VARIANCE AND THE NORMAL DISTRIBUTION

Descriptive Statistics

Measures of Variation

Range

= largest value - smallest value

The range is the difference between the maximum value and the minimum value in the data.

Interquartile-range (IQR)

= Q3-Q1

- The interquartile range is the difference between the upper and lower quartiles. It is the middle 50% of the data.
- The **lower quartile** (LQ, Q1) separates the smallest 25% (the first quarter) of data values from the rest of the data set, i.e. the lower quartile is the 25th percentile.
- The **upper quartile** (UQ, Q3) separates the largest 25% (the third quarter) of data values from the rest of the data set, i.e. the upper quartile is the 75th percentile.
- The **second quartile** (Q2) is the median, 50% of the values are below the median and 50% above i.e. 50th percentile.
- Note: The IQR can help to determine potential outliers. An extreme value or outlier is a value located far away from the mean. A value is suspected to be a potential outlier if it is less than 1.5*IQR below the first quartile or more than 1.5*IQR above the third quartile. Potential outliers always require further investigation.

Outliers: Observations that fall below Q1-1.5(IQR) or above Q3+1.5(IQR)

Variance

Measure of variation based on squared deviations from the mean; directly related to the standard deviation.

Standard deviation

Measure of variation based on squared deviations from the mean; directly related to the variance. The standard deviation is a measure of how a set of data is clustered or distributed around its mean.

Coefficient of variation

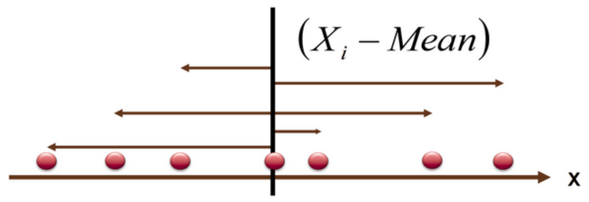
This is a relative measure of variation that is expressed as a percentage rather than in terms of the units of the data; it shows the deviation in the data relative to the mean.

Deviation

Deviation is the distance between a data point and the mean.

Its purpose is to measure and describe the variation in a set of data.

Deviation = (X - Mean)



Mean Absolute Deviation (MAD)

$$\text{MAD} = \frac{\sum |x - \overline{x}|}{n}$$

Degrees of Freedom

- Related to number of observations and the number of decisions to be made.
- The degrees of freedom will vary according to the statistical test and parameters used.
- They are the independent elements
- Sometimes abbreviated as 'DOF' or 'DF'

Sample Deviation(s) Formula

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\boxed{n-1}}}$$

Population Standard Deviation

Measure of the “average” deviation about the mean.

Population Standard Deviation (σ) Formula

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

$\sum_{i=1}^{N}$ = Sum all values from the first to last

Population implies ALL data or a census

N is count of ENTIRE population, μ is mean of ENTIRE population

$(X_i - \mu)$	Deviation
$(X_i - \mu)^2$	Deviation squared
$\sum(X_i - \mu)^2$	Sum of the Squares = SS
$\frac{\sum(X_i - \mu)^2}{N}$	Variance
$\sqrt{\frac{\sum(X_i - \mu)^2}{N}}$	Standard Deviation

Sample Standard Deviation

Measure of the “average” about the mean.

Sample Standard Deviation (s) Formula:

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

Sample implies a SUBSET of the data in the population

n is the count of the sample

is the mean of the sample

Example:

Determine the sample standard deviation for the following data set:

1 2 3

The mean is **2**

n = 3

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

$$s = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{2}}$$

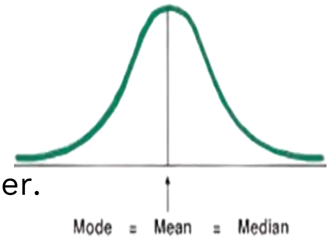
$$s = \sqrt{1} = 1.00$$

Descriptive Statistics

Measures of Shape

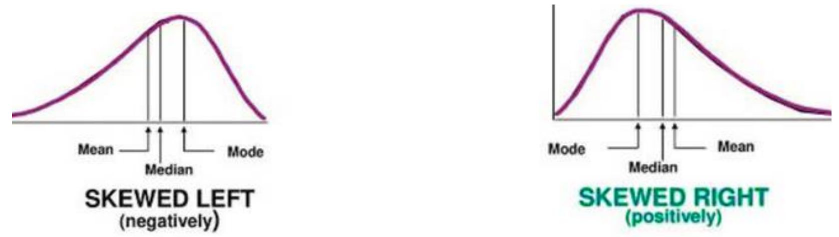
Symmetrical (bell-shaped) or normal

A distribution is symmetrical (bell-shaped or normal) if the lower and upper halves of the graph are mirror images of each other.



Not Symmetrical (skewed)

If the distribution is not symmetrical, skewed.



Skewed to the left or negatively skewed: There are relatively few small data values, and most of the values are concentrated in the upper portion of the distribution. The tail extends toward smaller values (to the left).

e.g. the age disribution in a retirement community

Skewed to the right or positively skewed: There are relatively few large data values, and most values are concentrated in the lower portion of the distribution. The tail extends toward larger values (to the right).

E.g., income distribution

Measurements of asymmetry

Skewness

It measures the lack of symmetry in the data and is based on a statistic that is a function of the cubed differences around the mean. Positive and negative values indicate positive or negative skewness.

Kurtosis

It measures whether the data points in the distribution have heavier, or lighter tails compared to a normal distribution.

This material is provided to you as a Macquarie University student for your individual research and study purposes only. You cannot share this material publicly online without permission. Macquarie University is the copyright owner of (or has a license to use) the intellectual property in this material. Legal and /or disciplinary actions may be taken if this material is shared without the university's written permission.