# Descriptive Statistics

Measures of Variation

# Descriptive Statistics

Summarizing Data:

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

# Range

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students

102
128
131
98
140
93
110

115
109
89
106
119
97

**Class A Range = 140 - 89 = 51**

Class B--IQs of 13 Students

127
131
96
80
93
120
109

162
103
111
109
87
105

**Class B Range = 162 - 80 = 82**

# Interquartile Range

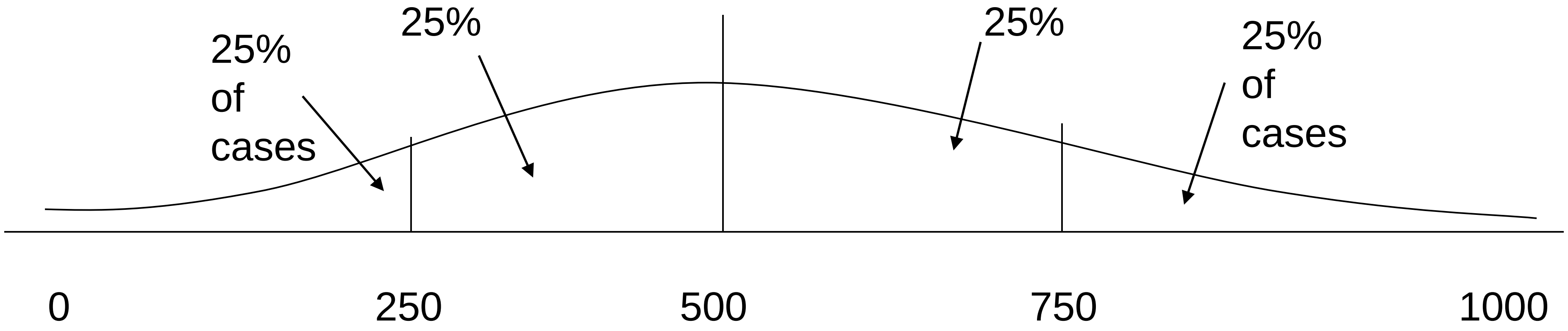
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.

75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

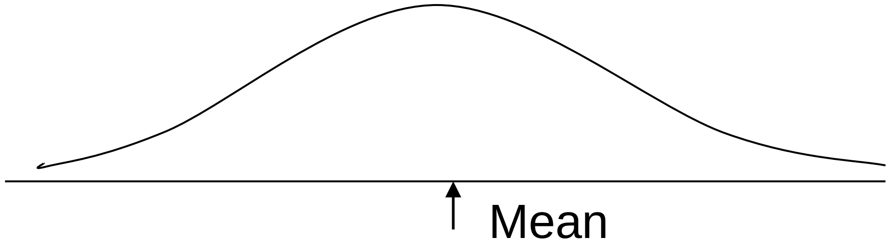
The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



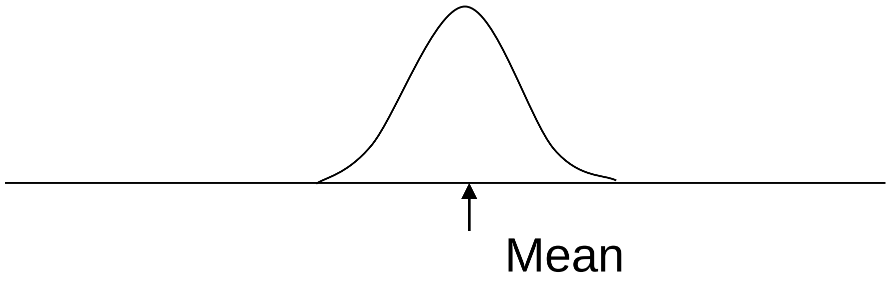
# Variance

A measure of the spread of the recorded values on a variable.  A measure of dispersion.

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



# Variance and Standard Deviation of a Population

- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate negative signs.

$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} \equiv \sigma^2,$$

$$\sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}} \equiv \sigma$$

# Variance, S.D. of a Sample

$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1} \equiv s^2,$$

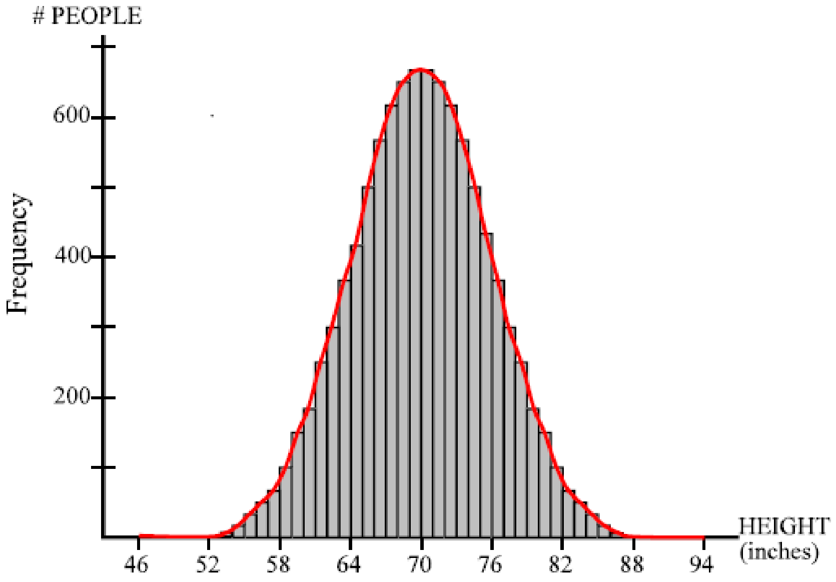
**Degrees of freedom**

$$\sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1}} \equiv s$$

# Standard Deviation

- Note about computational formulas:
  - Your book provides a useful short-cut formula for computing the variance and standard deviation.
  - This is intended to make hand calculations as quick as possible.
  - They obscure the conceptual understanding of our statistics.

# Normal distribution example

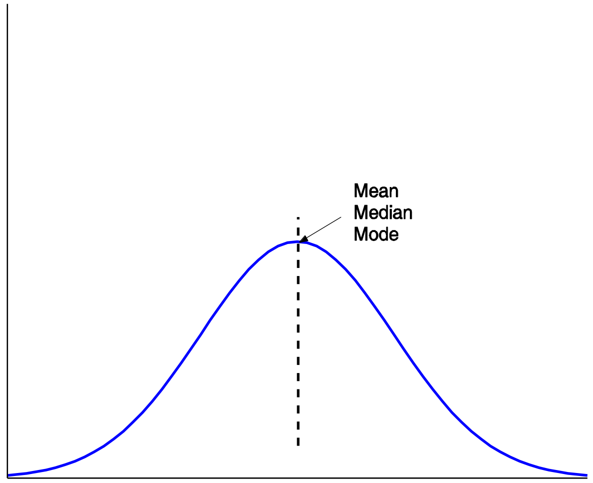


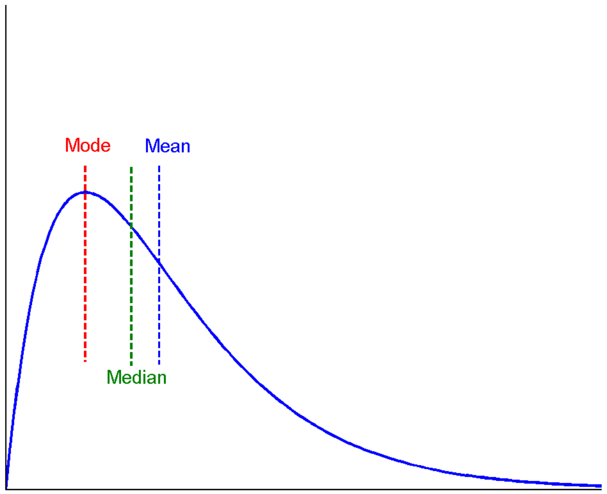
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

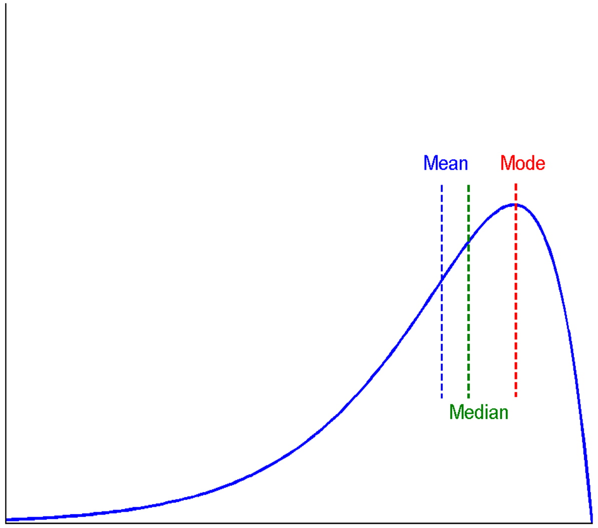
- "No skew"
- "Zero skew"
- Symmetrical
- Mean = median = mode

# Symmetric vs. Skewed Data

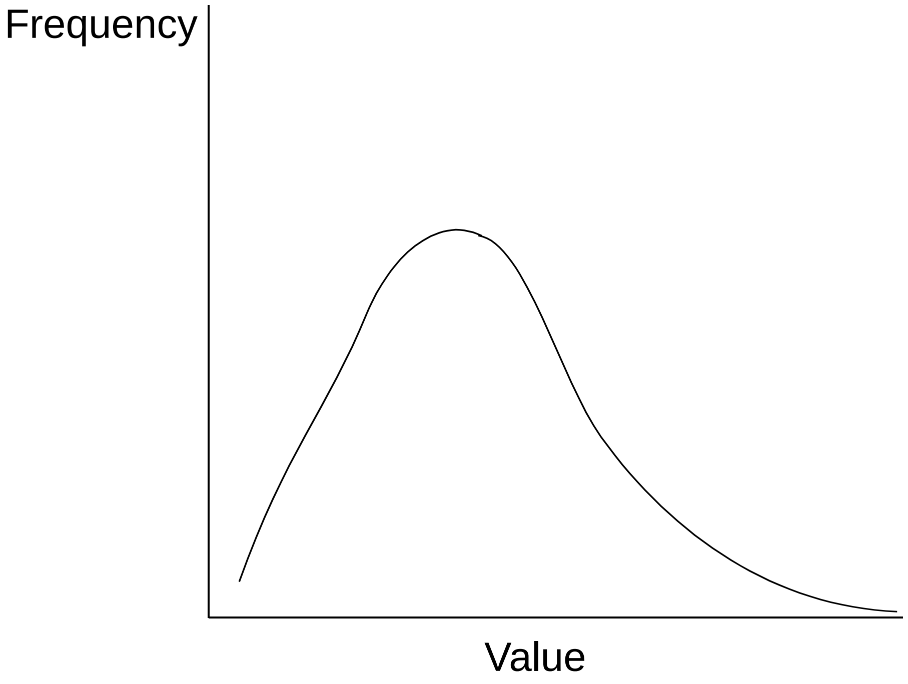
- Median, mean and mode of symmetric, positively and negatively skewed data







# Skewness
## Asymmetrical distribution

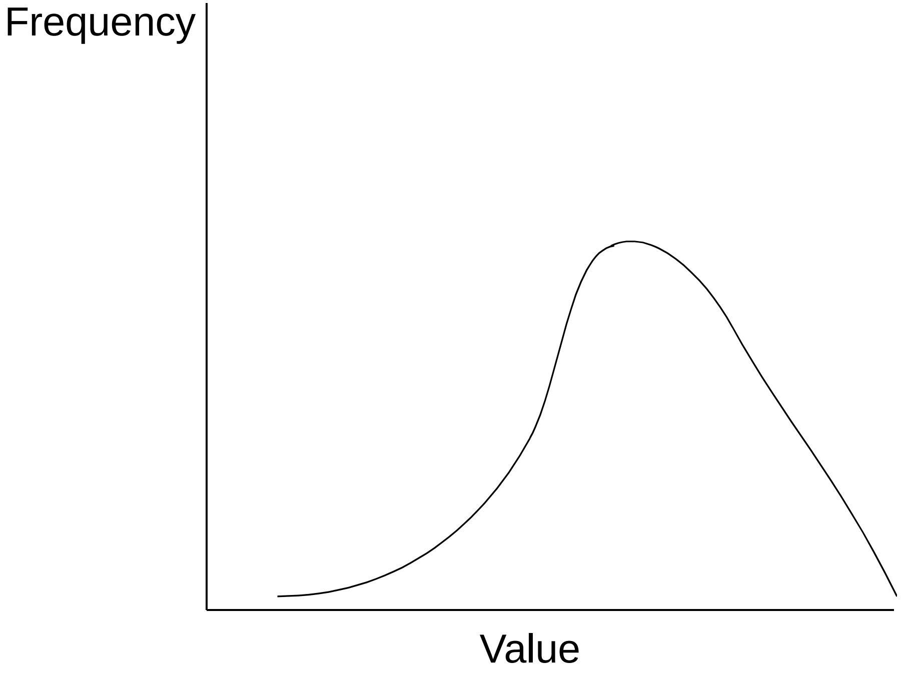


- Income
- Contribution to candidates
- Populations of countries
- “Residual vote” rates

- “Positive skew”
- “Right skew”

# Skewness

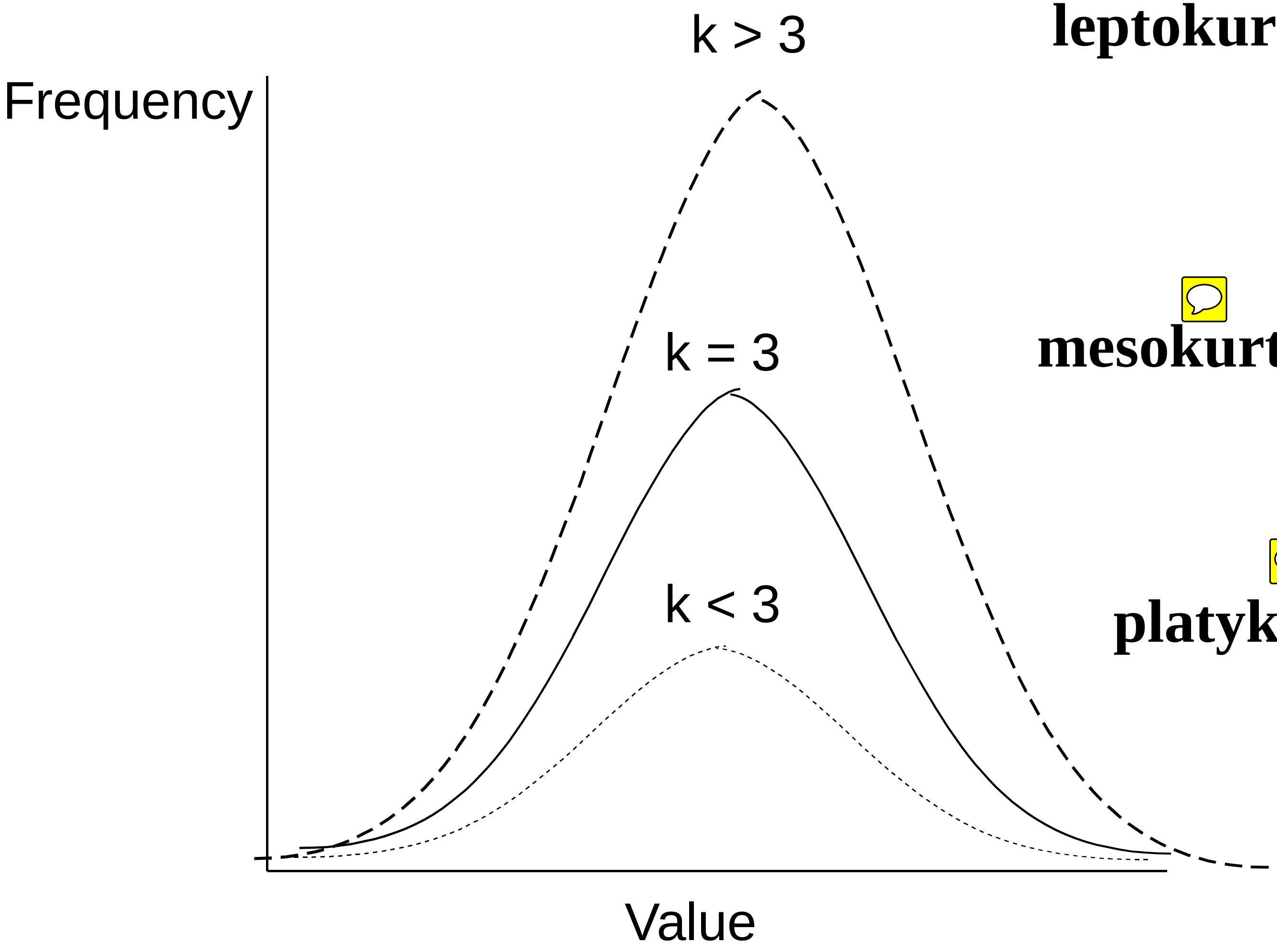
## Asymmetrical distribution



- GPA of NUS students

- “Negative skew”
- “Left skew”

# Kurtosis

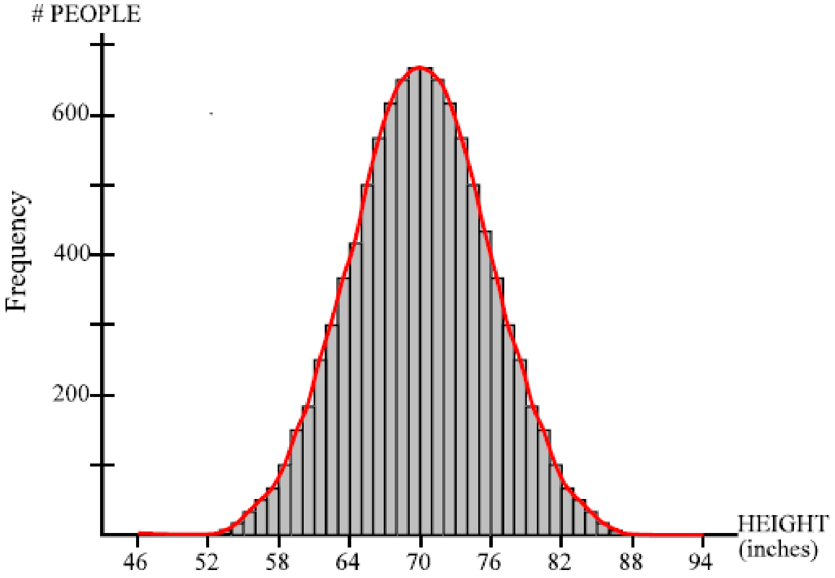


**leptokurtic**

**mesokurtic**

**platykurtic**

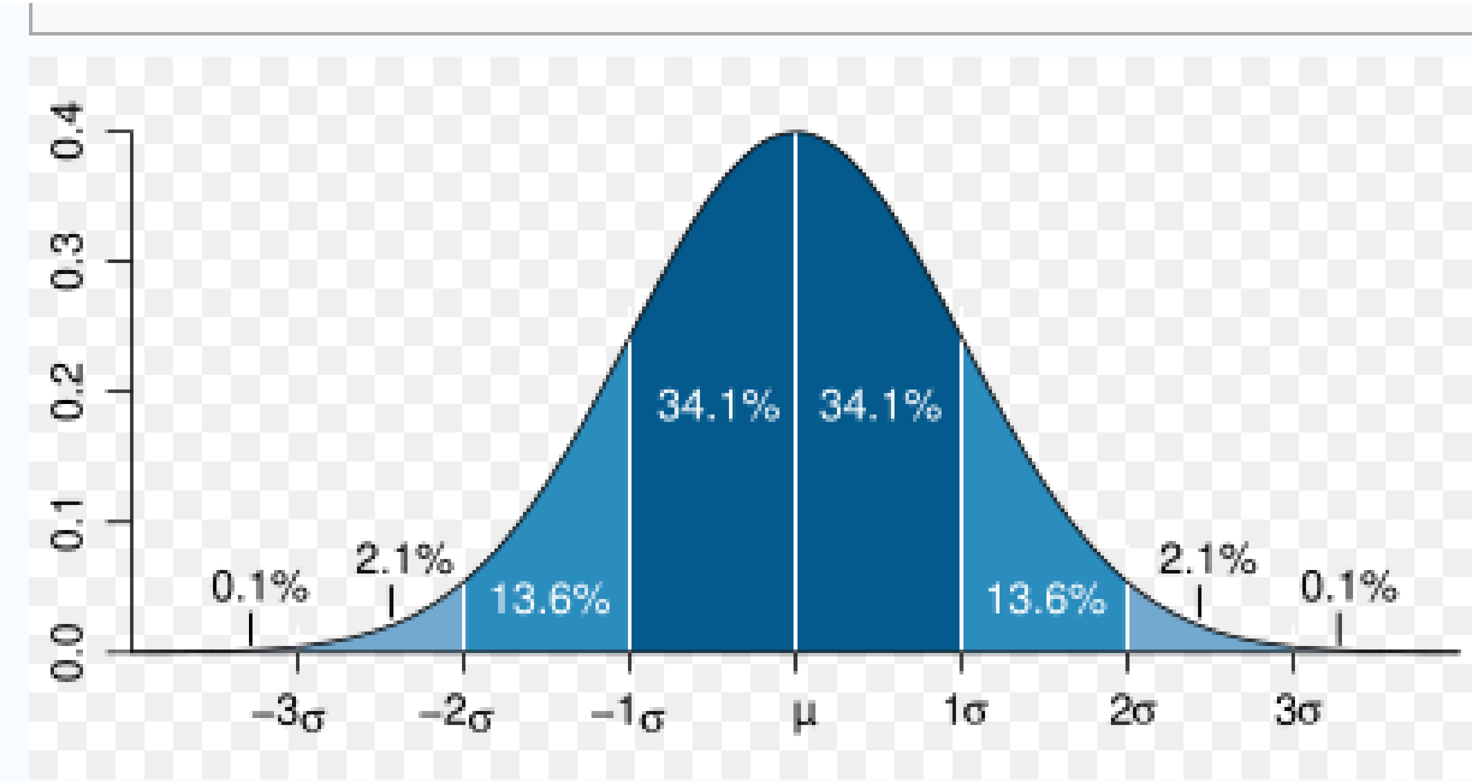
# Normal distribution



- Skewness = 0
- Kurtosis = 3

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

# More words about the normal curve



# Summary: Measures of the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: $Q_1$ (25th percentile), $Q_3$ (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, $Q_1$, M, $Q_3$, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

  - Standard deviation *s (or σ)* is the square root of variance *$s^2$ (or $\sigma^2$)*

# Descriptive Statistics

Summarizing Data:

- ✓Central Tendency (or Groups' "Middle Values")
  - ✓Mean
  - ✓Median
  - ✓Mode

- ✓Variation (or Summary of Differences Within Groups)
  - ✓Range
  - ✓Interquartile Range
  - ✓Variance
  - ✓Standard Deviation
  - ✓Higher moments (Skewness, Kurtosis, …)

# Box-Plots

A way to graphically portray almost all the descriptive statistics at once is the box-plot.

A box-plot shows:

Upper and lower quartiles

Mean

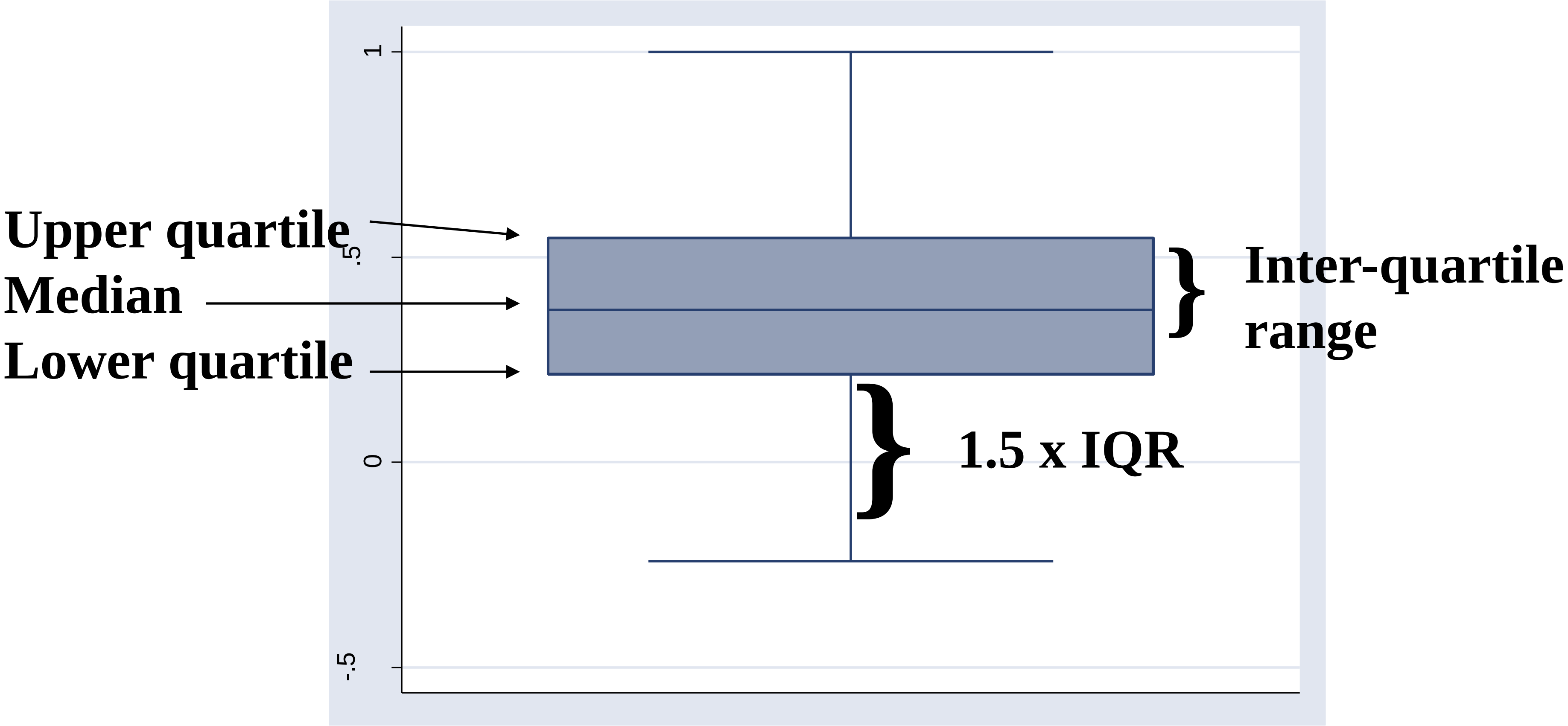
Median

Range

Outliers (1.5 IQR)

# Draw the graph with a box plot

```
. graph box totalscore
```



# IQV—Index of Qualitative Variation

To calculate:

$$\text{IQV} = \frac{K(100^2 - \Sigma \text{ cat.}\%^2)}{100^2(K - 1)}$$

K=# of categories

Cat.% = percentage in each category

# IQV—Index of Qualitative Variation

- For nominal variables
- Statistic for determining the dispersion of cases across categories of a variable.
- Ranges from 0 (no dispersion or variety) to 1 (maximum dispersion or variety)
- 1 refers to even numbers of cases in all categories, NOT that cases are distributed like population proportions
- IQV is affected by the number of categories