Distributions of data

Distribution

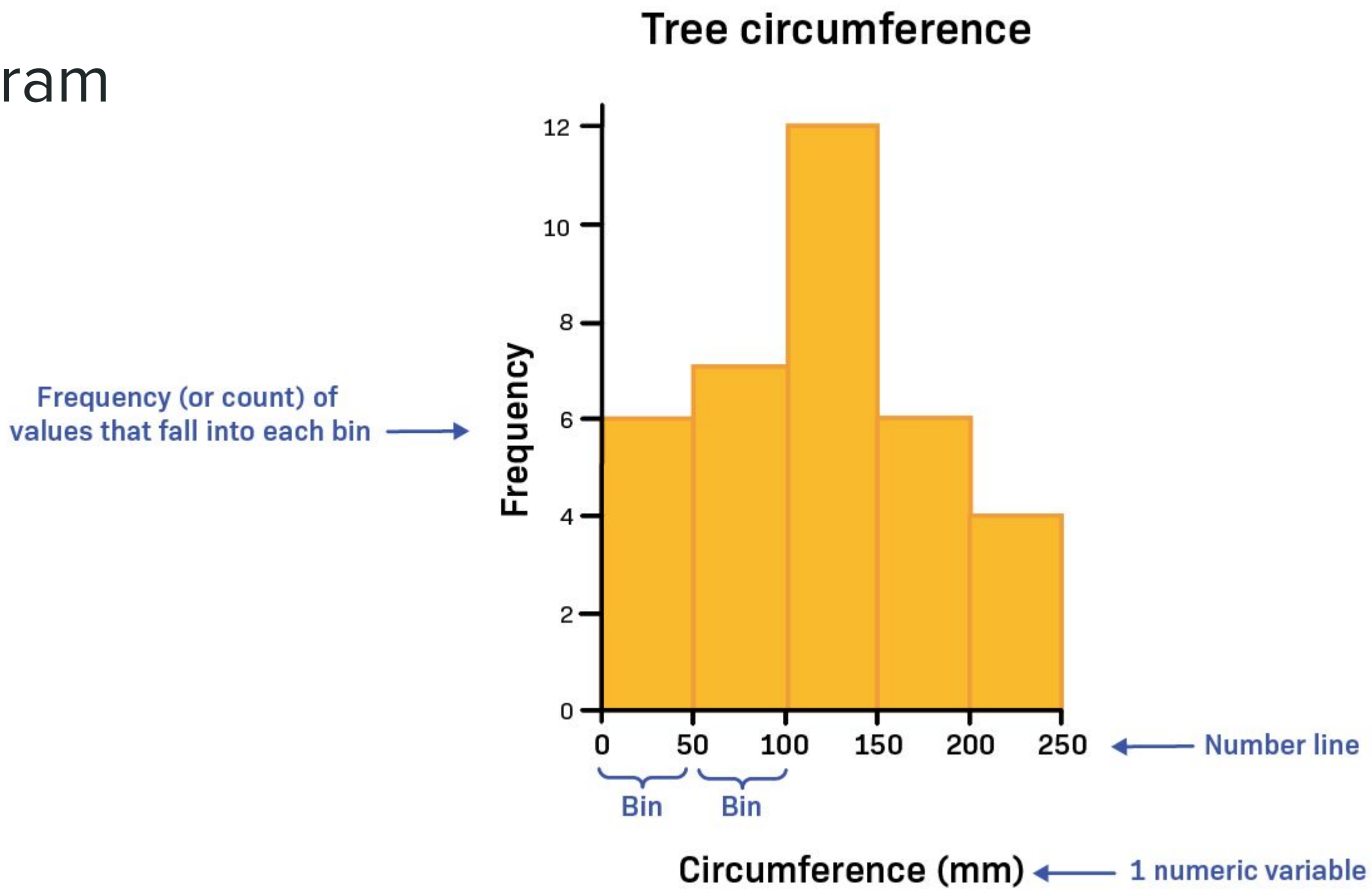
: how values within a column of data are spread or dispersed

- We can learn a lot about the quantity measured by exploring the distribution of its values
- identify patterns, trends, and anomalies

Histogram

- A visual tool to view the data distribution
- Values are binned
- Height of bars shows the count or frequency of values in that bin

Histogram

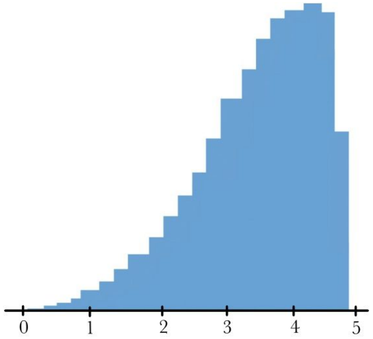


Frequency (or count) of values that fall into each bin

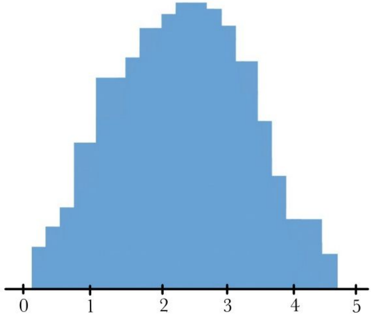
Number line

1 numeric variable

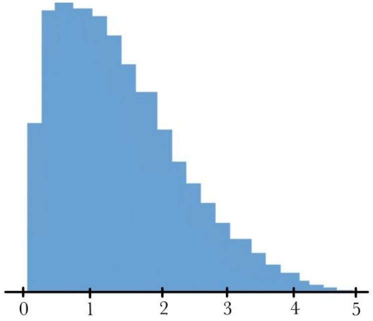
https://www.labxchange.org/library/pathway/lx-pathway



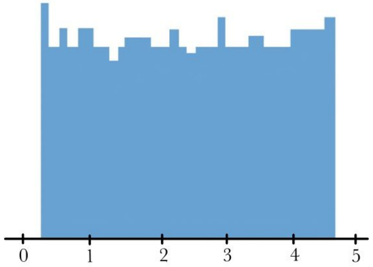
skew left



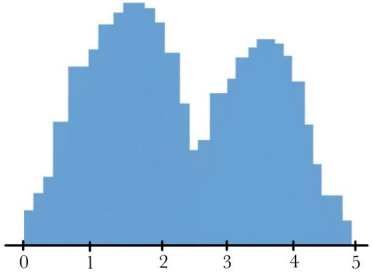
symmetric, unimodal



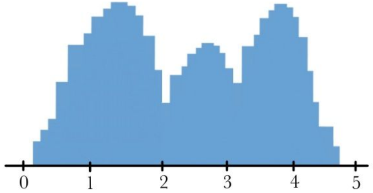
skew right



uniform



bimodal



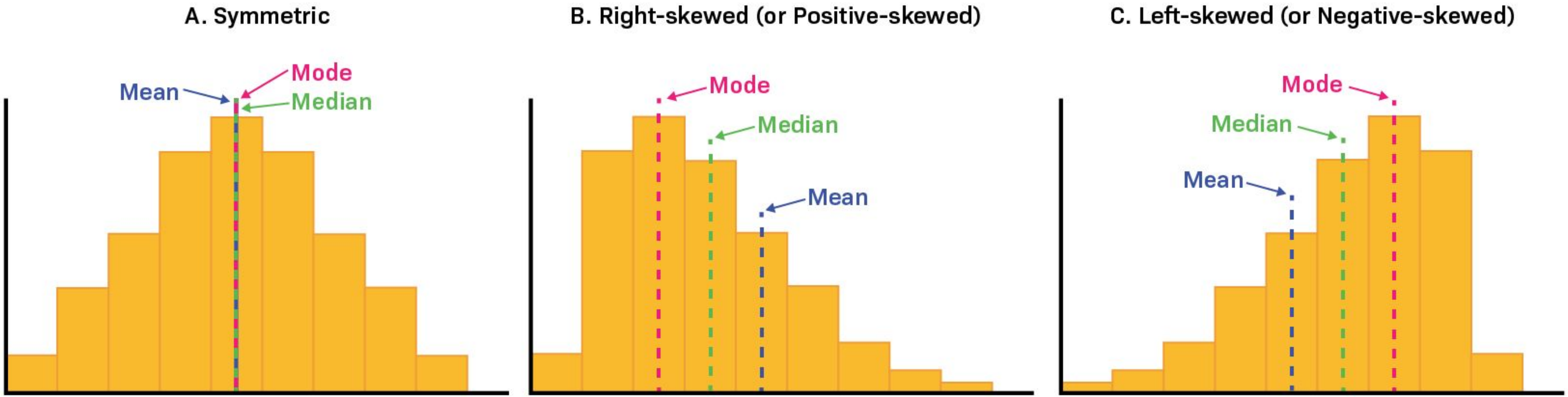
multimodal

https://serc.carleton.edu/details/images/338431.html

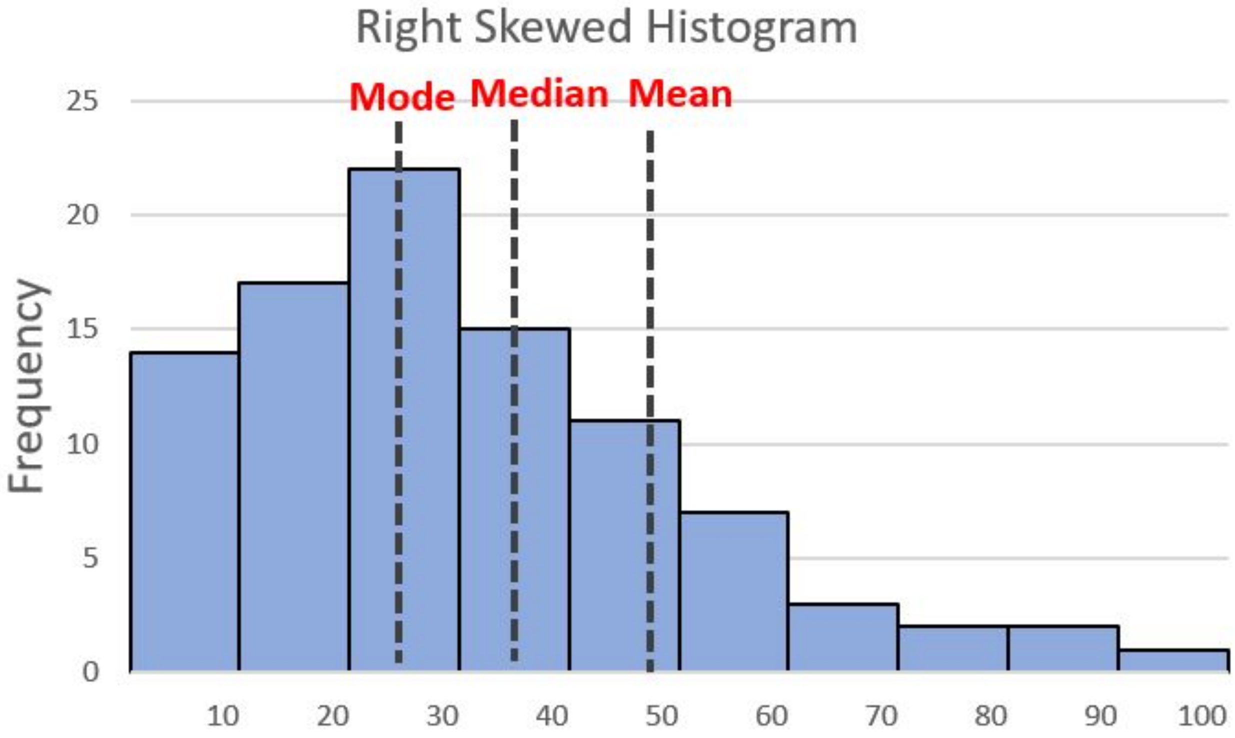
Interpreting/exploring histogram

- Are items in the data mostly very similar, or un-alike/heterogeneous?
- Are values clustered around a single central value, or are there multiple “peaks”/common values in the data?
- Are there gaps?
- Is data symmetric or skewed?
- Are there outliers?

Skew

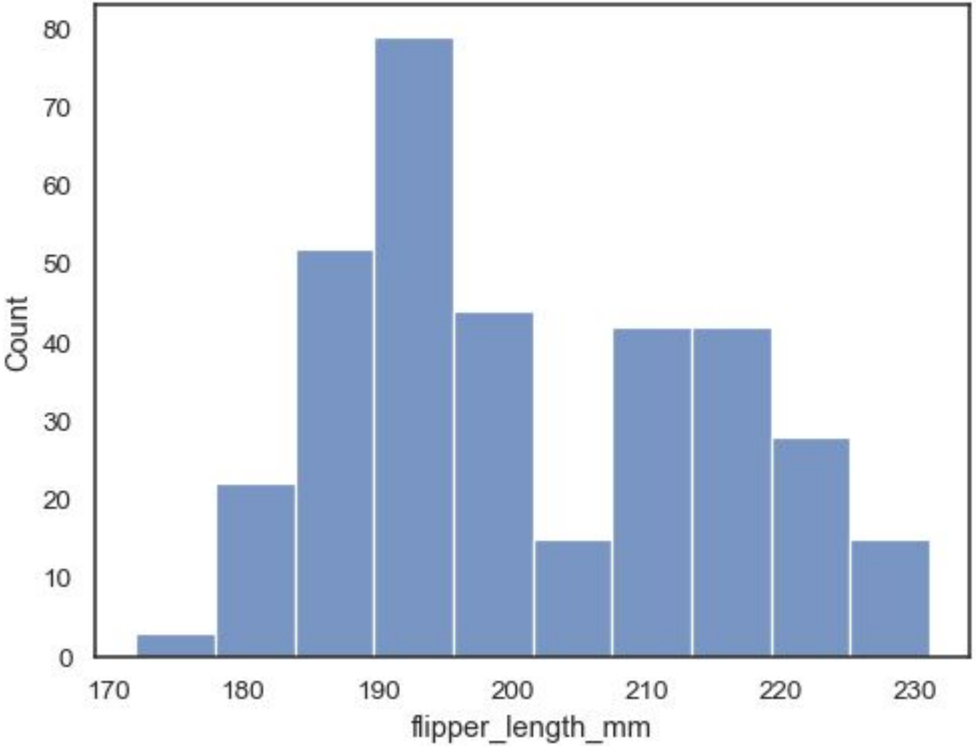


https://www.labxchange.org/library/pathway/lx-pathway

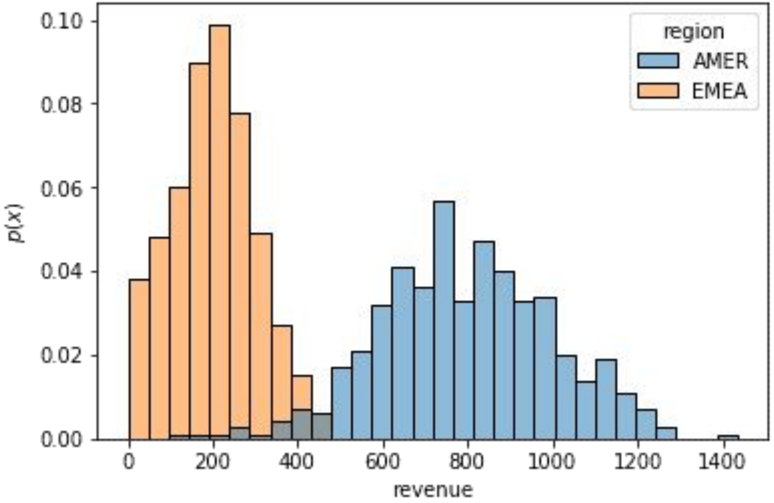


https://www.listendata.com/2023/08/right-skewed-histogram.html

Multiple peaks

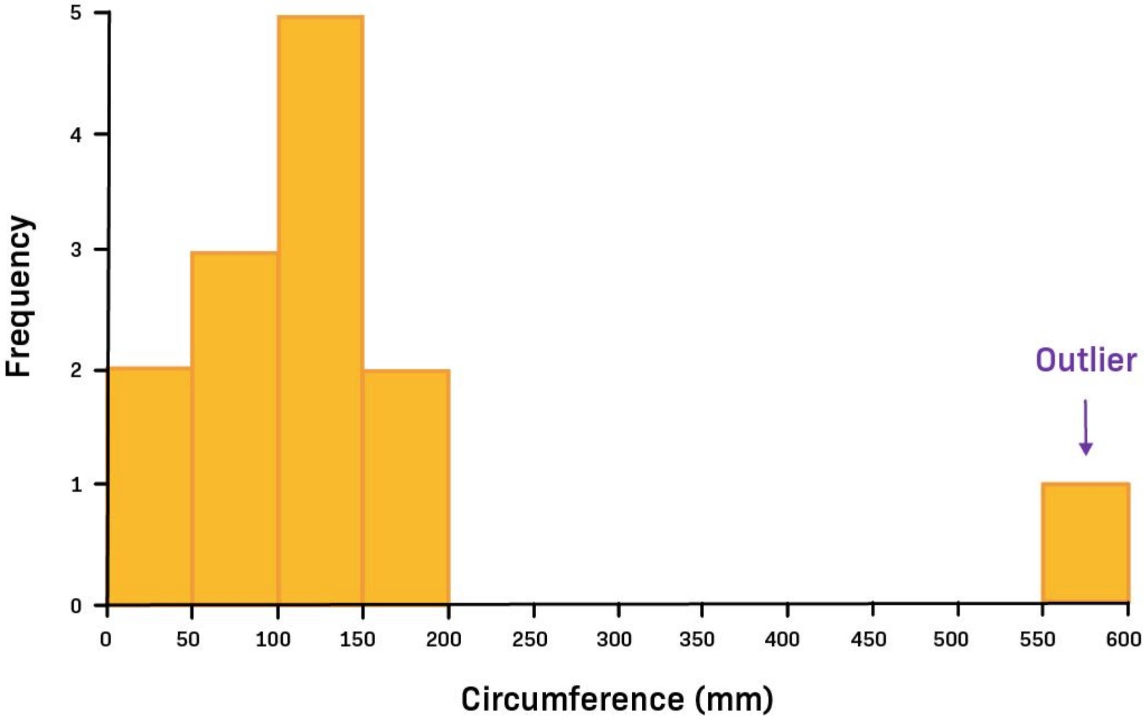


Compare groups



https://sparrow.dev/seaborn-histogram/

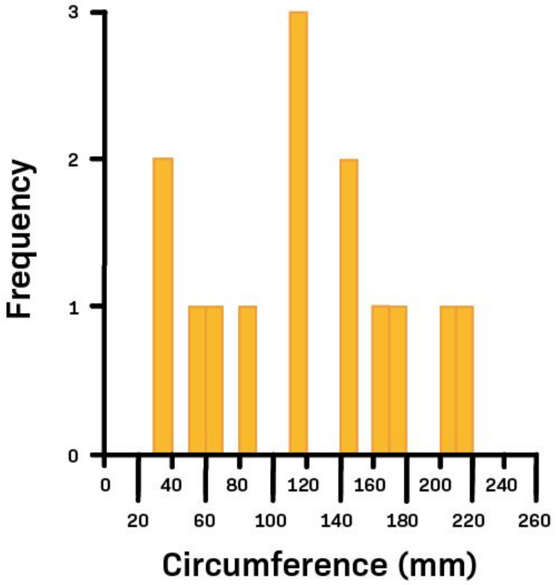
Spot outliers



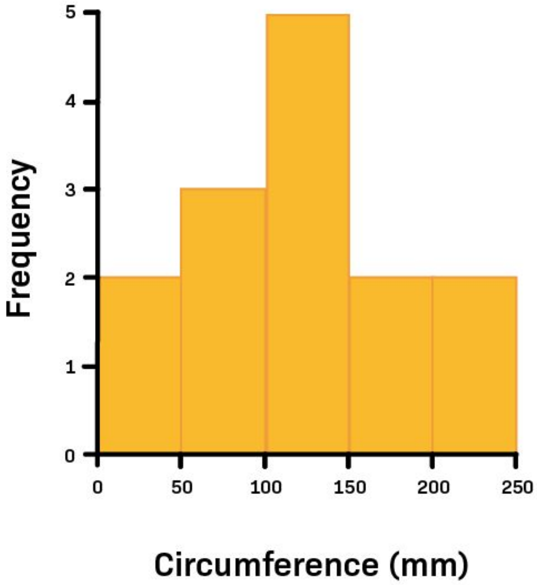
https://www.labxchange.org/library/pathway/lx-pathway

Choose the most helpful bin width

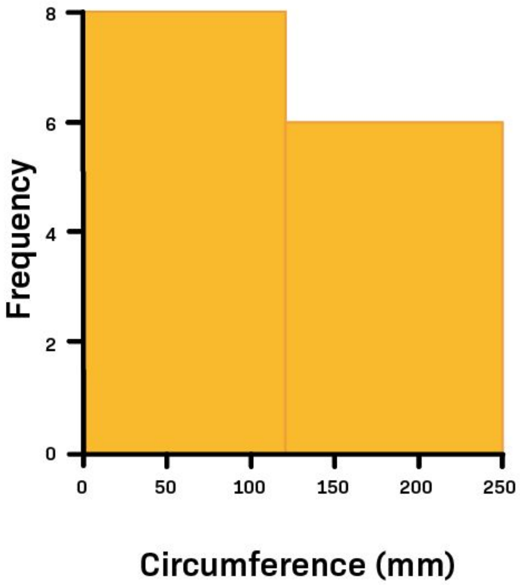
A. Bins Too Narrow



B. Bins Effective



C. Bins Too Wide



https://www.labxchange.org/library/pathway/lx-pathway

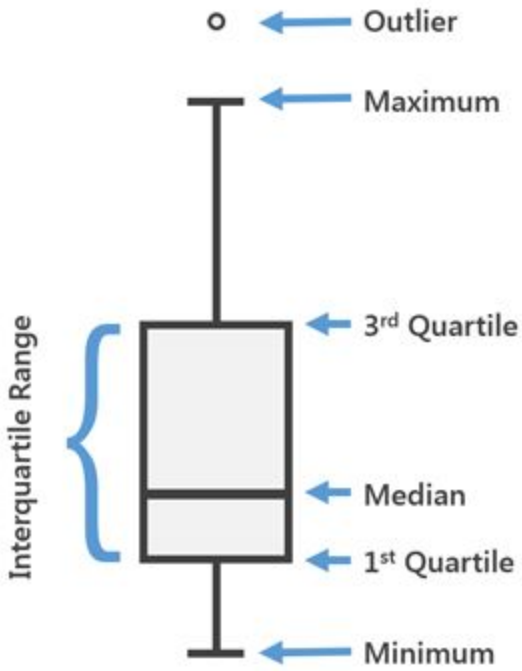
Percentiles

- The percent of a distribution equal or below the value
- median = 50th percentile = half of values are above, half are below
- e.g. 75th percentile: 75% of values are less than or equal to, 25% of values are greater than

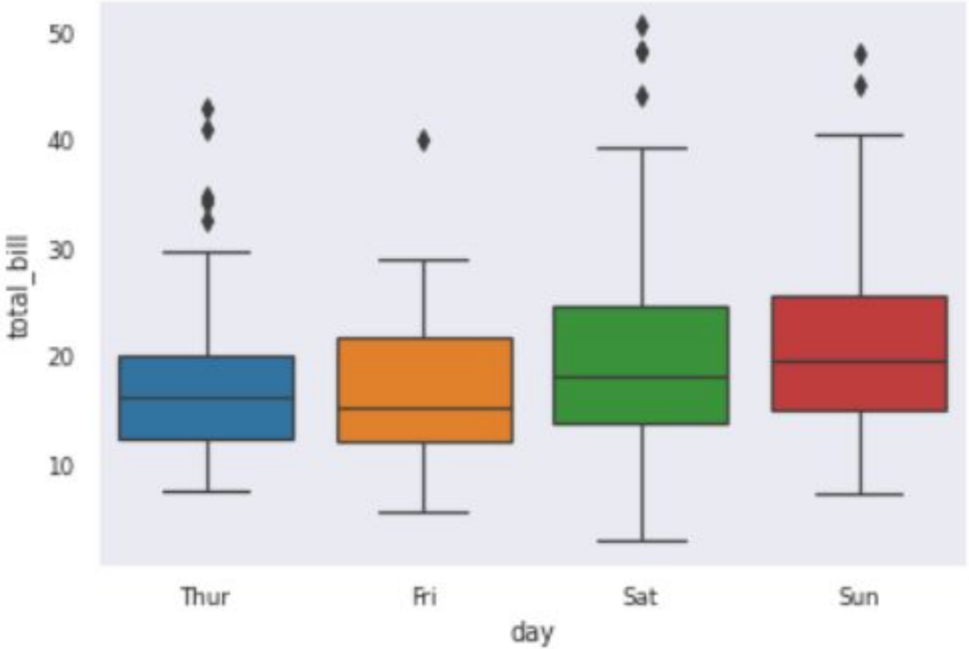
Interquartile range

- *Quartile*
  - divide date into quarters
  - i.e. 25th, 50th. 75th percentiles
- *Interquartile range*
  - middle half of values
  - i.e. between 25th and 75th percentiles

Box plot


Outlier
Maximum
3rd Quartile
Median
1st Quartile
Minimum
Interquartile Range

Box plots



Interpreting box plots

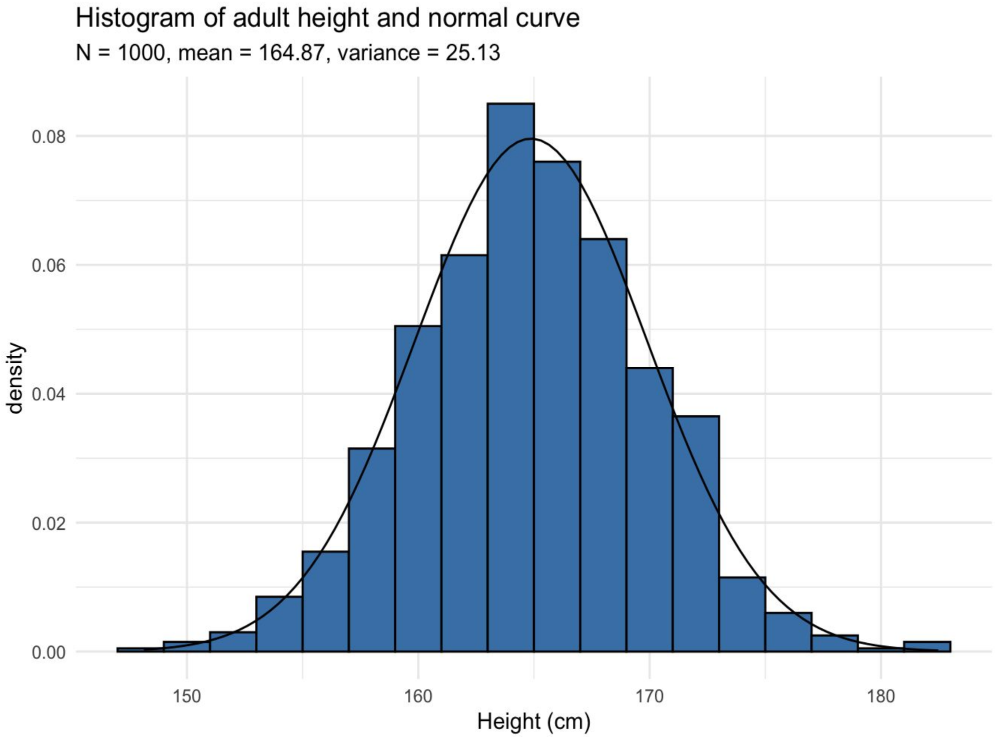
- How wide is the IQR? how concentrated are the values?
- Is the median centered? are there more values above or below?
- Are there outliers?

Why visualize distribution?

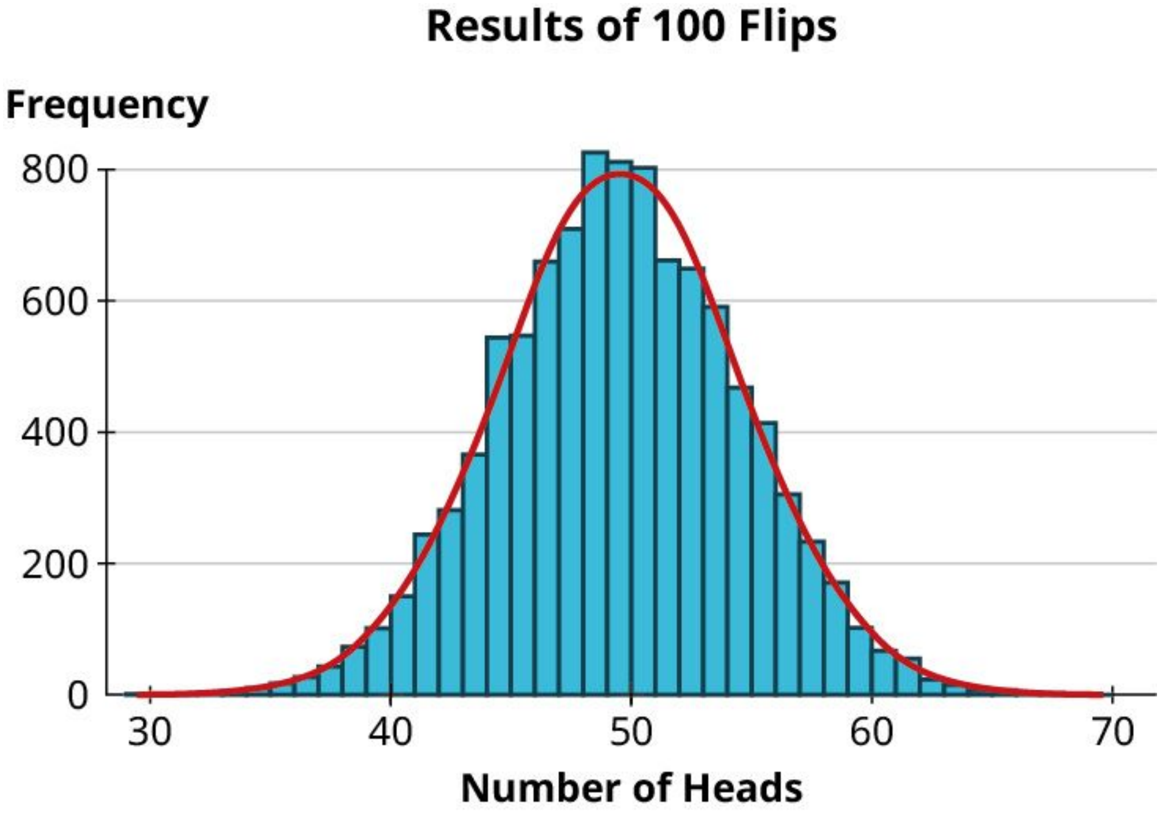
- Check for outlier values (and potentially invalid values)
- Check for skew or symmetry
- Check for dispersion

What is normal?

Normal distribution



Normal distribution



Variance

A measure of dispersion of the data

: average square difference from the mean

- mathematically useful but not intrinsically interpretable

Standard deviation

: square root of variance

- so at the *same scale* as the data values and the mean

Standard deviation

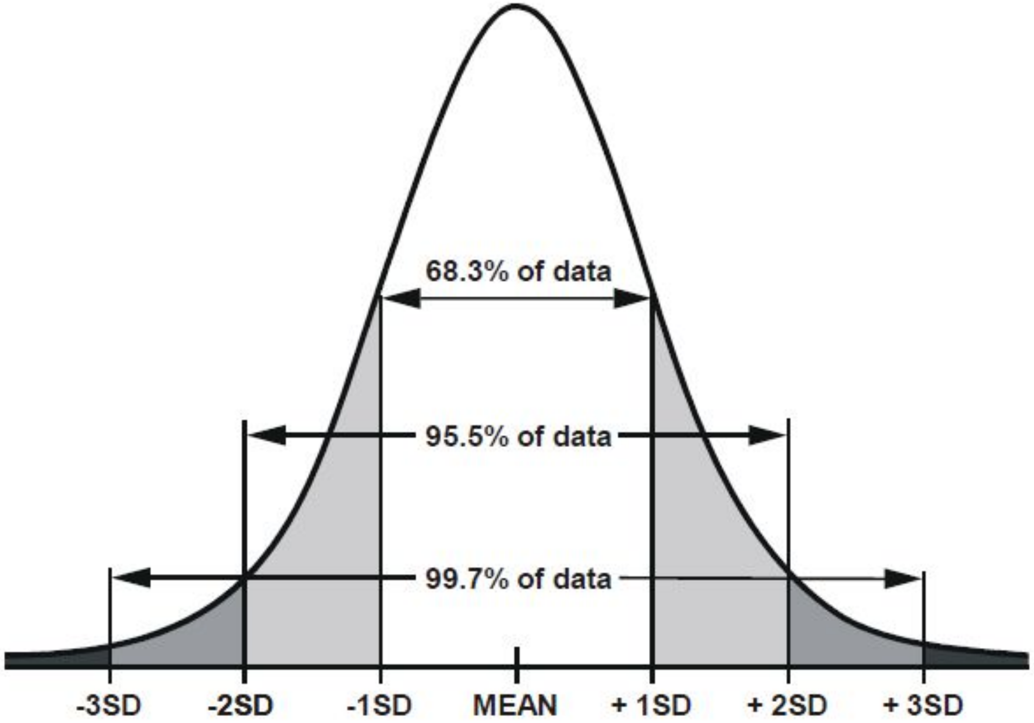
measure of amount of dispersion of values around the mean

- low standard deviation : values clustered near the mean
- high standard deviation : values spread far from the mean

Empirical rule

Figure 3.9

Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean


68.3% of data
95.5% of data
99.7% of data
-3SD
-2SD
-1SD
MEAN
+ 1SD
+ 2SD
+ 3SD