Measures of Variability

LECTURE 4

Objectives

- Define terms.
- Diagram relative differences in central tendency and variability.
- Calculate the range, sum of squares, variance, and standard deviation.
- Calculate the diversity of a sample of nominal observations.

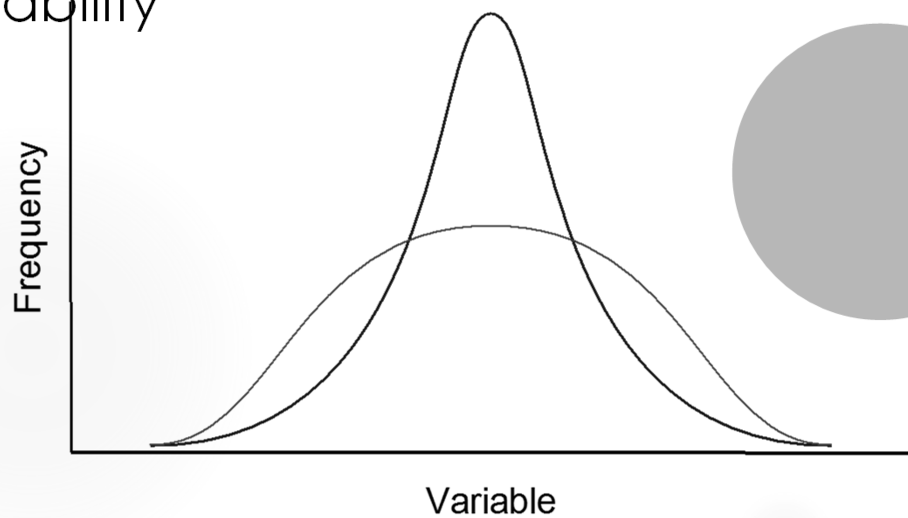
Variability

- A measure of variability is an indication of the spread of measurements around the center of the distribution.

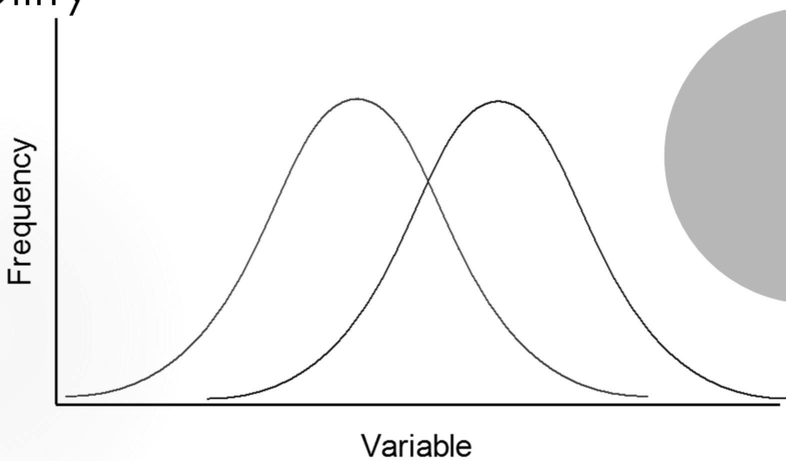
Variability

- Measures of variability are parameters of the population.
- Sample measures that estimate the variability of the population are statistics.
  - Range
  - Variance
  - Standard deviation

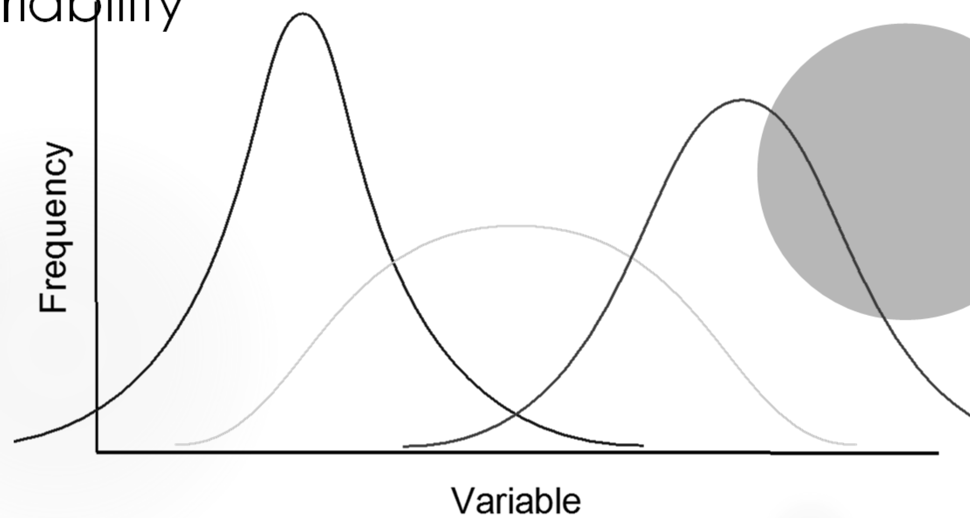
Measures of centrality and variability



Measures of centrality and variability



Measures of centrality and variability



Range

- The difference between the highest and lowest measurements in a group of data.
  - Statistical range vs. mathematical range
- In a data array (smallest to largest):

Sample range = X_n-X_1

Range

- Gives some indication of the variability of data, but it only depends on the extreme values of the data array (largest and smallest).
- It is unlikely that the sample will contain the extreme values of the population so the sample range will consistently underestimate the population range (a biased estimate).

Biased and unbiased

- A biased estimator will consistently under- or over-estimate the value of a parameter.
- An unbiased estimator will not always (or even often) give the correct value of a parameter, but it will over-estimate the parameter as often as it underestimates the parameter.

Sum of squares

- Since the mean is a useful measure of central tendency, it is possible to express variability in terms of deviation from the mean.

$$(X_i - \overline{X})$$

Sum of squares

- The sum of all deviations from the mean will always equal zero.

$$\sum (X_i - \overline{X}) = 0$$

- Positive deviations are cancelled by negative deviations.

Sum of squares

- Squaring the deviations from the mean is one way of eliminating the signs from the deviations.
- The sum of the squares of the deviations from the mean is called the sum of squares (SS).

Sum of squares

- Population sum of squares

$$SS = \sum (X_i - \mu)^2$$

- Sample sum of squares

$$SS = \sum (X_i - \overline{X})^2$$

Variance

- The mean sum of squares is called the variance.
  - Squared units
- Population variance

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

Variance

- Sample variance

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Note difference in the denominator

Standard deviation

- The standard deviation is the positive square root of the variance.
- Population standard deviation

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Standard deviation

- Sample standard deviation

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

Coefficient of variation

- The coefficient of variation expresses sample variability relative to the sample mean.

$$CV = \frac{s}{\overline{X}}$$

Diversity

- For nominal scale data there is no mean to serve as a reference for variability. Instead the concept of diversity (the distribution of observations among categories) is used.
- The index is used in a relative fashion (comparison not absolute measure)

Diversity

- Takes into account both numbers of species and the evenness of the distribution among categories.

Diversity

- n = number of individuals in the sample
- fi = number of observations in category i

$$H' = \frac{n(\log n) - \sum_{i=1}^{k} f_i \log f_i}{n}$$

Evenness

- K = number of categories (species)

$$J' = \frac{H'}{H_{\max}}$$

$$H_{\max} = \log K$$