Measures of Central Tendency

Mode: most frequent score
- best average for nominal data
- sometimes none or more than one mode in a sample
- bimodal or multimodal distributions indicate several groups included in sample
- easy to determine

Midrange: mean of highest and lowest scores
- easy to compute, rough estimate, rarely used

Median: value that divides distribution in half
- best average for ordinal data
- more appropriate average for skewed ratio or interval data or data on salaries
- difficult to compute because data must be sorted
- unaffected by extreme data

Arithmetic Mean: centre of balance of data
- sum of numbers divided by *n* (sample) or *N* (population)
- best average for unskewed ratio or interval data
- easy to compute

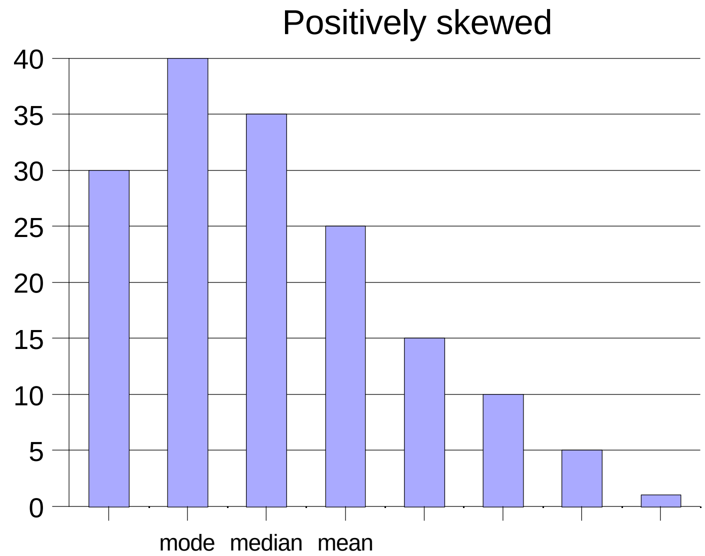
sample mean= $\overline{X} = \dfrac{\Sigma X}{n}$ population mean = $\mu = \dfrac{\Sigma X}{N}$

Other Measures: harmonic mean, geometric mean and quadratic mean, also called root mean square (RMS)

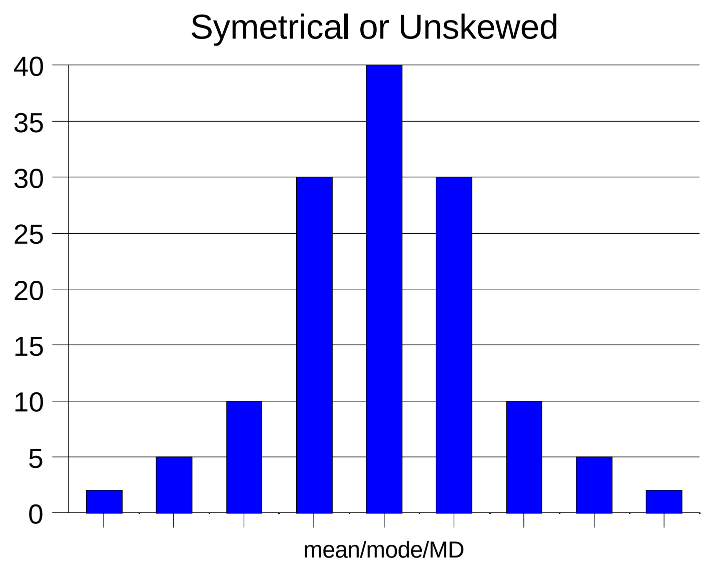
$$\text{RMS} = \sqrt{(\Sigma X^2 / n)}$$

Skewed Data

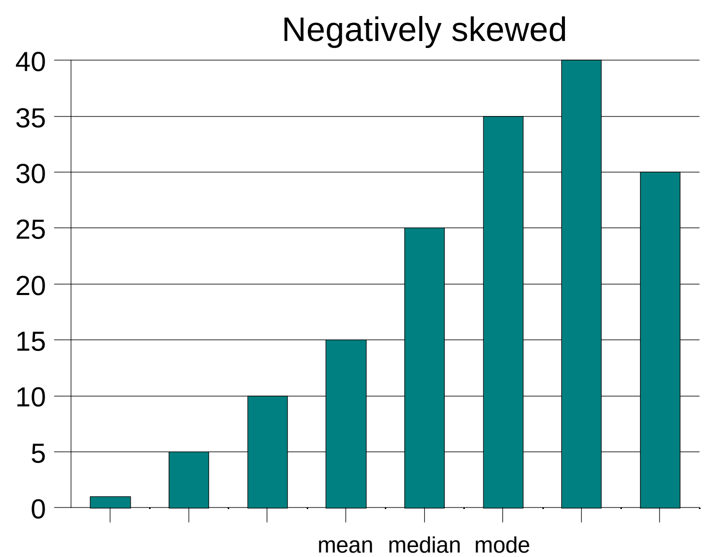
- direction of skew is the direction of the tail
- positive direction of a number line is to the right, left is negative direction



- mean, mode and median (MD) are the same for symmetrical distributions



- notice mean is closest to the tail (i.e., more influenced by extreme values)



Measures of Variation

Range: -highest minus lowest values

- used for ordinal data

R = highest - lowest

Interquartile Range: -75th minus 25th percentile

- used for determining "outliers"

$$\mathbf{IQR = Q_3 - Q_1}$$

Variance: -mean of squared differences between scores and the mean

- used on ratio or interval data
- used for advanced statistical analysis (ANOVAs)

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

Standard Deviation: -has same units as raw data

- used on ratio or interval data
- most commonly used measure of variation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$$

Coefficient of Variation: -percentage of standard deviation to mean

- used to compare variability among data with different units of measure

$$\text{CVar} = \frac{\sigma}{|\mu|} \times 100\% \quad \text{or} \quad \text{CVar} = \frac{s}{|\overline{X}|} \times 100\%$$

Biased and Unbiased Estimators

- sample mean is an unbiased estimate of the population mean
- same variances and standard deviations are **biased** estimators because the sample mean is used in their computation

Why? Last score can be determined from mean and all other scores, therefore, it is not free to vary or add to variability. To compensate divide sums of squares by n-1 instead of n.

Computing Formulae

- Instead of using the standard formula, a computing formula is used so that running totals of scores and scores squared may be used to compute variability.

Variance: s^2 = sample variance

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}$$

Standard Deviation: s = sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}}$$

Measures of Position

Percentile: score which exceeds a specified percentage of the population

- suitable for ordinal, ratio or interval data
- median (MD or Q_2) is 50^{th} percentile
- first and third quartiles (Q_1 and Q_3) are 25^{th} and 75^{th} percentiles
- easier for nonstatisticians to understand than z-scores
- scores are all positive numbers

$$\text{percentile} = \frac{\text{number of values below X} + .05}{\text{total number of values}} \times 100\%$$

Standard or z-scores: based on mean and standard deviation and the “normal” distribution

- suitable for ratio and interval numbers
- approximately 68% of scores are within 1 standard deviation of the mean, approximately 95% are within 2 standard deviations and approximately 99% are within 3 standard deviations
- half the scores are negative numbers
- mean score is zero
- excellent way of comparing measures or scores which have different units (i.e., heights vs. weights, metric vs. Imperial units, psychological vs. physiological measures)

$$z = \frac{X - \mu}{\sigma} \quad \text{or} \quad z = \frac{X - \overline{X}}{s}$$

Measures of Position and Outliers

Other Measures of Position:

Deciles: 10^{th}, 20^{th}, ... 100^{th} percentiles (D_1, D_2, ...D_{10})

- often used in education or demographic studies

Quartiles: 25^{th}, 50^{th} and 75^{th} percentiles (Q_1, Q_2, Q_3)

- frequently used for exploratory statistics and to determine outliers (Q_2 is same as median)

Outliers:
extreme values that adversely affect statistical measures of central tendency and variation

Method of Determining Outliers:

- compute interquartile range (IRQ)
- multiply IRQ by 1.5
- lower bound is Q_1 minus $1.5 \times$ IRQ
- upper bound is Q_3 plus $1.5 \times$ IRQ
- values outside these bounds are outliers and may be removed from the data set
- it is assumed that outliers are the result of errors in measurement or recording or were taken from an unrepresentative individual

Alternate Method for Normally Distributed Data:

- values outside +/- 4 or 5 standard deviations are considered outliers