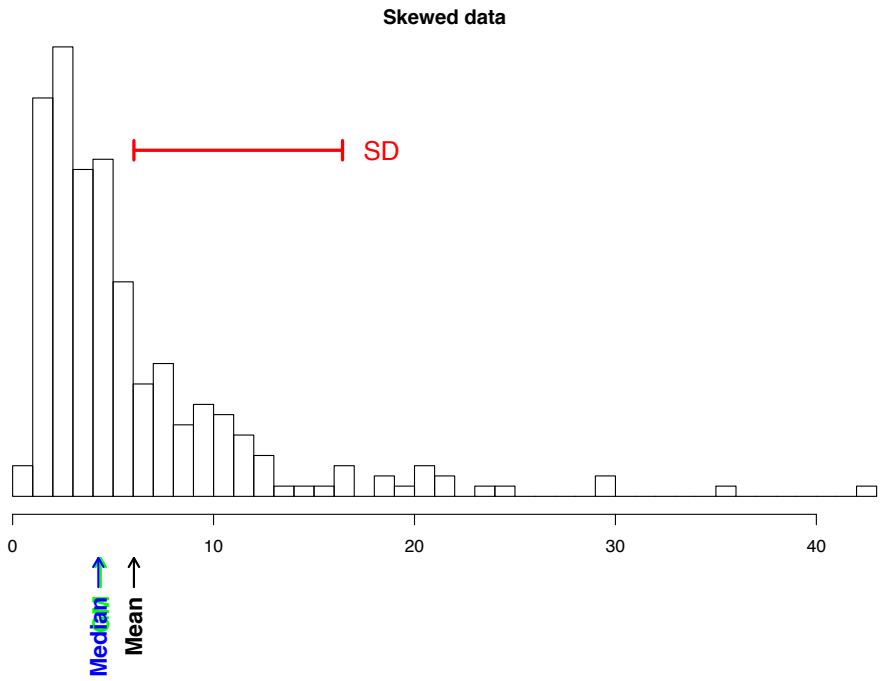
= sort of like ave$\{|x_i - \bar{x}|\}$

$\longrightarrow$ Remember: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

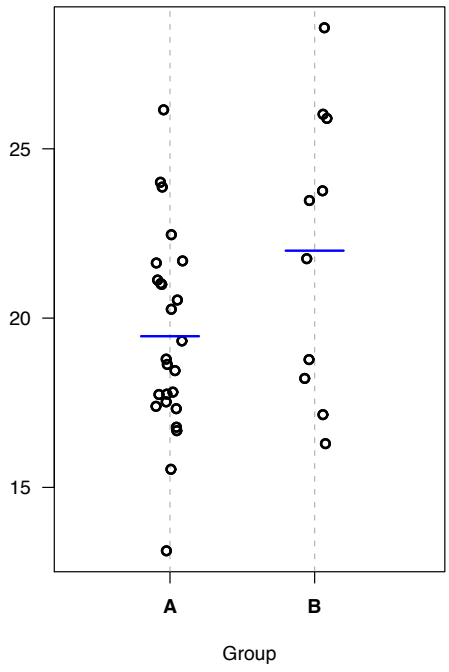
Standard deviation (SD)



Standard deviation (SD)

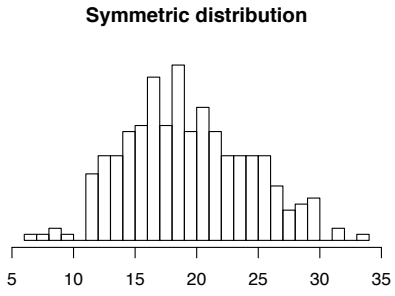


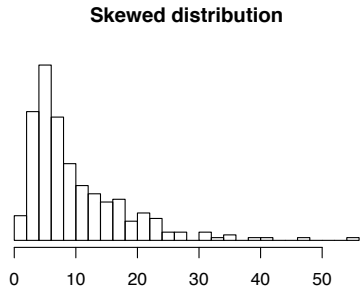
Dotplots



- Few data points per group.
- Possibly many groups.

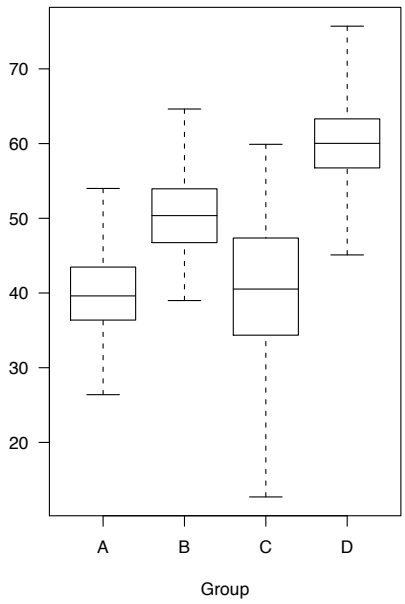
Histograms





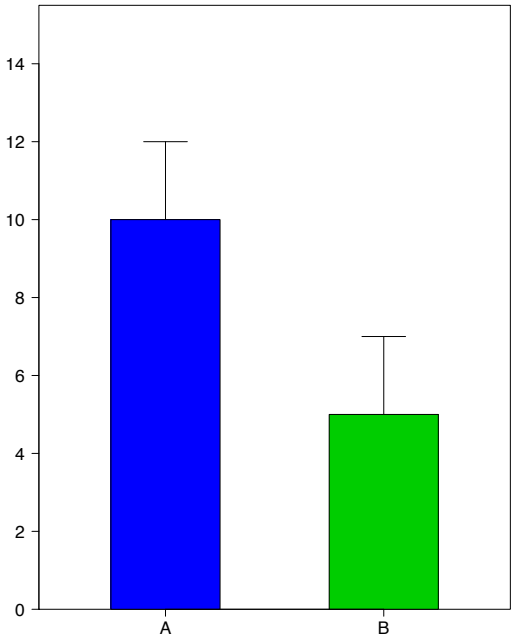
- Many data points per group.
- Few groups.
- Area of the rectangle is proportional to the number of data points in the interval.
- Typically $2\sqrt{n}$ bins is a good choice.

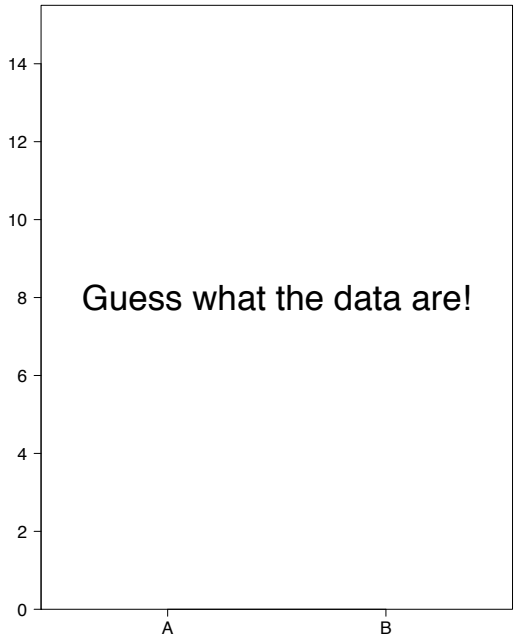
Boxplots



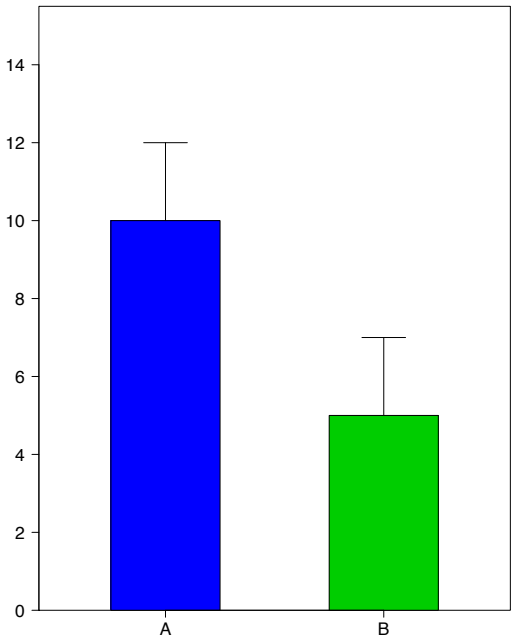
- Many data points.
- Possibly many groups.
- Displays the minimum, lower quartile, median, upper quartile, and the maximum.

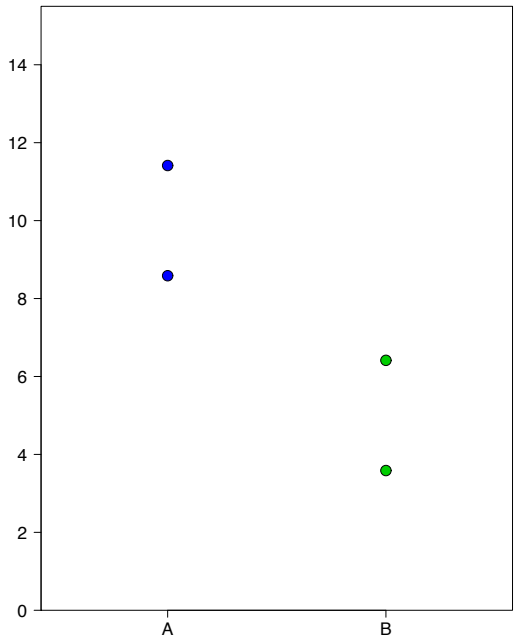
Skyscraper-with-antenna plots



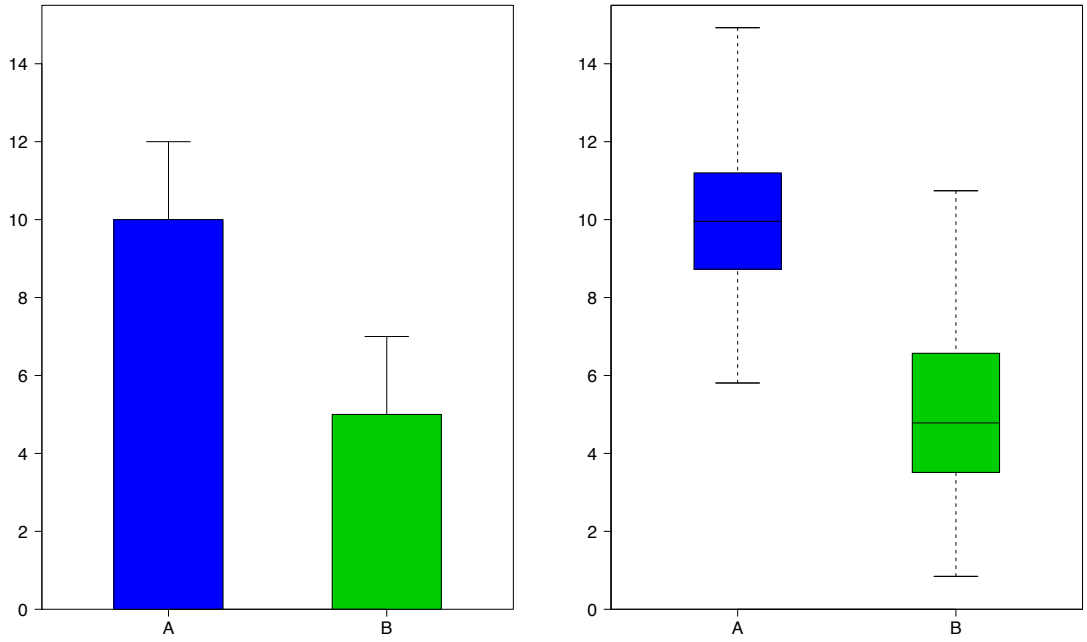


Skyscraper-with-antenna plots





Skyscraper-with-antenna plots



Skyscraper-with-antenna plots