Data displays and summary statistics

Module 2

Statistics 251: Statistical Methods

Updated 2022

Have data will, er. . . ?

Now what to do with the data?

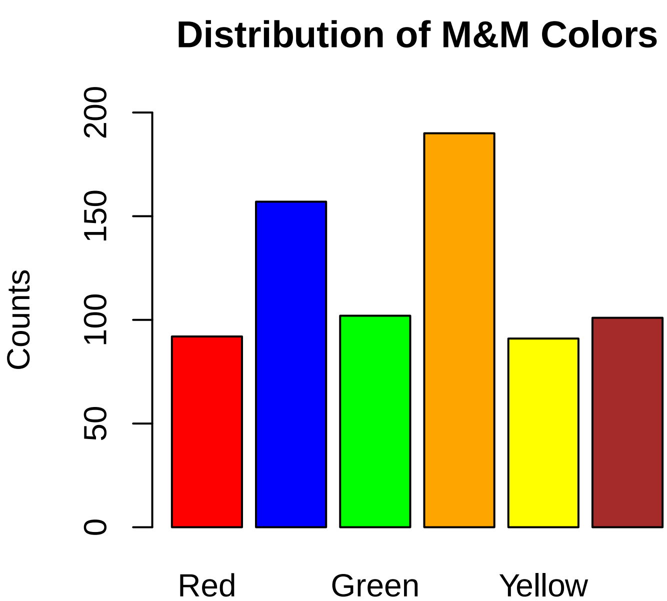
- Graphs visualize what the data is telling us
- see trends and other features
- knowing what the data looks like in graph form tells us what analysis to use

Qualitative Graphs

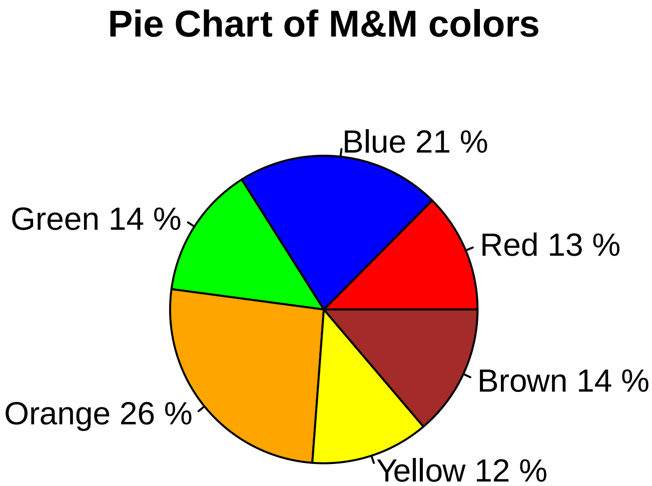
Bar graph consist of bars that are separated from each other and are usually rectangular; bars represent the categories

Pie graph shows parts of a whole, in pie form. Not very useful and easy to manipulate; not used in this class beyond an example

Barplot



Pie chart

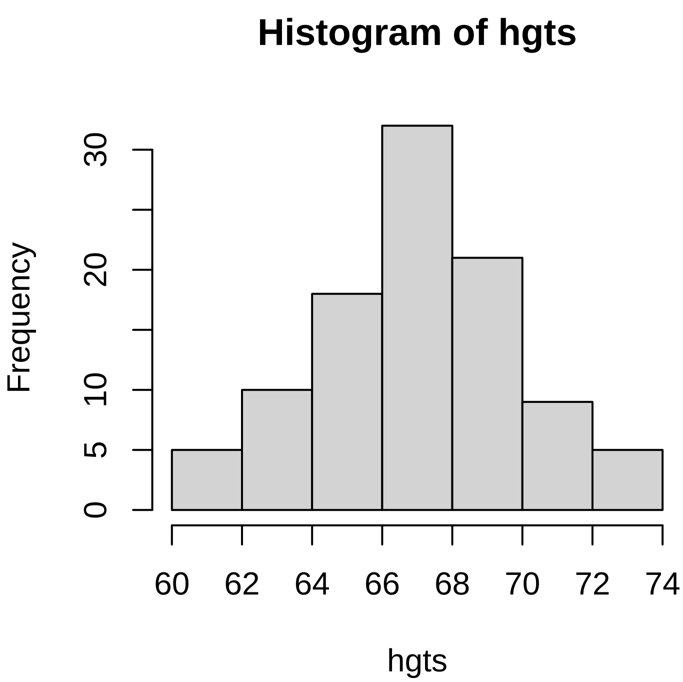


Quantitative graphs I

histogram consists of adjoining rectangles. The horizontal axis (x) is the ranges of data values and the vertical axis (y) is the frequency (counts) or relative frequency (percents or probability, denoted as rf).

stemplot (stem-and-leaf plot) comes from the field of exploratory data analysis and is good when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit.

Histogram



Stemplot

```
  The decimal point is 1 digit(s) to the right of the |
```

```
0 | 02
0 | 68
1 | 00013
1 | 557889
2 | 0011233344
```

Quantitative graphs II

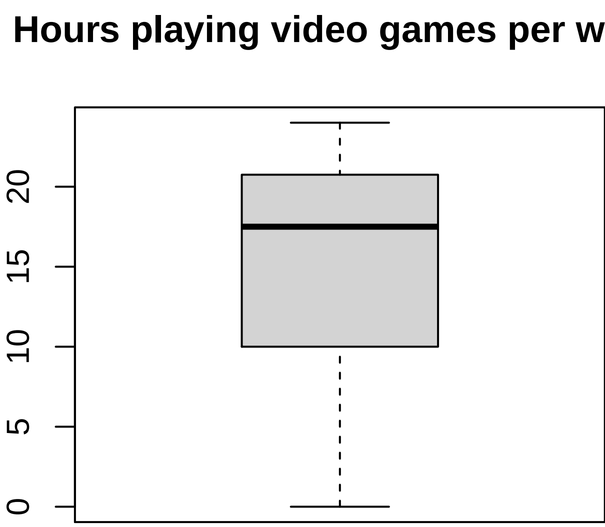
boxplot also called box-and-whisker plots or box-whisker plots. Gives a good graphical image of the concentration of the data and also show how far the extreme values are from most of the data. A boxplot is constructed from five values: the minimum value (min), the first quartile ($Q1$), the median ($Median$), the third quartile ($Q3$), and the maximum value (max); the collection of these five summaries are called the "5 Number Summary".

scatterplot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data, an x and y. Is useful for identifying trends/associations between two variables.

time series also called line plot. Shows the distribution of a varaible over a specified time period.

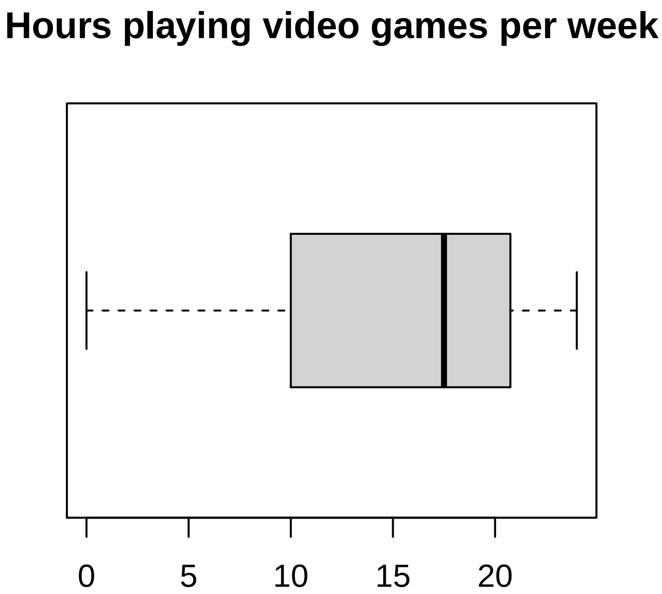
Boxplot

Vertical boxplot

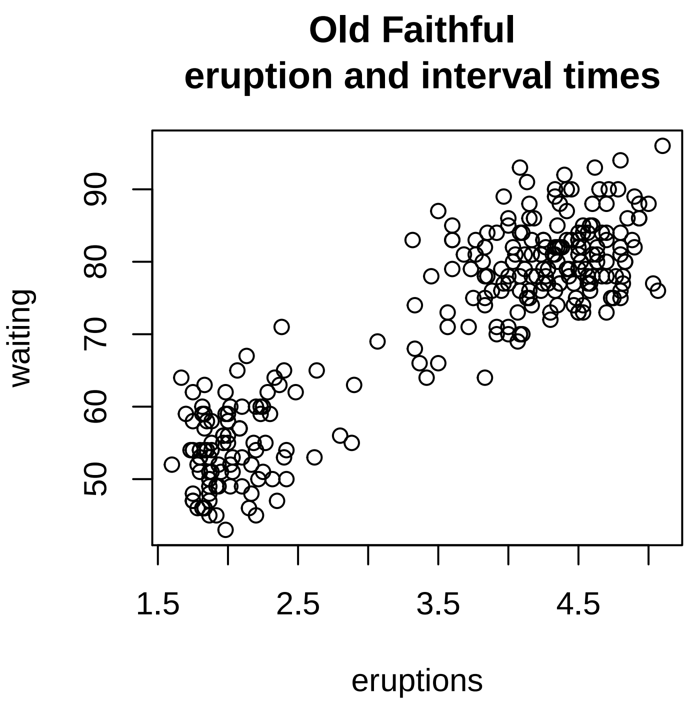


Boxplot

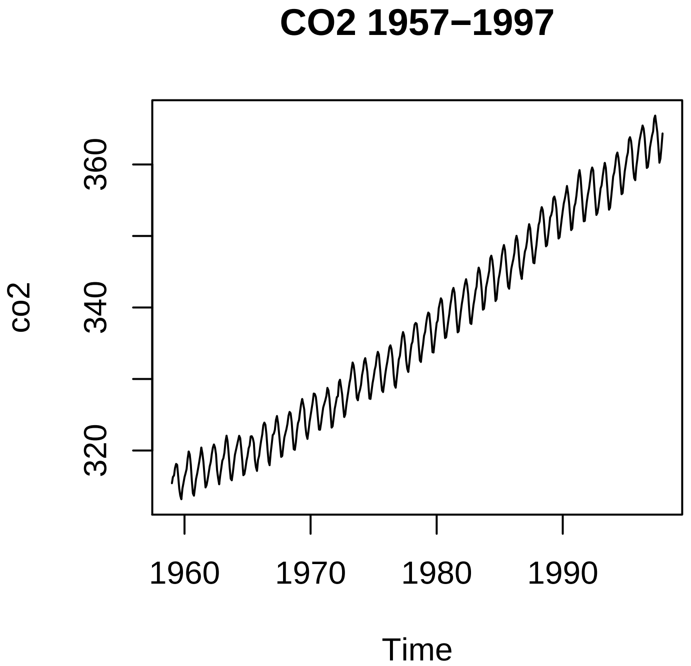
Horizontal boxplot



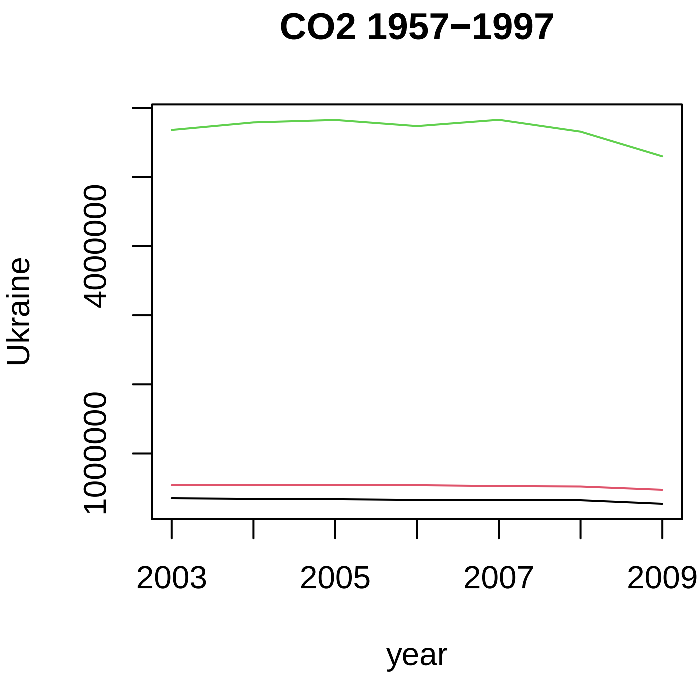
Scatterplot



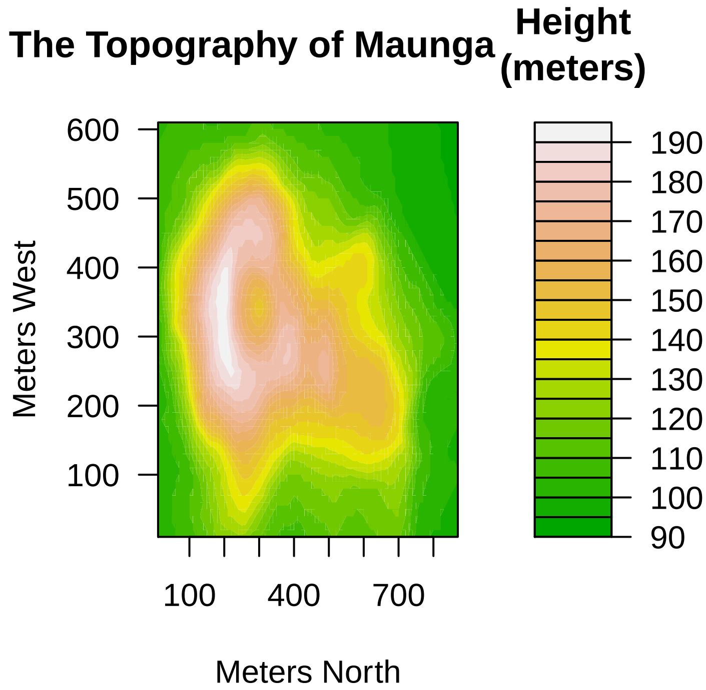
Time series (line plot)



Multiple time series



Contour plot



Describing distributions I

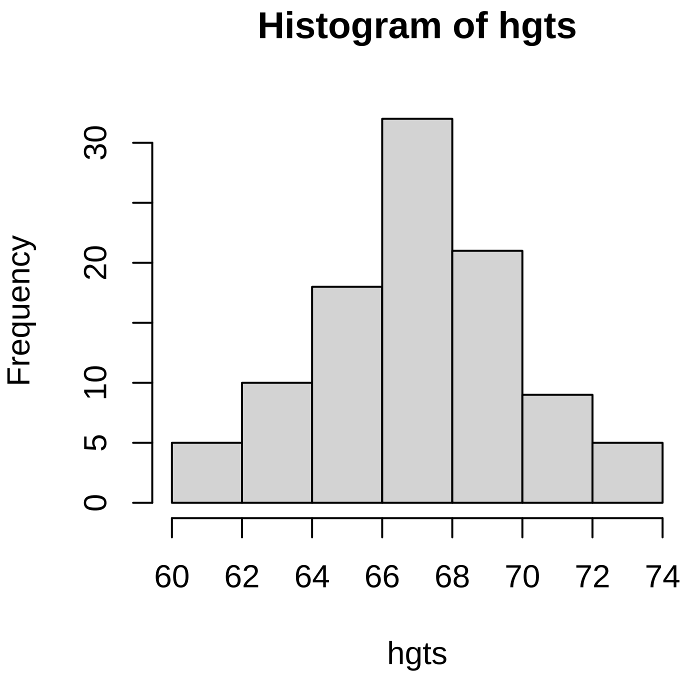
symmetric: if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. In symmetric distributions, the mean and median are the same (approximately equal) and the mode(s) are generally in the center as well

left (or negative) skew: when it looks like the graph is "pulled" to the left (fewer observations to the left than the right); mode(s) generally on right side

right (or positive) skew: when it looks like the graph is "pulled" to the right (fewer observations to the right than the left); mode(s) generally on left side

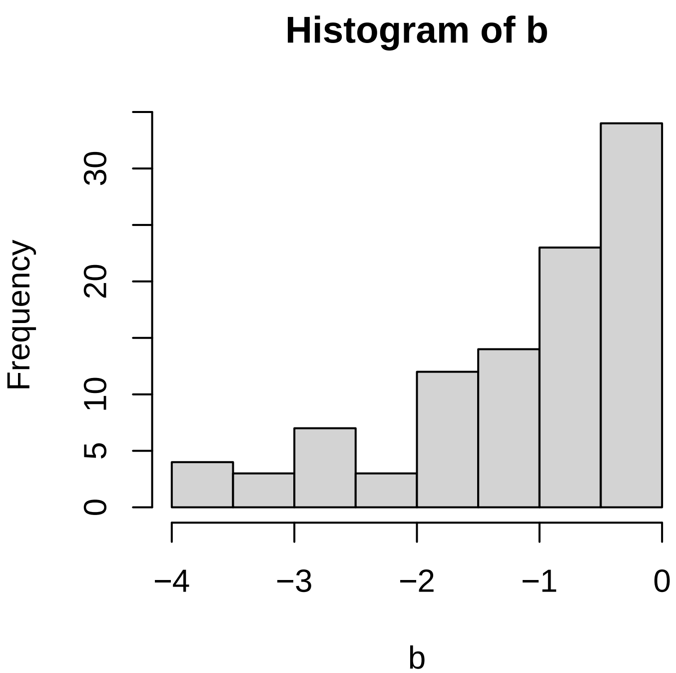
Symmetric

```
hist(hgts)
```



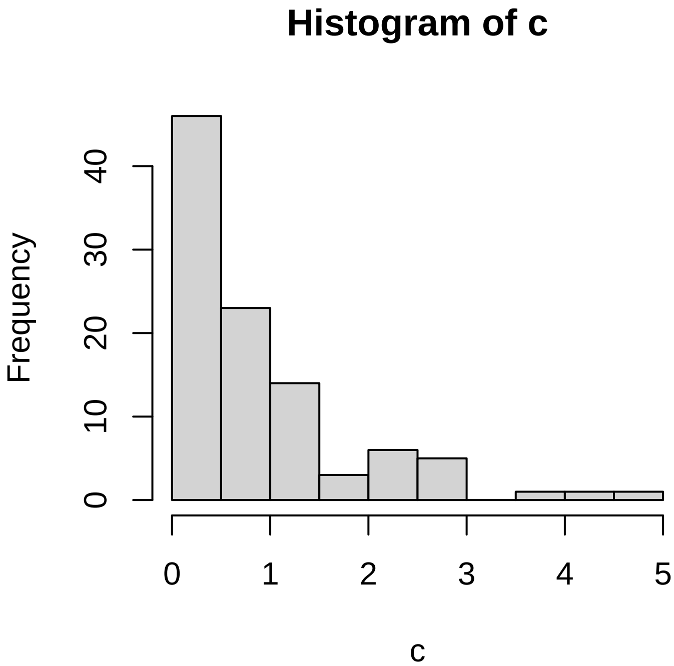
Left skewed

```
hist(b)
```



Right skewed

```
hist(c)
```



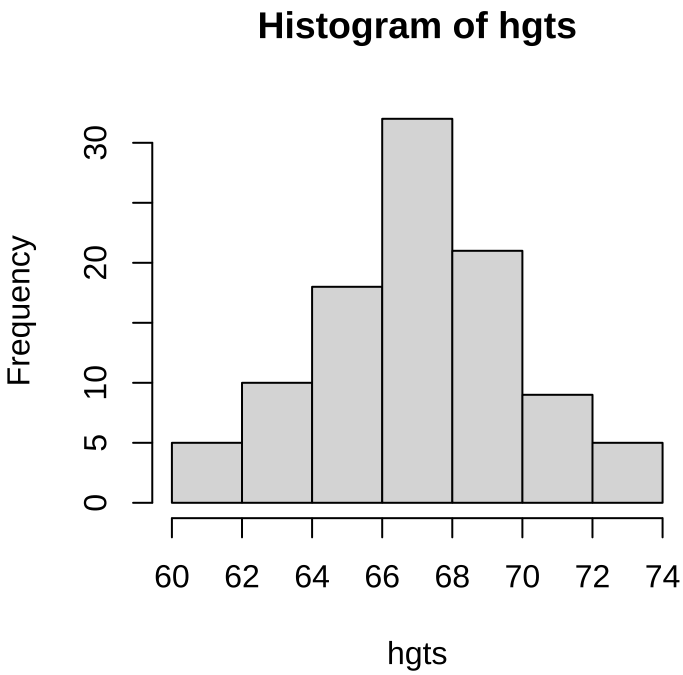
Describing distributions II

unimodal: one main mode in the data set

bimodal: two modes (more than two is multimodal)

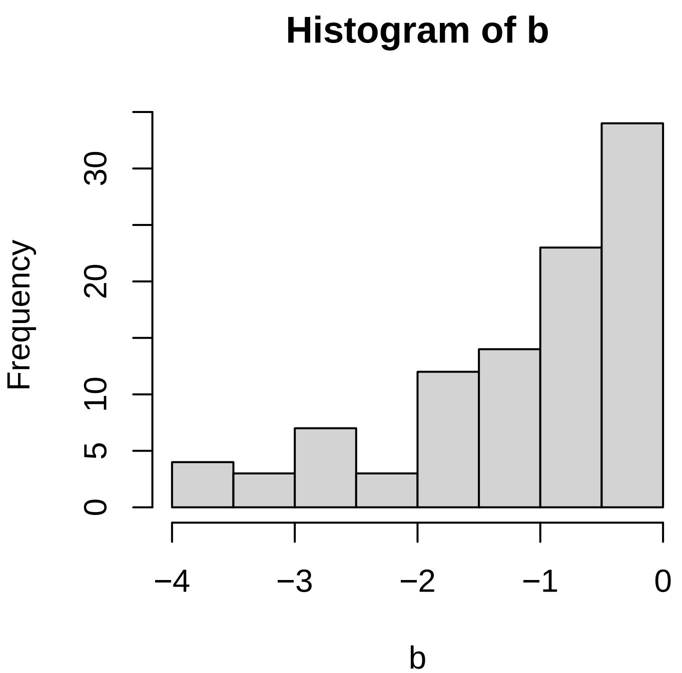
Unimodal

```
hist(hgts)
```



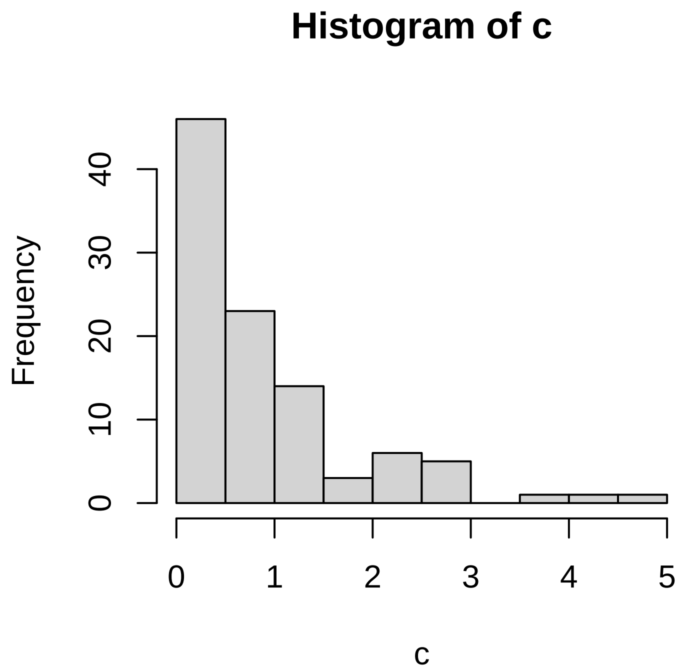
Also unimodal

```
hist(b)
```



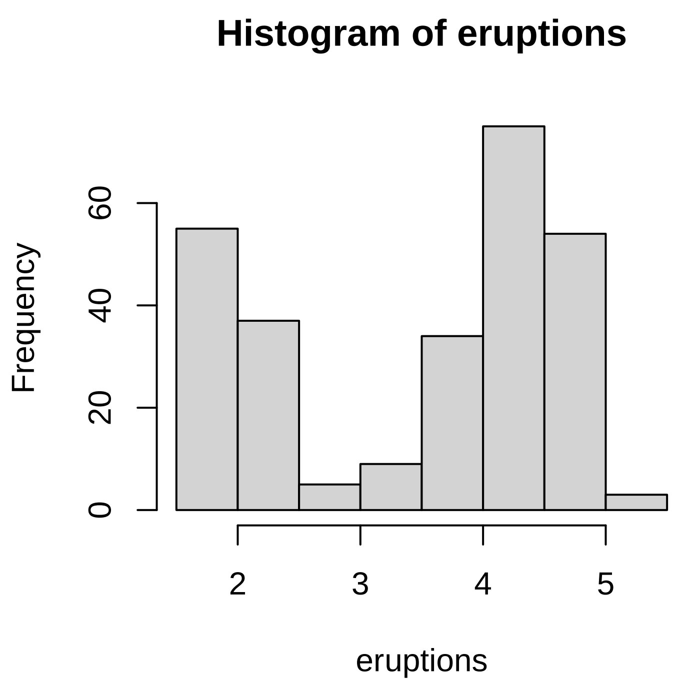
Still unimodal

```
hist(c)
```



Bimodal

```
with(faithful,hist(eruptions))
```



Measures of Location I

percentiles: divide ordered data into hundredths; useful for comparing values, especially with large populations. Ex: unemployment rates, SAT scores

quartiles: divide ordered data into quarters
- $Q1$: quartile 1 refers to the 25^{th} percentile (25% of the data is less than $Q1$ and 75% of the data is more than $Q1$)
- M: the *median* refers to the 50^{th} percentile, the center-most value of the ordered dataset (50% of the data is less than M and 50% of the data is more than M).
- $Q3$: quartile 3 refers to the 75^{th} percentile (75% of the data is less than $Q3$ and 25% of the data is more than $Q3$)

Measures of Location II

mean: the mathematical average. μ represents the population mean and $\overline{X}$ represents the sample mean. It is calculated by the sum of the observation values divided by the number of observations

minimum, maximum: the smallest (*min*) and largest (*max*), respectively value of the dataset

mode: the most frequently ocurring observation(s); can have more than 1, can also have 0

population size: the number of elements in a population; denoted as N (always upper case)

sample size: the number of observations in a dataset; denoted as n (always lower case)

Formula definitions

$Q1$ = median of lower half of ordered dataset
$Q3$ = median of upper half of ordered dataset
min the smallest value
max the largest value
mode the observation(s) that occur most frequently N population size (most often not known or difficult to find)
n the sample size; the number of observations in a dataset

Examples

Dataset: {1,11,6,7,4} and $n = 5$

The Mean

Population values (parameters) are usually denoted with letters from the Greek alphabet, while sample values (statistics) are *usually* denoted with English letters ($\pm$ a few extra symbols here and there)

Sample mean: $\overline{X} = \frac{\sum x_i}{n} =$ where x_i are the values of each observation in the *sample*

Example of sample mean: $\overline{X} = \frac{\sum x_i}{n} = \frac{1+11+6+7+4}{5} = 5.8$

Measures of Variability

IQR (interquartile range): shows the "spread" of the middle 50% of the data; also used to help identify outliers

outlier: a data point that is significantly different than the other data points. Some could be due to data entry errors, some are unique and usually require more investigation

range: the difference between the *max* and *min* values; shows entire "spread" of the data

variance: the average *squared* distance each data point is from its mean

standard deviation: the average distance each data point is from its mean; is used most often as the main measure of variability (thus its name)

Formulas II

$IQR = Q3 - Q1$

A value is considered a potential outlier if it is $< Q1 - IQR(1.5)$ or if $> Q3 + IQR(1.5)$

Sample variance: $s^2 = \frac{\sum(x_i - \overline{X})^2}{n-1} = \frac{(x_1-\overline{X})^2+(x_2-\overline{X})^2+\cdots+(x_n-\overline{X})^2}{n-1}$ where $n - 1 = df$, the degrees of freedom (more to come on df later)

Sample variance: $s = \sqrt{\frac{\sum(x_i-\overline{X})^2}{n-1}} = \sqrt{s^2}$

Note: s^2, and s MUST ≥ 0

More Examples

Dataset: {1,11,6,7,4}

$IQR = Q3 - Q1 = 11 - 1 = 10$; $IQR(1.5) = 10(1.5) = 15$; lower outlier boundary: $= Q1 - IQR(1.5) = 1 - 15 = -14$ and upper outlier boundary: $= Q3 + IQR(1.5) = 11 + 15 = 26$. Since no observations are outside those boundaries, there are no potential outliers in the dataset

Continued Example

$range = max - min = 11 - 1 = 10$

$s^2 = \frac{\sum(x_i-\overline{X})^2}{n-1} = \frac{(1-5.8)^2+(11-5.8)^2+(6-5.8)^2+(7-5.8)^2+(4-5.8)^2}{5-1} = 13.7$

$s = \sqrt{\frac{\sum(x_i-\overline{X})^2}{n-1}} = \sqrt{s^2} = \sqrt{13.7} = 3.7014$

With no idea of population size nor population values, we cannot compute the parameters (population values) but will use the statistics calculated above as estimates of the parameters

Appropriate summaries for different distributions

When looking at data that has a non-symmetric distribution (skewed, bimodal, multimodal, ...), the best statistics to use would be the 5# summary with the IQR, with a boxplot.

When looking at data that has a symmetric distribution (espeically the "normal" distribution – bell curve), the best statistics to use would be the mean and standard deviation, with a histogram (usually – boxplots are ok too when n is small).

Empirical Rule

The Empirical Rule (ER) is appropriate *only* with symmetric distributions. ER states:

68% of observations are within the interval $\overline{X} \pm 1s$
95% of observations are within the interval $\overline{X} \pm 2s$
99.7% of observations are within the interval $\overline{X} \pm 3s$

ER example

Example: $\overline{X} = 15$, $s = 2$
68% of observations are within the interval $\overline{X} \pm 1s = 15 \pm 2 = (13, 17)$
95% of observations are within the interval $\overline{X} \pm 2s = 15 \pm 2(2) = (11, 19)$
99.7% of observations are within the interval $\overline{X} \pm 3s = 15 \pm 3(2) = (9, 21)$