M2S2 - Distributions

Professor Jarad Niemi

STAT 226 - Iowa State University

August 29, 2018

Outline

- Population
  - Location
  - Spread
  - Modality: unimodal, bimodal
  - Skewness: symmetric, right-skewed, left-skewed
- Sample
  - Boxplot
  - Histogram
  - Summary statistics
- Outliers

Population

Definition

The population is the entire group of individuals that we want to say something about.

Definition

Individuals are the subjects/objects of interest.

Definition

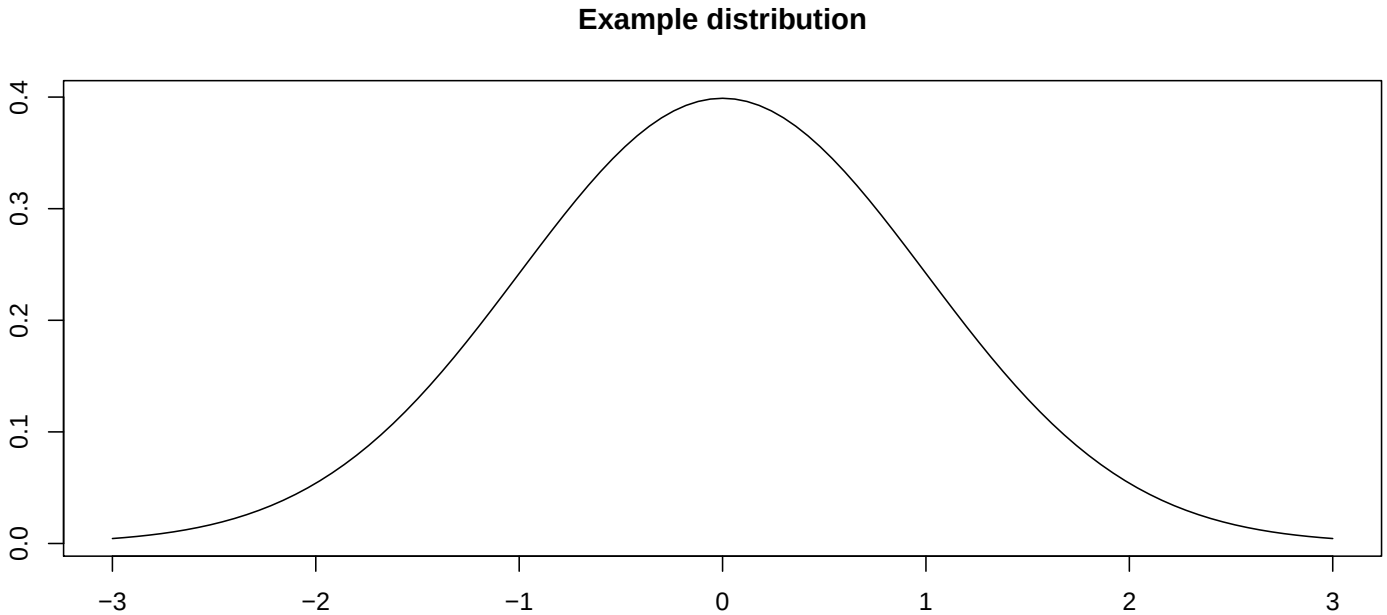
A variable is any characteristic of an individual that we are interested in.

Distribution

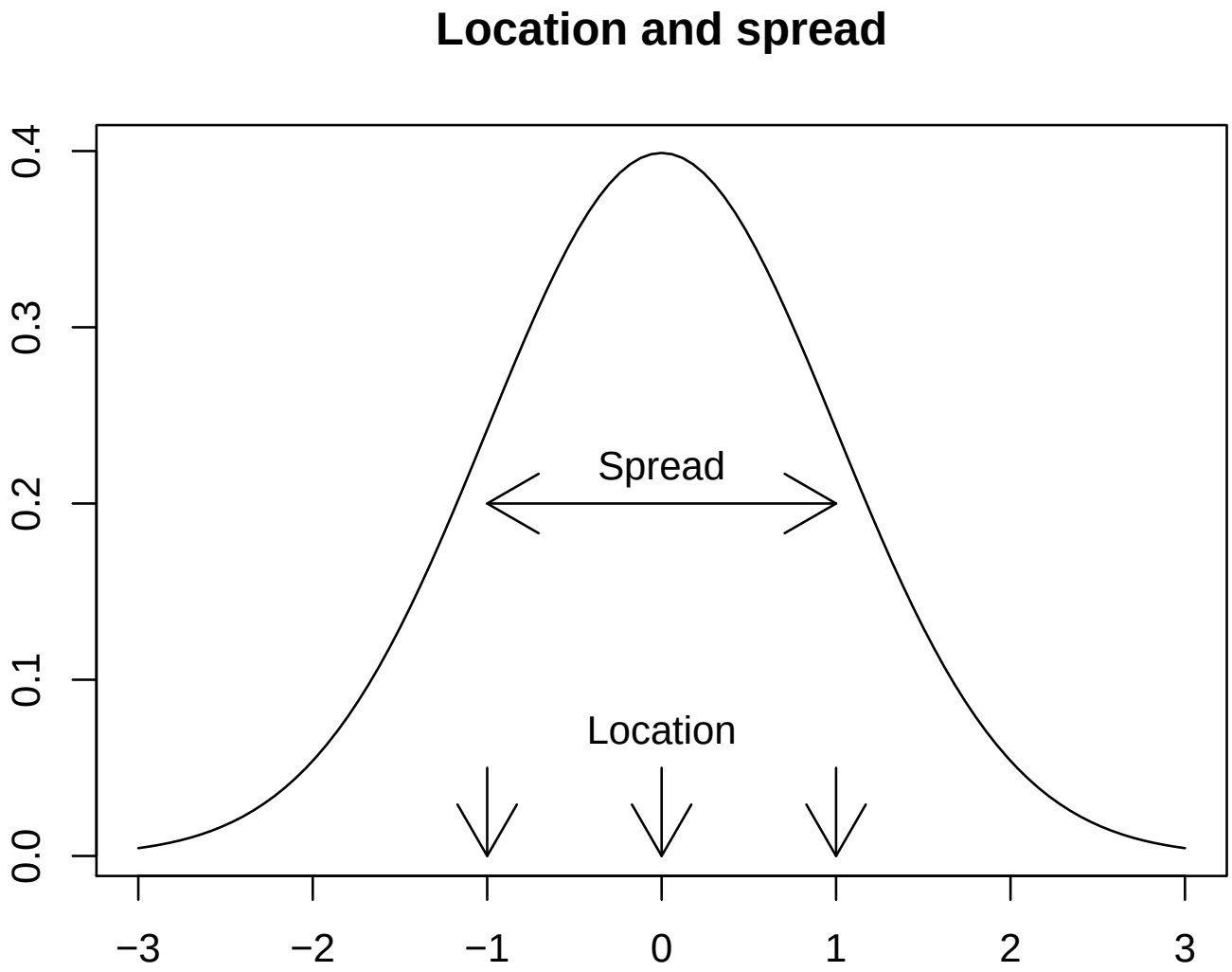
Definition

The distribution of a variable is the collection of possible values the variable can take and how often each value occurs **in the population**.

Enumerating the values may be possible for categorical variables, but typically will not work for numerical variables. Instead we depict the distribution graphically, e.g.



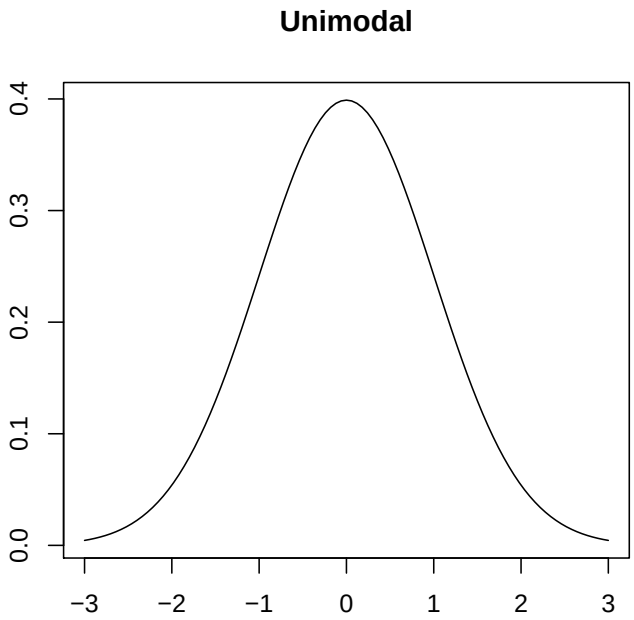
Distribution location and spread

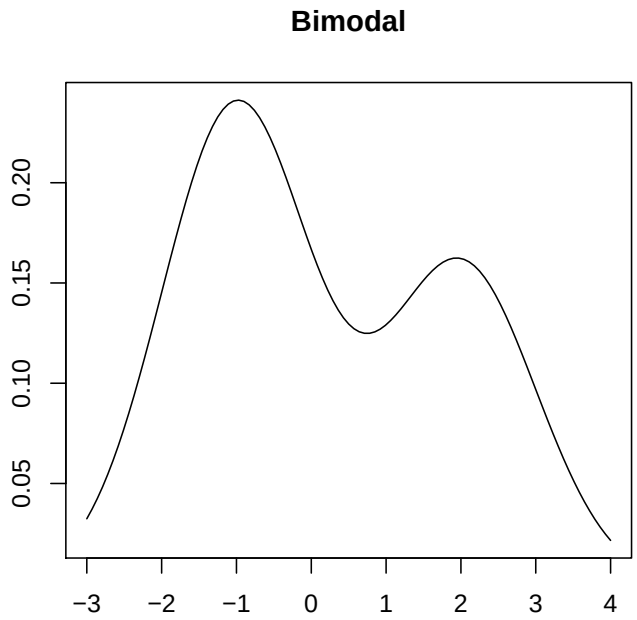


Modality

Definition

A unimodal distribution has one peak. A bimodal distribution has two peaks.

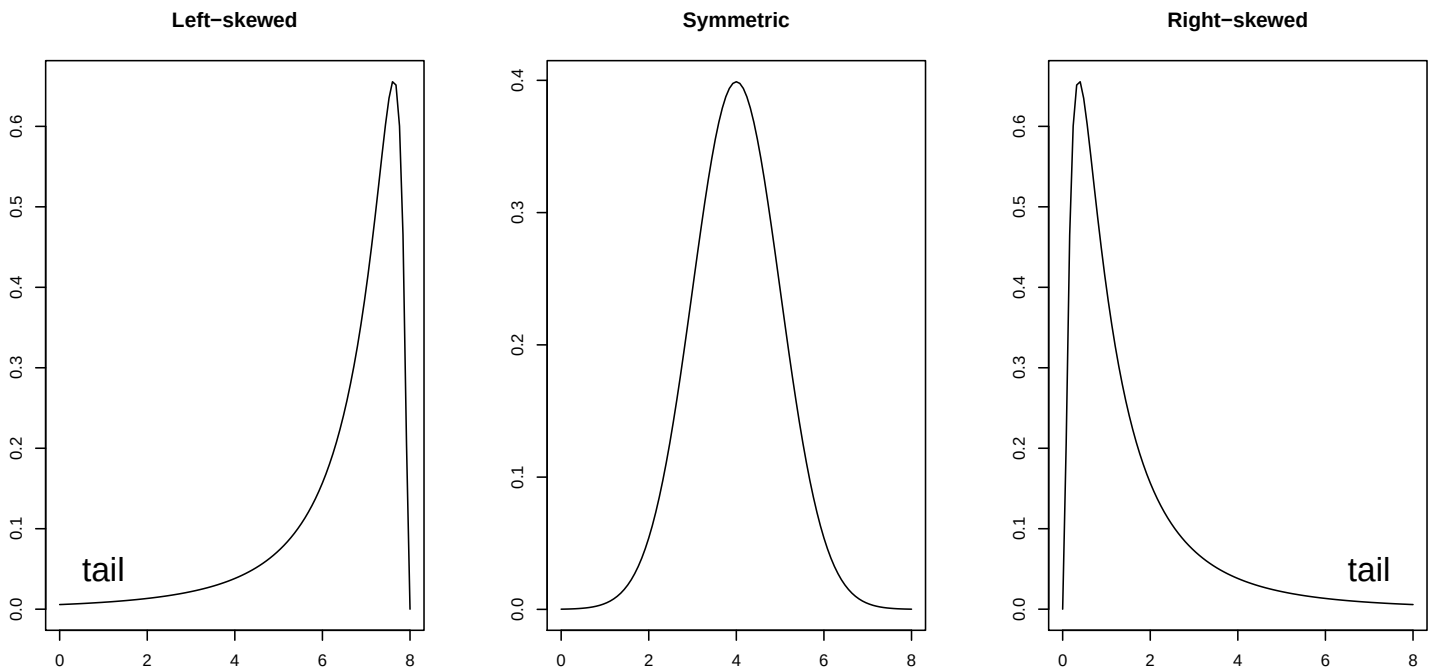




Skewness

Definition

A distribution is symmetric if there is some vertical line where the graph is a mirror reflection. A distribution is right skewed if the tail of the distribution is longer to the right. A distribution is left skewed if the tail of the distribution is longer to the left.



Sample

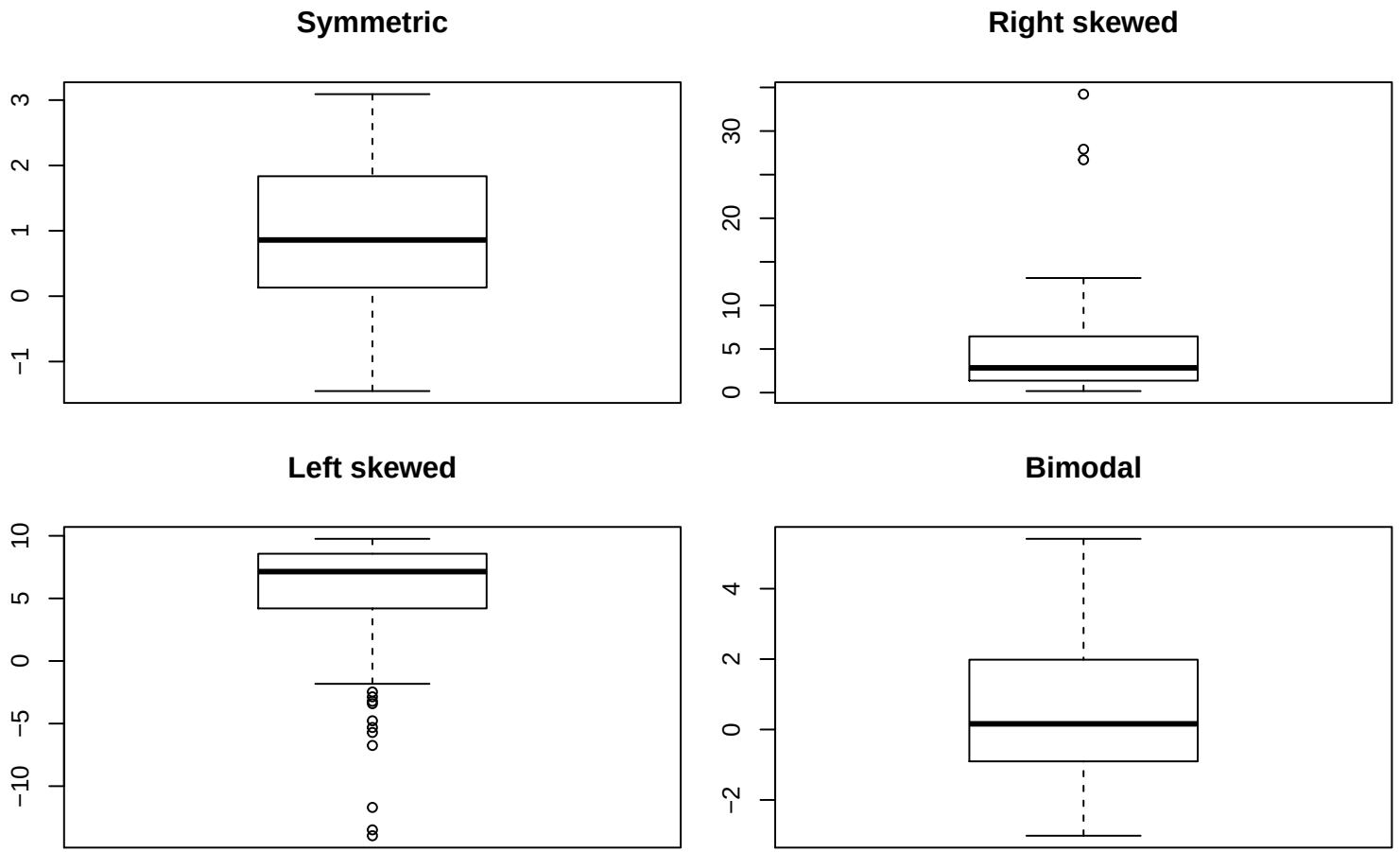
We never see the population!

Thus we often try to infer details about the population from our sample. We use our sample to infer the distribution's

- location,
- spread,
- modality, and
- skewness.

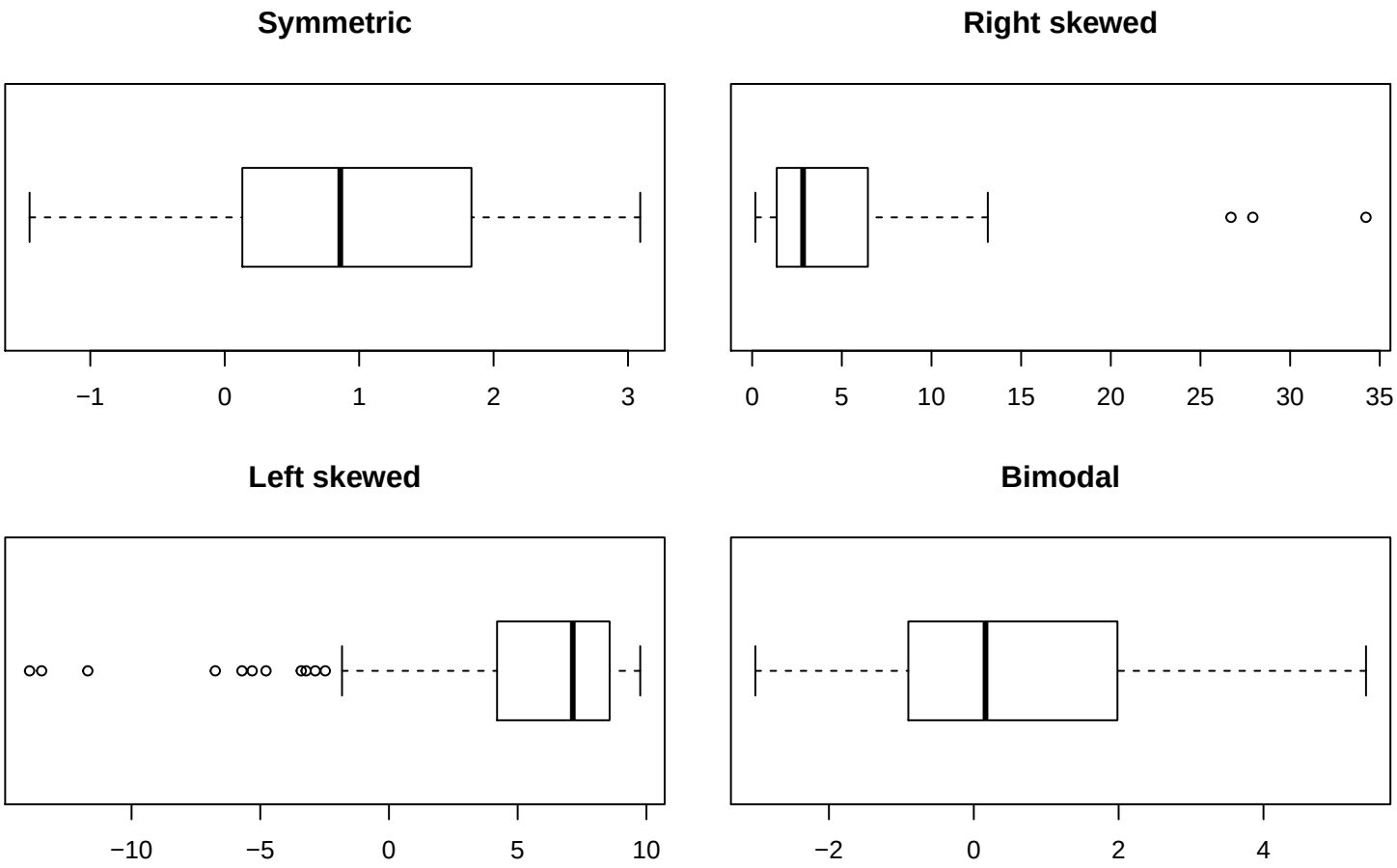
Vertical Boxplots

A boxplot can be used to help infer location, spread, and skewness, e.g.



Horizontal Boxplots

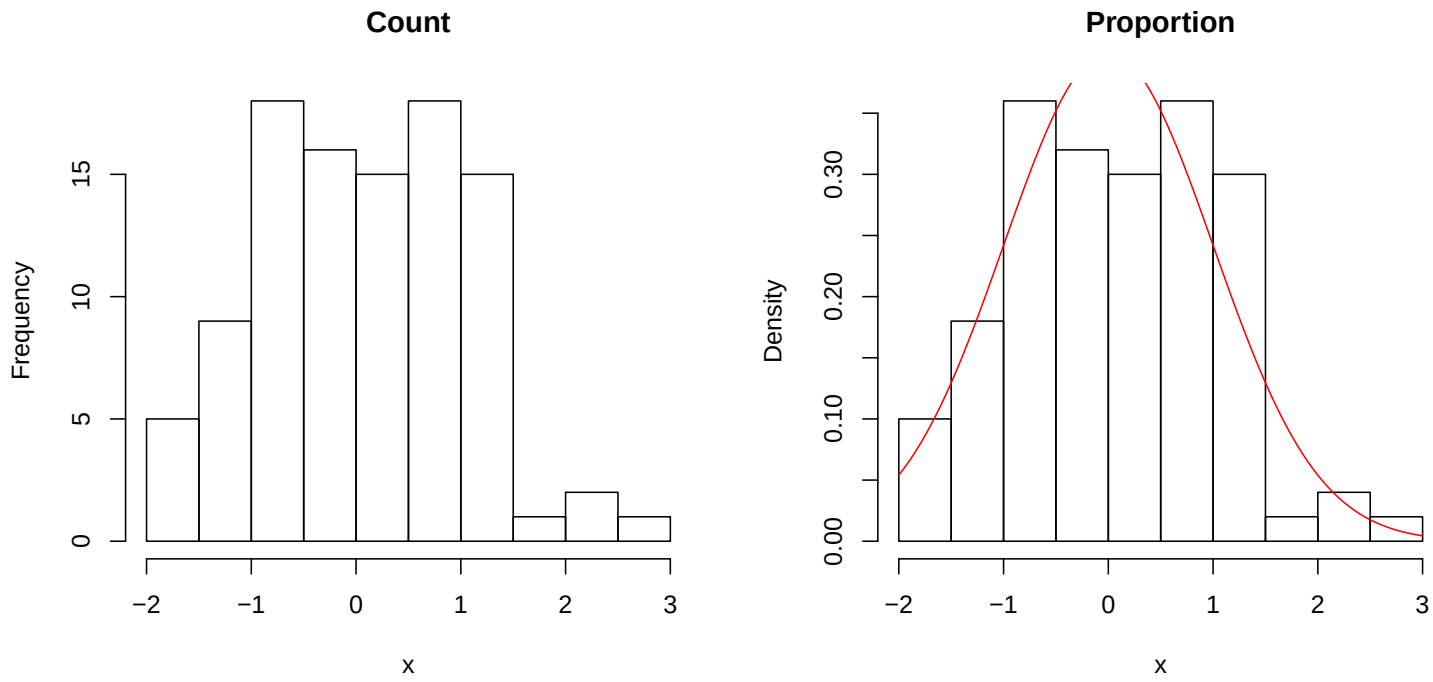
A boxplot can be used to help infer location, spread, and skewness, e.g.



Histogram

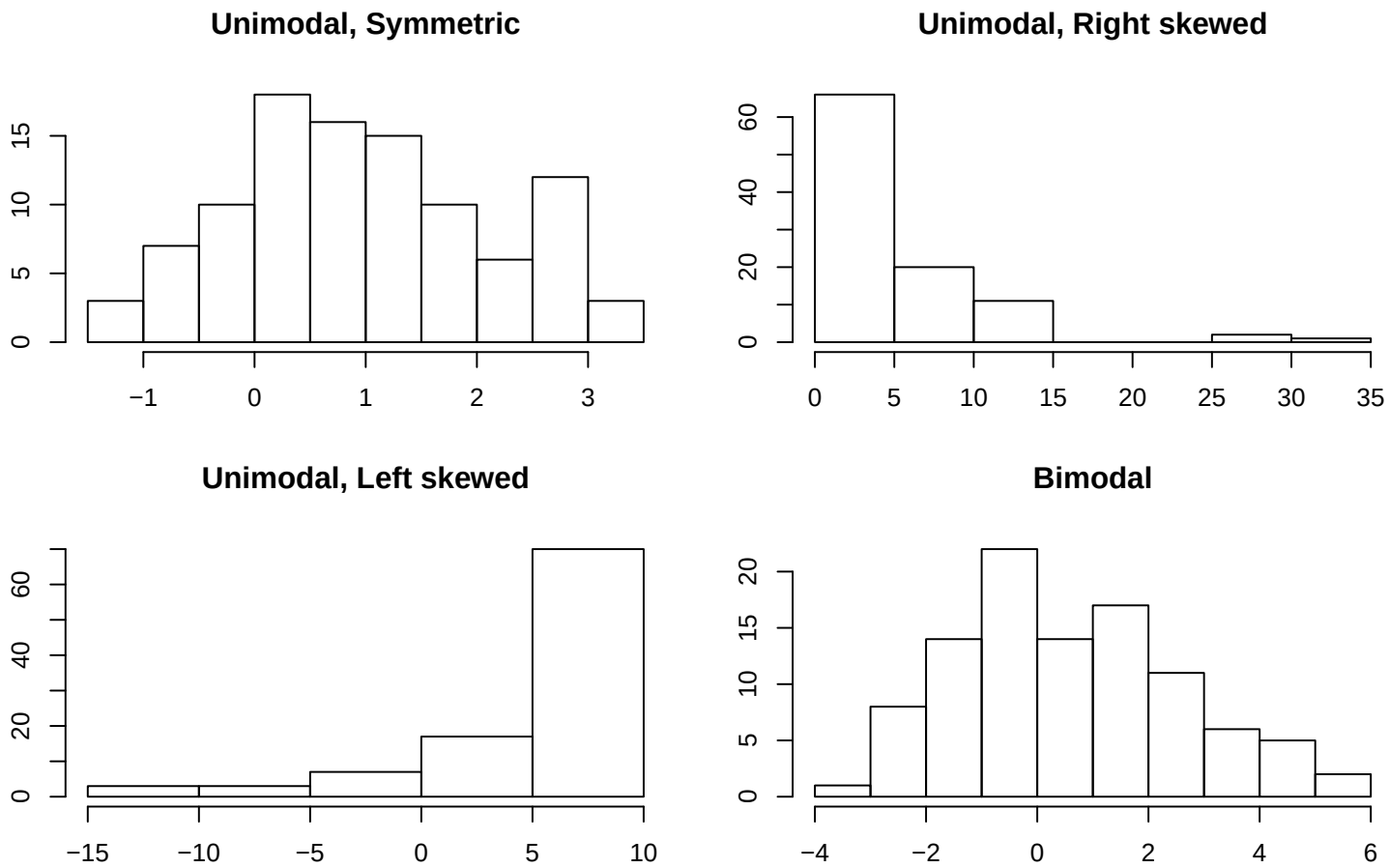
Definition

A histogram is a graphical display of numerical data that counts the number of observations in each bin where the bins are determined by the user.



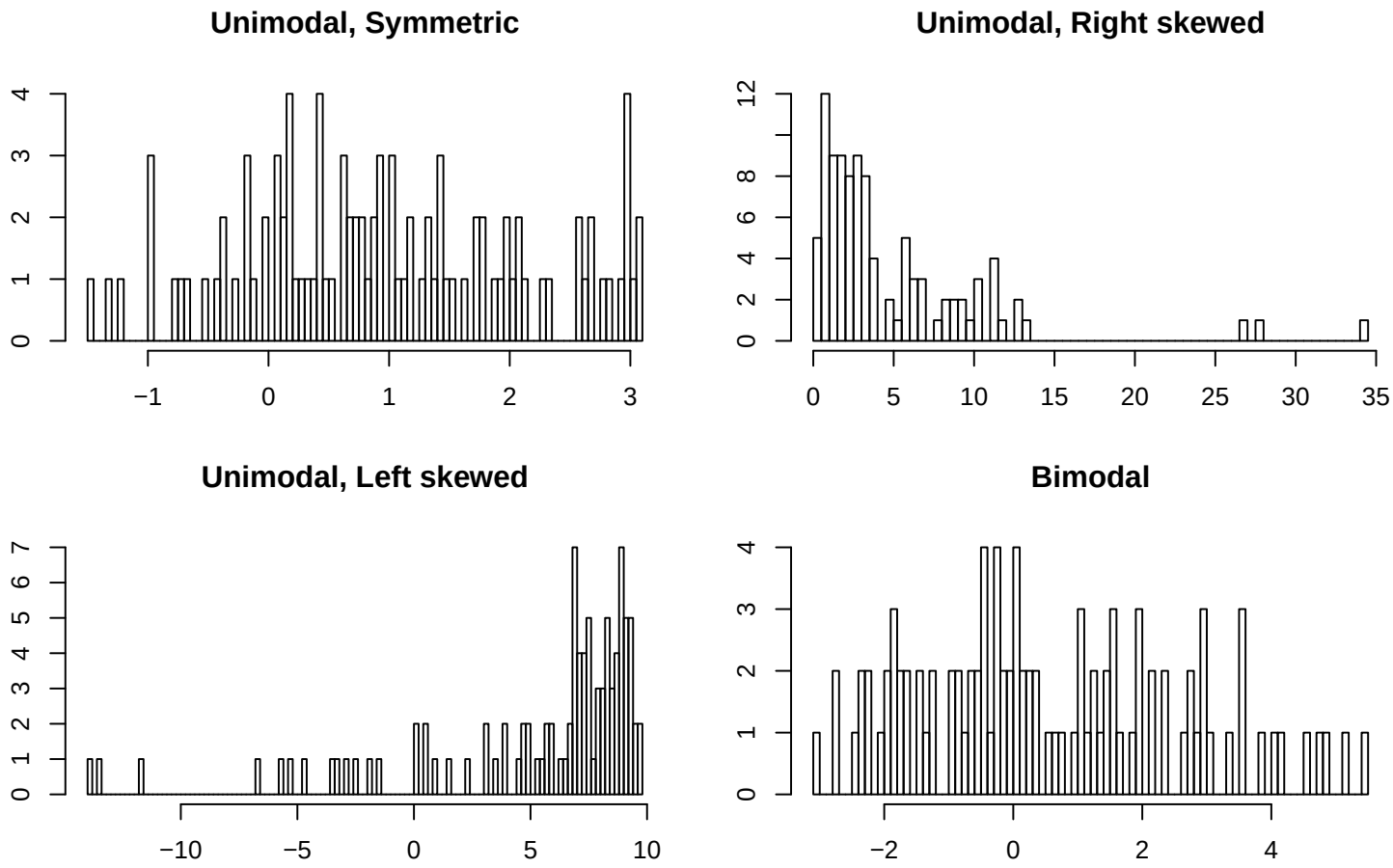
Histograms

A histogram can be used to help infer location, spread, skewness, and modality, e.g.



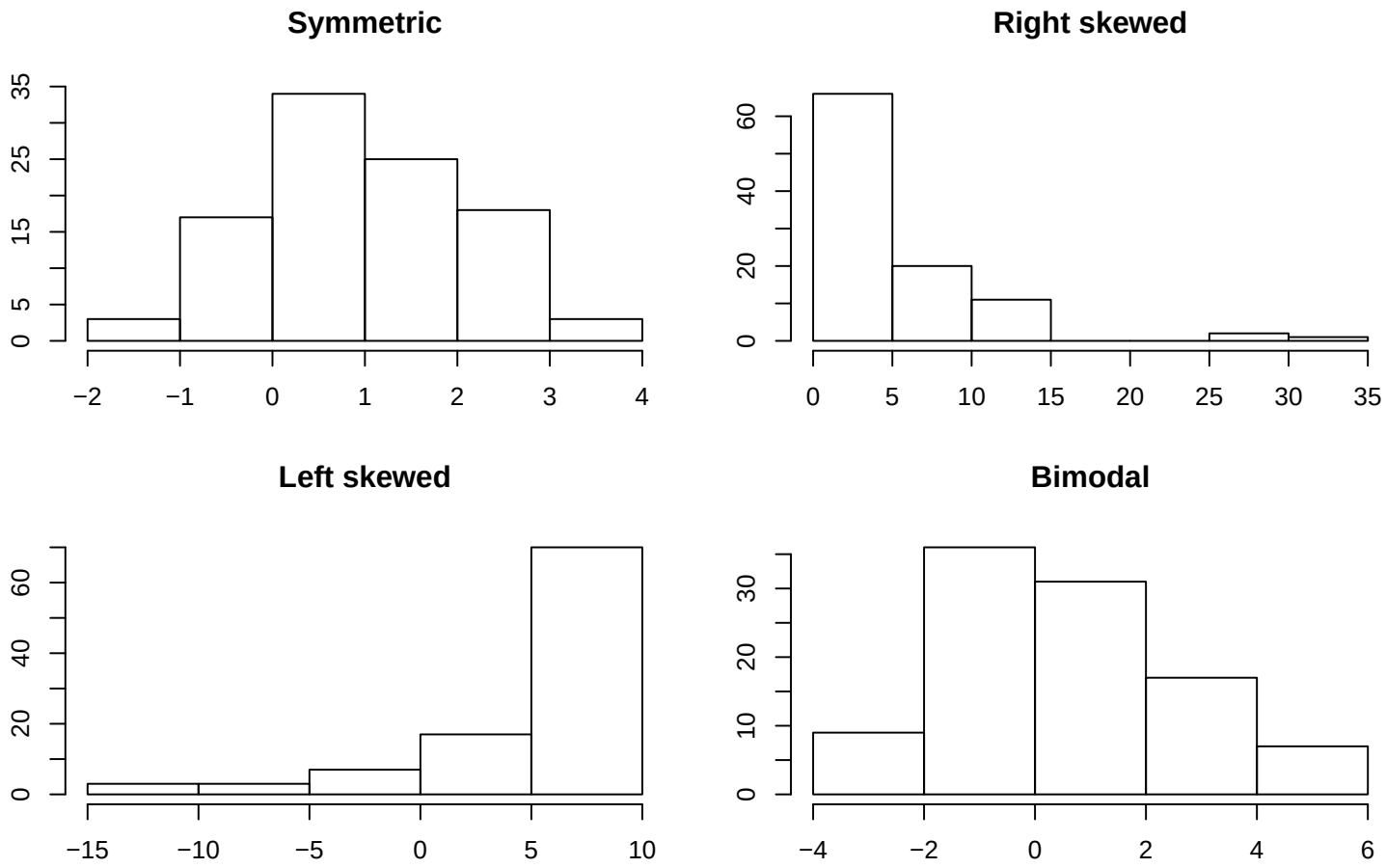
Histograms

Histograms are affected by the choice of bins



Histograms

Histograms are affected by the choice of bins



Measures of location

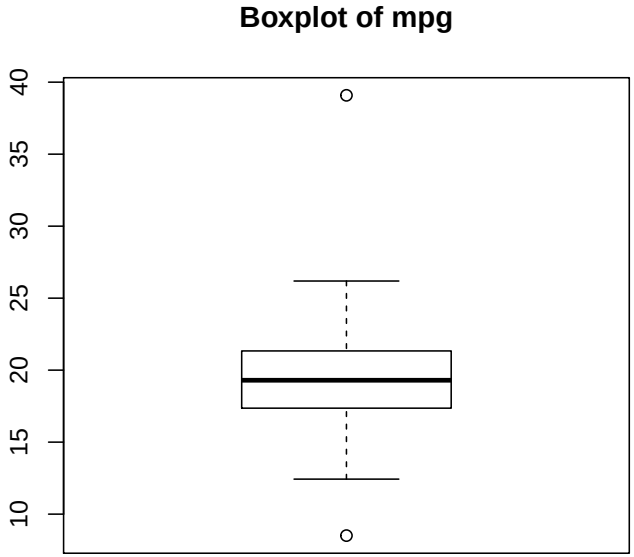
Distribution	min	Q1	median	mean	Q3	max
bimodal	-3.02	-0.90	0.16	0.57	-0.90	5.42
left_skew	-13.96	4.36	7.14	5.24	4.36	9.76
right_skew	0.18	1.39	2.84	4.89	1.39	34.23
symmetric	-1.45	0.14	0.86	0.97	0.14	3.09

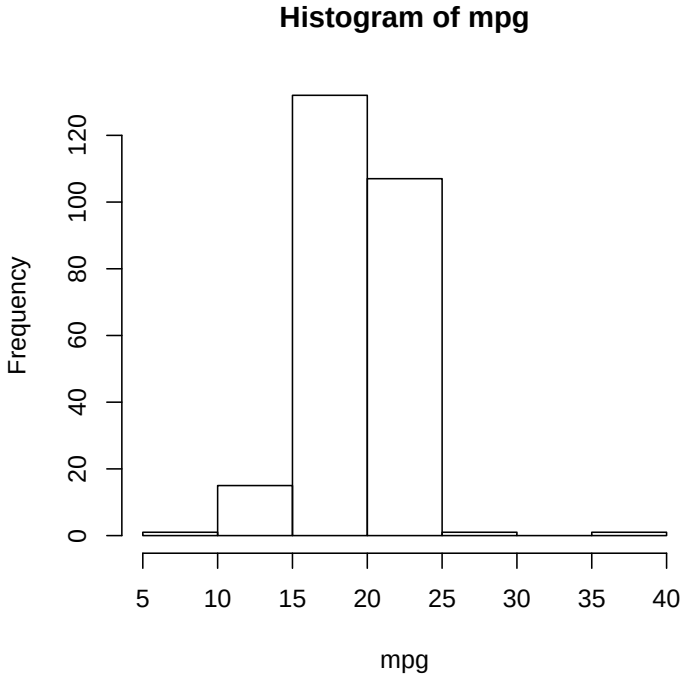
- Right-skew: mean $>$ median
- Left-skew: mean $<$ median
- Symmetric: mean $\approx$ median

Measures of spread

Distribution	variance	standard_deviation	range	interquartile_range
bimodal	4.20	2.05	8.43	2.88
left_skew	26.25	5.12	23.72	4.19
right_skew	31.57	5.62	34.05	5.04
symmetric	1.35	1.16	4.54	1.67

Toyota Sienna Miles per Gallon





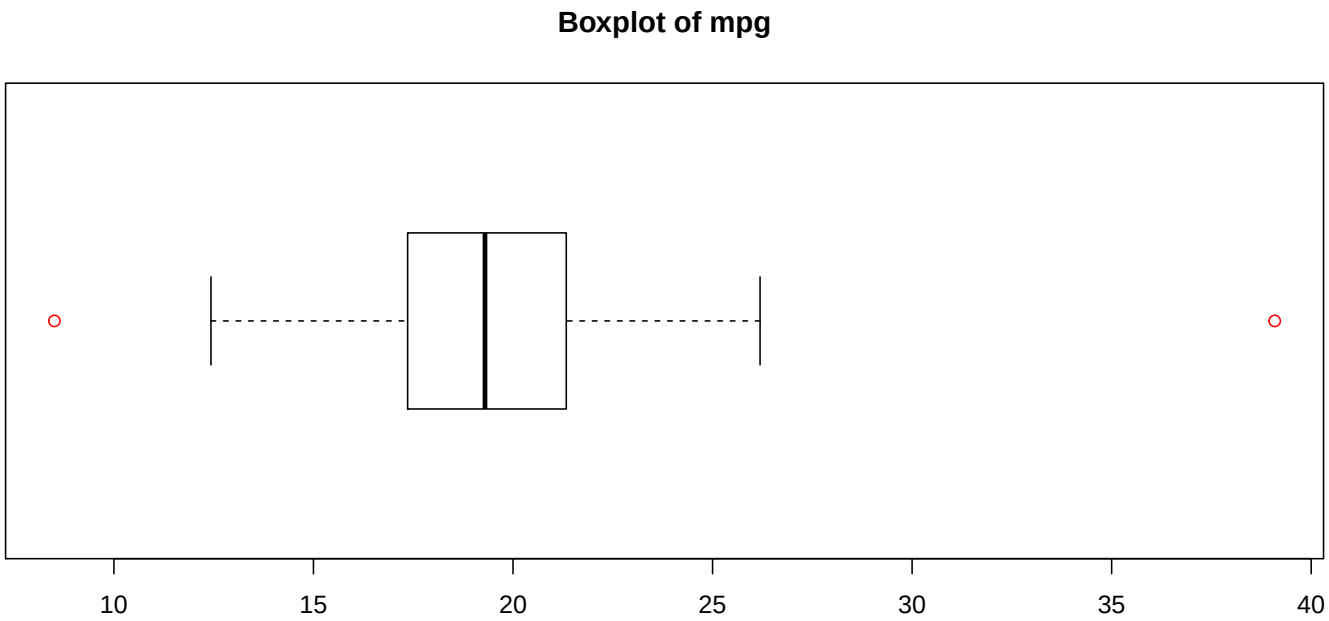

```
summary(dd$mpg)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.509  17.359  19.298  19.313  21.334  39.086
```

Outliers

Definition

An outlier is an observation that is distant from other observations. Sometimes, any observation below Q1-1.5×IQR or above Q3+1.5×IQR is called an outlier.



Summary statistic choice

Choice of an appropriate measure of location/spread depends on

- shape of the distribution
- presence of outliers.

Generally,

- symmetric with no outliers $\Longrightarrow$ mean and standard deviation
- skewed and/or outliers $\Longrightarrow$ median, IQR, 5-number summary