

### III. Descriptive Statistics - Numerical Methods

#### A. Measures of Location - Qualitative Data

1. Proportion - relative frequency that a characteristic occurs in a data set. The population proportion is

$$p = \frac{\text{number of population elements with characteristic}}{\text{total number of elements in population}} = \frac{\text{number of population elements with characteristic}}{N}$$

while the sample proportion is

$$\bar{p} = \frac{\text{number of sample observations with characteristic}}{\text{total number of observations in sample}} = \frac{\text{number of sample elements with characteristic}}{n}$$

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with:

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

the proportion of elements that are male is

$$\bar{p} = \frac{\text{number of males in sample}}{\text{number of observations in sample}} = \frac{6}{20} = 0.30$$

the proportion of elements with values in excess of 30 is

$$\bar{p} = \frac{\text{number of sample observations with with a value over 30}}{\text{total number of observations in sample}} = \frac{4}{20} = 0.20$$

Note that  $0 \leq p \leq 1$  and  $0 \leq \bar{p} \leq 1$  !

---

---

---

---

---

---

---

---

#### B. Measures of Location - Quantitative Data

1. Midrange - value half the distance between the minimum and maximum values in a data set.

$$\text{midrange} = \text{minimum value} + \frac{\text{maximum value} - \text{minimum value}}{2}$$

Example - for the data array that we have been working with, the midrange is:

$$10 + \frac{36 - 10}{2} = 10 + \frac{26}{2} = 10 + 13 = 23$$

---

---

---

---

---

---

---

---

2. Arithmetic Mean - measure of central location calculated by summing all values in a data set and dividing by the number of summed values. The population mean is

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

while the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with, the mean is:

$$\bar{x} = \frac{10 + 11 + 12 + \dots + 36}{20} = \frac{413}{20} = 20.65$$

---

---

---

---

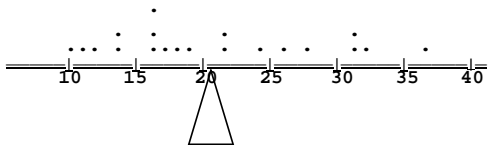
---

---

---

---

Note that the mean is the point at which you would place a fulcrum under the axis of a dot plot to 'balance' the data



that is, it is the point at which the sum of all positive differences from the mean and the absolute value of the sum of all negative differences from the mean are equal!

---

---

---

---

---

---

---

---

Why does this ALWAYS happen? Suppose you have N observations and subtract the mean  $\mu$  from each:

$$x_1 - \mu =$$

$$x_2 - \mu =$$

$$x_3 - \mu =$$

$$\cdot \cdot$$

$$\cdot \cdot$$

$$\cdot \cdot$$

$$x_{N-1} - \mu =$$

$$x_N - \mu =$$

$$\sum_{i=1}^N x_i - N\mu = \sum_{i=1}^N x_i - N \frac{\sum_{i=1}^N x_i}{N} = \sum_{i=1}^N x_i - \sum_{i=1}^N x_i = 0!$$

---

---

---

---

---

---

---

---

---

---

Example: for the data array that we have been working with:

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36



...so the mean distance of the data from their mean is 0.00 (as it always will be!).

---

---

---

---

---

---

---

---

---

---

3. T% Trimmed (Arithmetic) Mean - arithmetic mean that results after the most extreme (largest and smallest) T% of values have been eliminated from the data. The population T% trimmed mean is

$$\mu_{T\%} = \frac{\sum_{i=j+1}^{N-j} x_i}{N - 2j}$$

where the data have been arranged in ascending order and

used to calculate the start and end values of the index  $i$ :

$$j = \left\lfloor \frac{T}{200} N \right\rfloor$$

the Floor Operator

...and to calculate the denominator

is the largest integer that does not exceed  $\frac{T}{200} N$ .

---

---

---

---

---

---

---

---

---

---

The sample T% Trimmed (Arithmetic) Mean is

$$\bar{x}_{T\%} = \frac{\sum_{i=j+1}^{n-j} x_i}{n - 2j}$$

where the data have been arranged in ascending order and used to calculate the start and end values of the index  $i$ :  $j = \left\lfloor \frac{T}{200} n \right\rfloor$  ...and to calculate the denominator

is the largest integer that does not exceed  $\frac{T}{200} n$ .

In both the population and sample case, the trimming is performed to reduce the influence of extreme values.

---

---

---

---

---

---

---

---

Example - if we want to find the 15% trimmed mean for the data array that we have been working with:

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

we must first use the value of  $j$  to calculate the start and end values of the index  $i$ :

$$j = \left\lfloor \frac{T}{200} n \right\rfloor = \left\lfloor \frac{15}{200} 20 \right\rfloor = \lfloor 1.50 \rfloor = 1.0$$

so the trimmed mean is

$$\bar{x}_{15\%} = \frac{\sum_{i=j+1}^{n-j} x_i}{n - 2j} = \frac{\sum_{i=2}^{19} x_i}{20 - 2} = \frac{\sum_{i=2}^{19} x_i}{18} = \frac{11 + 12 + \dots + 31 + 32}{18} = \frac{367}{18} = 20.38\bar{8}$$

---

---

---

---

---

---

---

---

Example - if we want to find the 20% trimmed mean for the data array that we have been working with:

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

we must first use the value of  $j$  to calculate the start and end values of the index  $i$ :

$$j = \left\lfloor \frac{T}{200} n \right\rfloor = \left\lfloor \frac{20}{200} 20 \right\rfloor = \lfloor 2.00 \rfloor = 2.0$$

so the trimmed mean is

$$\bar{x}_{20\%} = \frac{\sum_{i=j+1}^{n-j} x_i}{n - 2j} = \frac{\sum_{i=3}^{18} x_i}{20 - 4} = \frac{\sum_{i=3}^{18} x_i}{16} = \frac{12 + 14 + \dots + 31 + 31}{16} = \frac{324}{16} = 20.25$$

Note that trimmed means are often used in Olympic scoring to minimize the effects of extreme ratings possibly caused by biased judges.

---

---

---

---

---

---

---

---

What if we are interested in some mean rate of change. For example, suppose we have invested \$1000 in some stock on January 1, 2002. If the value of our investment was \$2,000 January 1, 2003, we earned a return of

$$R_1 = \frac{\$2000 - \$1000}{\$1000} = 1.00$$

or 100.0% during the first year (2002). If the value of our investment was \$1,000 on January 1, 2004, we earned a return of

$$R_2 = \frac{\$1000 - \$2000}{\$2000} = -0.50$$

or -50.0% during the second year (2003).

So is the mean rate of return  $\frac{1.00 + (-0.50)}{2} = 0.25$ ?

How can this be if we have the same amount we initially invested?

---

---

---

---

---

---

---

---

---

---

4. Geometric Mean - the  $n^{\text{th}}$  root of the product of  $n$  values. The geometric mean of a population is:

$$\mu_g = \sqrt[N]{\prod_{i=1}^N (1 + R_i)} = \sqrt[N]{(1 + R_1)(1 + R_2) \cdots (1 + R_{N-1})(1 + R_N)}$$

and the geometric mean of a sample is:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n (1 + R_i)} = \sqrt[n]{(1 + R_1)(1 + R_2) \cdots (1 + R_{n-1})(1 + R_n)}$$

The geometric mean is usually used to compute mean growth rates over multiple time periods.

---

---

---

---

---

---

---

---

---

---

Consider our previous example: We invested \$1000 in some stock on January 1, 2002; the value of our investment was \$2,000 on January 1, 2003 and \$1,000 on January 1, 2004. The geometric mean is

$$\begin{aligned} \mu_g &= \sqrt[N]{\prod_{i=1}^N (1 + R_i)} = \sqrt[2]{(1 + 1.00)(1 - 0.50)} \\ &= \sqrt{(2.00)(0.50)} = \sqrt{1.00} = 1.00 \end{aligned}$$

so we still have exactly what we invested (100%) so our return over the two year period (2002 and 2003) is 0.0%!

This makes sense!

---

---

---

---

---

---

---

---

---

---



Would the arithmetic mean of the individual annual returns (3% in the first year, 4% in the second and third years, 5% in the fourth year, and 15% in the fifth year) work?

The arithmetic mean of the individual annual returns is:

$$\bar{x} = \frac{.03 + .04 + .04 + .05 + .15}{5} = \frac{.32}{5} = 0.064$$

If this investment earns 6.4% annually for five years, it would earn a total of

$$(1.046)(1.064)(1.064)(1.064)(1.064)\$Y = (1.396288117)\$Y$$

or 39.6288117% over the entire five year period!

Note this is the same return we erroneously calculated when we simply found the arithmetic mean of the five year return!

---

---

---

---

---

---

---

---

Example: Suppose you have invested \$Y in a five-year certificate of deposit that guarantees a return on your investment of 25% in five years. What is the mean annual return on your investment?

At the end of five years, your investment would be worth 1.25 times its initial value. Thus, the geometric mean is

$$\bar{x}_g = \sqrt[5]{1.25} = 1.04564$$

...so the mean annual return is actually 4.564%.

Check the five year return -  $1.04564^5 = 1.25$  - the five year return is *exactly* 25%!

---

---

---

---

---

---

---

---

For the same investment (\$Y in a five-year certificate of deposit that guarantees a return on your investment of 25% in five years), the *arithmetic mean annual return* is

$$\bar{x} = \frac{.25}{5} = 0.05$$

However, if you actually earned 5% annually, your return on investment after five years would be

$$1.05 \cdot (1.05 \cdot (1.05 \cdot (1.05 \cdot (1.05(\$Y)))))) = 1.05^5 (\$Y) = 1.27628 (\$Y)$$

for a five year return of 27.628% (which exceeds the 25% you are actually earning) - the arithmetic mean is again misleading - it overstates the true annual return!

---

---

---

---

---

---

---

---

Many other means exist - these include:

- the Harmonic (or Subcontrary) Mean
- the Quadratic Mean
- the Winsorized Mean
- the General Mean
- the Weighted Mean
- the Heronian Mean

Each of these measures of central tendency/location are appropriate under specific circumstances.

---

---

---

---

---

---

---

---

5. Median - value in the middle of the data array. Often denoted  $M_d$  for a population and  $m_d$  for a sample.

- if the data set has an odd number of observations, the median is the  $(n+1)/2^{\text{th}}$  (or middle) value of the data array
- if the data set has an even number of observations, the median is the mean value of the  $n/2^{\text{th}}$  and  $(n/2)+1^{\text{th}}$  (or middle two) values of the data array

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

middle two observations

the median is

$$m_d = \frac{18 + 19}{2} = \frac{37}{2} = 18.5$$

---

---

---

---

---

---

---

---

Extreme Value Elimination Method (an *easy* way to find the median) - systematically eliminate the most extreme values remaining in the data array until you are left with only one or two values - the mean of the remaining value(s) is the median

Example - for the data array that we have been working with

10 ~~11~~ ~~12~~ ~~14~~ ~~14~~ ~~16~~ ~~16~~ ~~16~~ ~~17~~ ~~18~~ ~~19~~ ~~21~~ ~~21~~ ~~24~~ ~~26~~ ~~28~~ ~~31~~ ~~31~~ ~~32~~ ~~36~~

and the median is 18.5.

Example - for the data array with an odd number of observations

~~14~~ ~~16~~ ~~16~~ ~~16~~ ~~17~~ ~~18~~ ~~19~~ ~~21~~ ~~21~~ ~~24~~ ~~26~~ ~~28~~ ~~31~~

and the median is 19.0.

---

---

---

---

---

---

---

---

6. Mode - most frequently occurring value(s) in the data array. Often denoted  $M_o$  for a population and  $m_o$  for a sample.

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

the mode is  $m_o = 16$ .

---

---

---

---

---

---

---

---

7. Percentile - the  $p^{\text{th}}$  percentile is the value that is at least as large as  $p$  percent of all observations in a data set and is no larger than  $(100 - p)$  percent of all observations in a data set.

To calculate the  $p^{\text{th}}$  percentile:

- create the data array (i.e., arrange the data in ascending order)
- compute an index  $i$

$$i = \left( \frac{P}{100} \right) n$$

- if  $i$  is not an integer, round up to the nearest integer. This is the position (in the data array) of the  $p^{\text{th}}$  percentile
- If  $i$  is an integer, the  $p^{\text{th}}$  percentile is the mean of the values occupying positions  $i$  and  $i + 1$  in the data array

---

---

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with, find the 15<sup>th</sup> percentile.

- create the data array (i.e., arrange the data in ascending order)

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

- compute an index  $i$

$$i = \left( \frac{15}{100} \right) 20 = 3$$

- $i$  is an integer ( $i=3$ ), so the 15<sup>th</sup> percentile is the mean of the values occupying positions  $i=3$  and  $i + 1=4$  in the data array

15% ≥ 15% of the data      85% ≥ (100-15)% = 85% of the data

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

13

or 13.0.

---

---

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with, find the 78<sup>th</sup> percentile.

- create the data array (i.e., arrange the data in ascending order)

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

- compute an index  $i$

$$i = \left( \frac{78}{100} \right) 20 = 15.6$$

- $i$  is not an integer ( $i=15.6$ ), so the 78<sup>th</sup> percentile is the value occupying the 16<sup>th</sup> position in the data array

75% ≥ 78% of the data      25% ≥ (100-78)% = 22% of the data

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

28

or 28.0.

---

---

---

---

---

---

---

---

---

---

Special percentiles include:

- the median or 50<sup>th</sup> percentile
- deciles or 10<sup>th</sup>, 20<sup>th</sup>, ..., 100<sup>th</sup> percentiles
- quintiles or 20<sup>th</sup>, 40<sup>th</sup>, 60<sup>th</sup>, 80<sup>th</sup>, 100<sup>th</sup> percentiles
- quartiles or 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, 100<sup>th</sup> percentiles (these are often denoted  $Q_1$ ,  $Q_2$ ,  $Q_3$ , and  $Q_4$ )

---

---

---

---

---

---

---

---

### C. Measures of Variability or Dispersion - Quantitative Data

1. Range - absolute difference between the minimum and maximum values in a data set

range = maximum value in a data set - minimum value in a data set

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

the range is

$$36 - 10 = 26$$

---

---

---

---

---

---

---

---

2. Interquartile Range (IQR) - absolute difference between the first and third quartiles in a data set, i.e.,

$$IQR = Q_3 - Q_1$$

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

the first and third quartiles are

$$Q_1 = 15.0 \text{ and } Q_3 = 27.0$$

so the interquartile range is

$$IQR = Q_3 - Q_1 = 27.0 - 15.0 = 12.0$$

---

---

---

---

---

---

---

---

3. Mean Absolute Deviation (MAD) - measure of relative dispersion for a data set based on the average distance that the observations in a data set lie from their mean. The MAD is calculated by

$$MAD = \frac{\sum_{i=1}^N |x_i - \mu|}{N}$$

for a population and by

$$mad = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

for a sample.

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

for which we have already calculated the sample mean to be 20.65, the MAD is

$$mad = \frac{|10 - 20.65| + \dots + |36 - 20.65|}{20} = \frac{128.3}{20} = 6.415$$

---

---

---

---

---

---

---

---

4. Variance - measure of relative dispersion based on the squared distance that the observations in a data set lie from their mean. The variance is calculated by

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

for a population and by

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

for a sample.

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

for which we have already calculated the sample mean to be 20.65, the variance is

$$s^2 = \frac{(10 - 20.65)^2 + \dots + (36 - 20.65)^2}{20 - 1} = 59.503$$

---

---

---

---

---

---

---

---

5. Standard Deviation - measure of relative dispersion that is equal to the positive square root of the variance. The standard deviation is calculated by

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

for a population and by

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

for a sample.

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

for which we have already calculated the sample mean to be 20.65, the standard deviation is

$$s = \sqrt{s^2} = \sqrt{59.503} = 7.714$$

---

---

---

---

---

---

---

---

6. Coefficient of Variation - measure of relative dispersion that standardized in relation to its mean. The coefficient of variation is calculated by

$$CV = \left( \frac{\sigma}{\mu} \right) * 100$$

for a population and by

$$cv = \left( \frac{s}{\bar{x}} \right) * 100$$

for a sample.

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

the coefficient of variation is

$$cv = \left( \frac{7.714}{20.65} \right) * 100 = 37.355$$

---

---

---

---

---

---

---

---

#### D. Using Measures of Relative Location to Identify Outliers

1. Outlier - an observation associated with an unusually extreme (either small or large) value of a variable
2. z-Score - number of standard deviations an observation ( $x_i$ ) lies from the mean. Often referred to as the standardized value, it is calculated by

$$z_i = \frac{x_i - \bar{x}}{s}$$

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

the z-score for the value  $x_3 = 12$  is

$$z_3 = \frac{12 - 20.65}{7.714} = -1.12$$

Note that the z-score can be interpreted as the number of standard deviations the observation  $x_3 = 12$  lies from its mean (i.e.,  $x_3$  lies  $z_3 = -1.12$  standard deviations from its mean of 20.65)

---

---

---

---

---

---

---

---

z-Scores have some special properties. They include

Chebyshev's Theorem - at least

$$1 - \frac{1}{z^2}$$

of the observations in any data set will be within  $z$  standard deviations of the mean, where  $z \geq 1$ . Thus we have that

- at least 75% of all observations in a data set must be within  $z = 2$  standard deviations of the mean
- at least 89% of all observations in a data set must be within  $z = 3$  standard deviations of the mean
- at least 94% of all observations in a data set must be within  $z = 4$  standard deviations of the mean

---

---

---

---

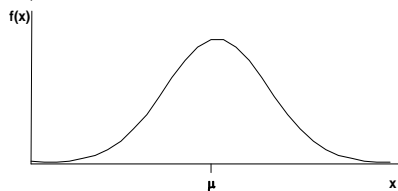
---

---

---

---

The Empirical Rule - for data with a bell-shaped (normal) distribution,



- approximately 68% (68.26%) of all observations in a data set are within  $z = 1$  standard deviation of the mean
- approximately 95% (95.44%) of all observations in a data set are within  $z = 2$  standard deviations of the mean
- over 99% (99.72%) of all observations in a data set are within  $z = 3$  standard deviations of the mean

---

---

---

---

---

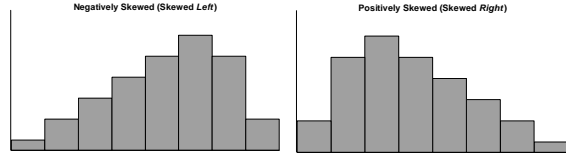
---

---

---

## E. Other Characteristics of Data Distribution Shapes

### 1. Skewness - degree to which a data distribution is asymmetric



By skewed left, we mean that the left tail is longer than the right tail. Similarly, skewed right means that the right tail is longer than the left tail.

Note that  $\mu > M_d$  for a positively skewed population and  $\mu < M_d$  for a negatively skewed population (why?).

---

---

---

---

---

---

---

---

Skewness is commonly defined as:

$$SK = \frac{\sum_{i=1}^N (x_i - \mu)^3}{N\sigma^3}$$

for a population and

$$sk = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

for a sample.

Although many different formulas for calculating skewness exist

- this sample formula is used by Excel
- all variations rely on the cubed distance from the mean

---

---

---

---

---

---

---

---

Note that:

- The sign indicates the direction of skewness in the population
  - it will be positive if the population is positively skewed
  - negative if the population is negatively skewed
  - close to 0 if the population is symmetric
- A general guideline for Excel's skewness measure is that the distribution is approximately symmetric if the value is between -1 and +1.

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

Excel would calculate the skewness to be

$$sk = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$
$$= \frac{20}{(20-1)(20-2)} \frac{(10-20.65)^3 + (11-20.65)^3 + \dots + (36-20.65)^3}{7.714^3} = 0.5312$$

Thus, this skewness coefficient suggests the sample data are relatively symmetric (or slightly right-skewed).

---

---

---

---

---

---

---

---

Why does cubing the distances of the observations in the population from their mean  $\mu$  provide a measure of skewness?

- cubed distances retain their direction (sign)
- large distances (either negative or positive) increase dramatically in magnitude when cubed - these correspond to observations far out in the tails
- small distances (either negative or positive) experience a less dramatic increase (or possibly even a decrease) in magnitude when cubed - these correspond to observations near the center of the distribution

---

---

---

---

---

---

---

---

Pearson suggested a less complex measure of skewness that takes advantage of the relationship between the population mean  $\mu$  and population median  $M_d$  in skewed populations ( $\mu > M_d$  for a positively skewed population and  $\mu < M_d$  for a negatively skewed population). *Pearson's Second Coefficient of Skewness* is

$$SK = 3 \left( \frac{\mu - M_d}{\sigma} \right)$$

for a population and

$$sk = 3 \left( \frac{\bar{x} - m_d}{s} \right)$$

for a sample.

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

Pearson's Second Coefficient of Skewness is

$$sk = 3 \frac{\bar{x} - m_d}{s} = 3 \frac{20.65 - 18.50}{7.714} = 3(0.2787) = 0.8361$$

The positive value of Pearson's Second Coefficient of Skewness suggests the sample data are right-skewed.

Note that the disadvantage of Pearson's Second Coefficient of Skewness is that it does not consider all observations in the population or sample.

---

---

---

---

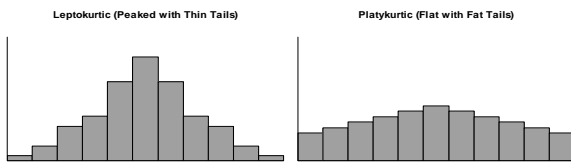
---

---

---

---

2. Kurtosis - degree of relative peakedness of a data distribution (and also a measure of the heaviness of the tails of a distribution).



- Leptokurtic - relatively peaked with thin tails
- Platykurtic - relatively flat with fat tails
- Mesokurtic - relatively smooth

---

---

---

---

---

---

---

---

Kurtosis is commonly defined as:

$$KUR = \frac{\sum_{i=1}^N (x_i - \mu)^4}{N\sigma^4}$$

for a population and

$$kur = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$$

for a sample.

---

---

---

---

---

---

---

---

Because these definitions of kurtosis yield a value of

$$\frac{3(n-1)^2}{(n-2)(n-3)}$$

for a *normal distribution*, the leading to the following adjusted formula

$$kur = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

for the kurtosis of a sample (which will have a value of zero for a normal distribution).

Many different formulas for calculating kurtosis exist, but

- this sample formula is used by Excel
- all variations rely on the quartic distance from the mean

---

---

---

---

---

---

---

---

---

---

For the adjusted measure of kurtosis relative to a *normal distribution*

- a symmetric distribution with positive (lepto) kurtosis suggests the distribution has less area in the tails and a sharper peak than that of a normal distribution
- a symmetric distribution with negative (platy) kurtosis suggests the distribution has more area in the tails and a flatter peak than that of a normal distribution
- a symmetric distribution with near-zero (meso) kurtosis suggests the distribution has area in the tails and a peak that are similar to that of a normal distribution

---

---

---

---

---

---

---

---

---

---

Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

Excel would calculate the (adjusted) kurtosis to be

$$kur = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$= \frac{20(20+1)}{(20-1)(20-2)(20-3)} \frac{(10-20.65)^4 + (11-20.65)^4 + L + (36-20.65)^4}{7.714^4} - \frac{3(20-1)^2}{(20-2)(20-3)}$$

$$= (0.0722)(37.17212446) - 3.539215686 = -0.8539$$

Thus, this kurtosis coefficient suggests the sample data are relatively symmetric (or perhaps slightly right-skewed).

---

---

---

---

---

---

---

---

---

---

Why does taking the fourth power of distances of the observations in the population from their mean  $\mu$  provide a measure of kurtosis?

- quartic distances are directionless (lose their sign)
- large distances (either negative or positive) increase dramatically in magnitude when taken to the fourth power - these values correspond to observations far out in the tails
- small distances (either negative or positive) experience a less dramatic increase (or possibly even a decrease) in magnitude when taken to the fourth power - these correspond to observations near the center of the distribution

Note that neither measures of skewness or kurtosis are commonly used (*but you should understand them*).

---

---

---

---

---

---

---

---

## F. Other Tools for Exploratory Data Analysis (EDA)

1. Five Number Summary - use the minimum, first quartile ( $Q_1$ ), median ( $Q_2$ ), third quartile ( $Q_3$ ), and maximum to summarize a data set.
2. Box Plot - Graphical display of the results of a five number summary and outliers. *One possible* set of steps to construct a Box Plot from a data array and its five numbers are:

---

---

---

---

---

---

---

---

- create a horizontal axis of an appropriate scale as the basis of the box plot
- draw a box with vertical ends located at the first quartile ( $Q_1$ ) and third quartile ( $Q_3$ )
- draw a vertical line through the box at the median ( $Q_2$ )
- develop the inner fences -  $Q_1 - 1.5(IQR)$  and  $Q_3 + 1.5(IQR)$  and draw vertical lines from the ends of the box to the smallest and largest values inside the fences
- classify all observations outside of the inner fences (i.e.  $< Q_1 - 1.5(IQR)$  or  $> Q_3 + 1.5(IQR)$ ) as outliers. If any outliers exist, indicate their locations (using some symbol such as a star or asterisk) on the box plot

---

---

---

---

---

---

---

---

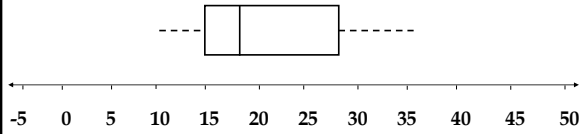
Example - for the data array that we have been working with

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 36

we have that

$Q_1 = 15.0, Q_2 = 18.5, Q_3 = 27.0,$  and  $IQR = 12.0$

so the box plot is




---

---

---

---

---

---

---

---

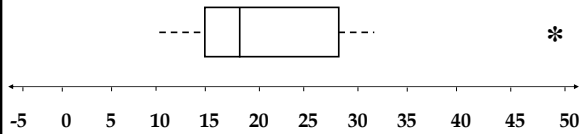
Example - what if the last data value were 49 (instead of 36)? The data array would look like this

10 11 12 14 14 16 16 16 17 18 19 21 21 24 26 28 31 31 32 49

we would still have that

$Q_1 = 15.0, Q_2 = 18.5, Q_3 = 27.0,$  and  $IQR = 12.0$

(why?), but the new box plot would be




---

---

---

---

---

---

---

---

## G. Measures of Association Between Two Variables

1. Covariance - numerical measure of linear association between two quantitative variables. It is calculated as

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

for a population or

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

for a sample.

---

---

---

---

---

---

---

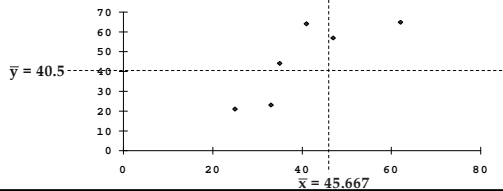
---

When will  $\sigma_{xy}$  or  $s_{xy}$  be negative? Positive? Zero?

Hint - look at the formulas!

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N} \quad \text{and} \quad s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Another hint - consider the scatterplot!




---

---

---

---

---

---

---

---

Example - for the data we collected and displayed on a scatter diagram, the covariance between age and income would be calculated as

AGE (x)	INCOME (y)	$x_i - 40.50$	$y_i - 45.67$	$(x_i - 40.50) * (y_i - 45.67)$
25	21	-15.500	-24.667	382.333
47	57	6.500	11.333	73.667
35	44	-5.500	-1.667	9.167
62	65	21.500	19.333	415.667
41	64	0.500	18.333	9.167
33	23	-7.500	-22.667	170.000
so we have that				1060.000

$$s_{xy} = \frac{1060.00}{6-1} = 212.00$$

---

---

---

---

---

---

---

---

2. Correlation - standardized numerical measure of linear association between two variables. Range is generally -1 to 1.

3. Pearson's Product Moment Correlation Coefficient - standardized numerical measure of linear association between two quantitative variables. Range is -1 to 1. It is calculated as

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

for a population or

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

for a sample.

---

---

---

---

---

---

---

---

When will  $\rho_{xy}$  or  $r_{xy}$  be negative? Positive? Zero?  
 Hint - look at the formulas!

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{and} \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

---

---

---

---

---

---

---

---

Example - for the data we collected and displayed on a scatter diagram, we can calculate

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(25-40.50)^2 + (47-40.50)^2 + (35-40.50)^2 + (62-40.50)^2 + (41-40.50)^2 + (33-40.50)^2}{6-1}} = 12.90$$

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{(21-45.67)^2 + (57-45.67)^2 + (44-45.67)^2 + (65-45.67)^2 + (64-45.67)^2 + (23-45.67)^2}{6-1}} = 19.82$$

Additionally, we have already calculated

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = 212.00$$

so the correlation between age and income would be calculated as

$$r_{xy} = \frac{212.00}{12.90 * 19.82} = 0.829$$

---

---

---

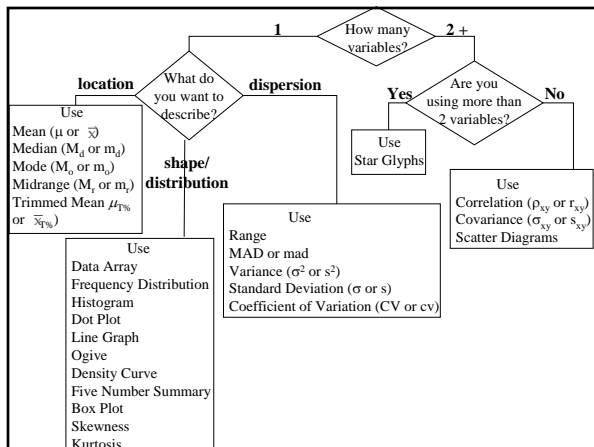
---

---

---

---

---




---

---

---

---

---

---

---

---