Chapter Four

Numerical Descriptive Techniques

Numerical Descriptive Techniques…

Measures of Central Location

Mean, Median, Mode

Measures of Variability

Range, Standard Deviation, Variance, Coefficient of Variation

Measures of Relative Standing

Percentiles, Quartiles

Measures of Linear Relationship

Covariance, Correlation, Determination, Least Squares Line

Measures of Central Location...

The ***arithmetic mean***, a.k.a. *average*, shortened to *mean*, is the most popular & useful measure of central location.

It is computed by simply adding up all the observations and dividing by the total number of observations:

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{Number of observations}}$$

Notation...

When referring to the number of observations in a ***population***, we use uppercase letter **N**

When referring to the number of observations in a ***sample***, we use lower case letter **n**

The arithmetic mean for a ***population*** is denoted with Greek letter "mu": μ

The arithmetic mean for a ***sample*** is denoted with an "x-bar": $\overline{x}$

Arithmetic Mean...

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

Population Mean

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

Sample Mean

The Arithmetic Mean...

…is appropriate for describing measurement data, e.g. heights of people, marks of student papers, etc.

…is seriously affected by extreme values called “outliers”. E.g. as soon as a billionaire moves into a neighborhood, the average household income increases beyond what it was previously!

Measures of Central Location…

The ***median*** is calculated by placing all the observations in order; the observation that falls in the *middle* is the median.

Data: {0, 7, 12, 5, 14, 8, 0, 9, 22} N=9 (odd)
Sort them bottom to top, find the middle:
0 0 5 7 **8** 9 12 14 22

Data: {0, 7, 12, 5, 14, 8, 0, 9, 22, 33} N=10 (even)

Sort them bottom to top, the middle is the simple average between 8 & 9:
0 0 5 7 **8 9** 12 14 22 33
median = (8+9)÷2 = **8.5**

Sample and population medians are computed the same way.

Measures of Central Location…

The ***mode*** of a set of observations is the value that occurs most *frequently*.

A set of data may have one mode (or modal class), or two, or more modes.

Mode is a useful for all data types, though mainly used for nominal data.

For large data sets the modal *class* is much more relevant than a single-value mode.

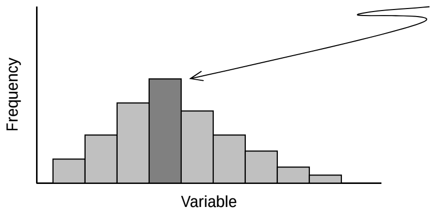
Sample and population modes are computed the same way.

Mode...

E.g. Data: {0, 7, 12, 5, 14, 8, 0, 9, 22, 33} N=10

Which observation appears most often?

The mode for this data set is **0**. How is this a measure of “central” location?



A modal class

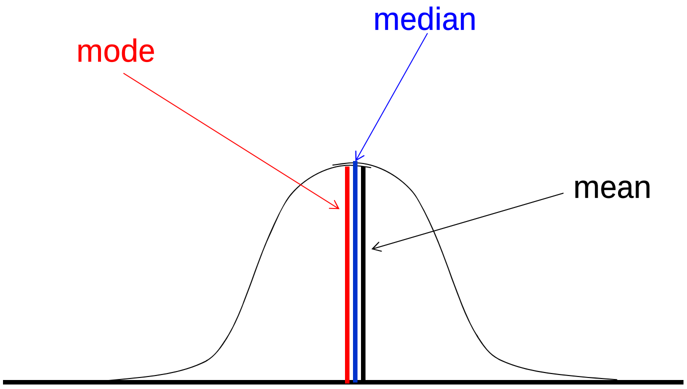
`=MODE(range)` in Excel...

Note: if you are using Excel for your data analysis and your data is multi-modal (i.e. there is more than one mode), Excel only calculates the smallest one.

You will have to use other techniques (i.e. histogram) to determine if your data is bimodal, trimodal, etc.

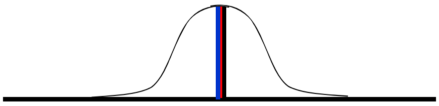
Mean, Median, Mode...

If a distribution is symmetrical,
the mean, median and mode may coincide…



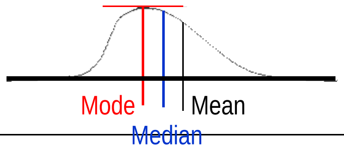
Mean, Median, Mode...

- If a distribution is symmetrical, the mean, median and mode coincide

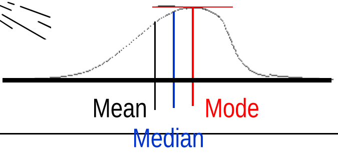


- If a distribution is non symmetrical, and skewed to the left or to the right, the three measures differ.

A positively skewed distribution
("skewed to the right")

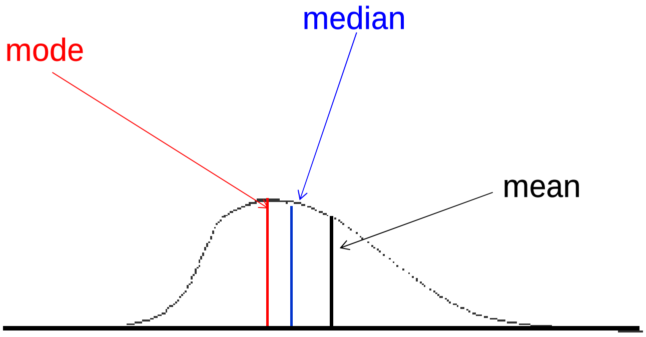


A negatively skewed distribution
("skewed to the left")



Mean, Median, Mode...

If a distribution is asymmetrical, say skewed to the left or to the right, the three measures may differ. E.g.:



Mean, Median, Mode: Which Is Best?

With three measures from which to choose, which one should we use?

The mean is generally our first selection. However, there are several circumstances when the median is better.

The mode is seldom the best measure of central location.

One advantage the median holds is that it not as sensitive to extreme values as is the mean.

Mean, Median, Mode: Which Is Best?

To illustrate, consider the data in Example 4.1.

The mean was 11.0 and the median was 8.5.

Now suppose that the respondent who reported 33 hours actually reported 133 hours (obviously an Internet addict). The mean becomes

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+7+12+5+133+14+8+0+22}{10} = \frac{210}{10} = 21.0$$

Mean, Median, Mode: Which Is Best?

This value is only exceeded by only two of the ten observations in the sample, making this statistic a poor measure of *central* location.

The median stays the same. When there is a relatively small number of extreme observations (either very small or very large, but not both), the median usually produces a better measure of the center of the data.

Mean, Median, & Modes for Ordinal & Nominal Data

For ordinal and nominal data the calculation of the mean is not valid.

Median is appropriate for ordinal data.

For nominal data, a mode calculation is useful for determining highest frequency but not “central location”.

Measures of Central Location • Summary…

Compute the Mean to

- Describe the central location of a single set of interval data

Compute the Median to

- Describe the central location of a single set of interval or ordinal data

Compute the Mode to

- Describe a single set of nominal data

Geometric Mean

The arithmetic mean is the single most popular and useful measure of central location.

However, there is another circumstance where neither the mean nor the median is the best measure.

When the variable is a growth rate or rate of change, such as the value of an investment over periods of time, we need another measure.

This will become apparent from the following illustration.

Geometric Mean

Suppose you make a 2-year investment of $1,000 and it grows by 100% to $2,000 during the first year.

During the second year, however, the investment suffers a 50% loss, from $2,000 back to $1,000.

The rates of return for years 1 and 2 are R_1 = 100% and R_2 = –50%, respectively. The arithmetic mean (and the median) is computed as

$$\overline{R} = \frac{R_1 + R_2}{2} = \frac{100 + (-50)}{2} = 25\%$$

Geometric Mean

But this figure is misleading. Because there was no change in the value of the investment from the beginning to the end of the 2-year period, the “average” compounded rate of return is 0%.

As you will see, this is the value of the *geometric mean*.

Geometric Mean

Let R_i denote the rate of return (in decimal form) in period i (i = 1, 2, …, n). The **geometric mean** R_g of the returns is defined such that

$$(1+R_g)^n = (1+R_1)(1+R_2)...(1+R_n)$$

Solving for R_g we produce the following formula.

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)...(1+R_n)} - 1$$

Geometric Mean

The geometric mean of our investment illustration is

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)...(1+R_n)} - 1$$

$$= \sqrt[2]{(1+1)(1+[-.50])} - 1 = 1 - 1 = 0$$

The geometric mean is therefore 0%. This is the single "average" return that allows us to compute the value of the investment at the end of the investment period from the beginning value. Thus, using the formula for compound interest with the rate = 0%, we find

Value at the end of the investment period = 1,000$(1 + R_g)^2$ = 1,000$(1 + 0)^2$ = 1,000

Geometric Mean

Thus, using the formula for compound interest with the rate = 0%, we find

Value at the end of the investment period
= 1,000$(1 + R_g)^2$ = 1,000$(1 + 0)^2$ = 1,000

Geometric Mean

The geometric mean is used whenever we wish to find the "average" growth rate, or rate of change, in a variable *over time*.

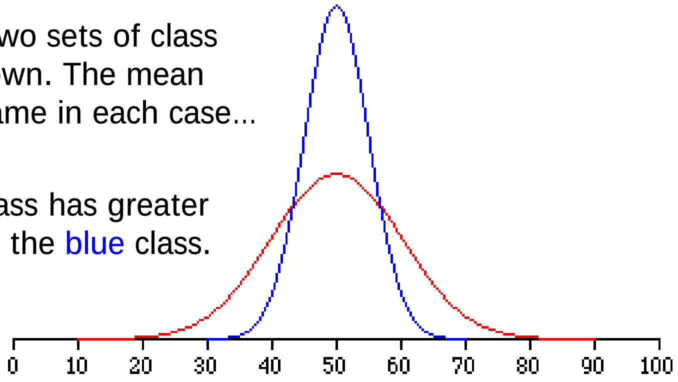
However, the arithmetic mean of n returns (or growth rates) is the appropriate mean to calculate if you wish to estimate the mean rate of return (or growth rate) for any *single* period in the future.

Measures of Variability…

Measures of central location fail to tell the whole story about the distribution; that is, how much are the observations spread out around the mean value?

For example, two sets of class grades are shown. The mean (=50) is the same in each case…

But, the red class has greater variability than the blue class.



Range...

The ***range*** is the simplest measure of variability, calculated as:

Range = Largest observation – Smallest observation

E.g.

Data: {4, 4, 4, 4, 50} Range = 46

Data: {4, 8, 15, 24, 39, 50} Range = 46

The range is the same in both cases,

but the data sets have very different distributions…

Range...

Its major advantage is the ease with which it can be computed.

Its major shortcoming is its failure to provide information on the dispersion of the observations between the two end points.

Hence we need a measure of variability that incorporates **all the data** and not just two observations. Hence…

Variance...

Variance and its related measure, standard deviation, are arguably the most important statistics. Used to measure variability, they also play a vital role in almost all statistical inference procedures.

Population variance is denoted by σ^2

(Lower case Greek letter “sigma” squared)

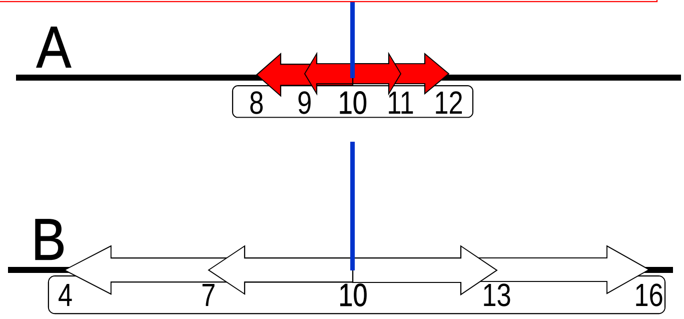
Sample variance is denoted by s^2

(Lower case “S” squared)

Why not use the sum of deviations?

Consider two small populations:

The sum of deviations is zero for both populations, therefore, is not a good measure of dispersion.



9-10= -1

11-10= +1

8-10= -2

12-10= +2

Sum = 0

4-10 = - 6

16-10 = +6

7-10 = -3

13-10 = +3

Sum = 0

Variance

Let us calculate the variance of the two populations

$$\sigma_A^2 = \frac{(8-10)^2+(9-10)^2+(10-10)^2+(11-10)^2+(12-10)^2}{5} = 2$$

$$\sigma_B^2 = \frac{(4-10)^2+(7-10)^2+(10-10)^2+(13-10)^2+(16-10)^2}{5} = 18$$

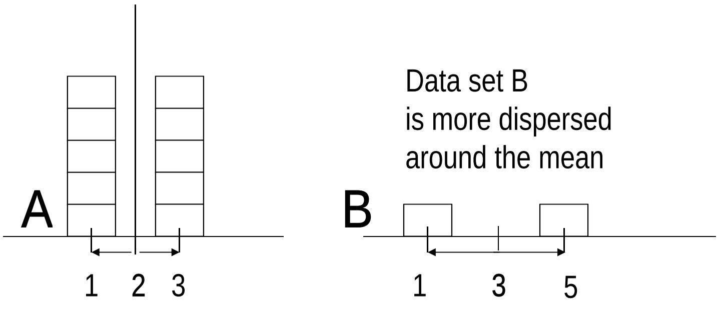
Why is the variance defined as the average squared deviation? Why not use the sum of squared deviations as a measure of variation instead?

After all, the sum of squared deviations increases in magnitude when the variation of a data set increases!!

Variance

Which data set has a larger dispersion?

Let us calculate the sum of squared deviations for both data sets

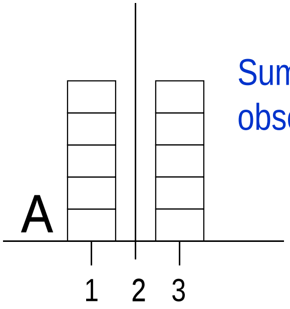


Data set B
is more dispersed
around the mean

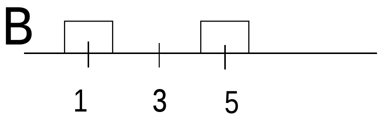
Variance

$\text{Sum}_A = (1-2)^2 + \ldots + (1-2)^2 + (3-2)^2 + \ldots + (3-2)^2 = 10$

$\text{Sum}_B = (1-3)^2 + (5-3)^2 = 8$

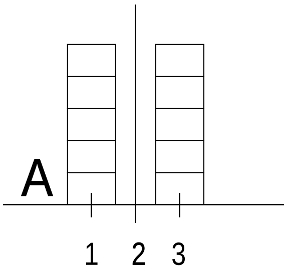


$\text{Sum}_A > \text{Sum}_B$. This is inconsistent with the observation that set B is more dispersed.



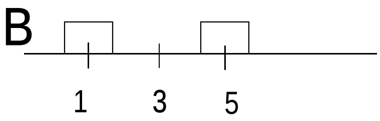
Variance

However, when calculated on “per observation” basis (variance), the data set dispersions are properly ranked.



$\sigma_A^2 = \text{Sum}_A/N = 10/5 = 2$

$\sigma_B^2 = \text{Sum}_B/N = 8/2 = 4$



Variance...

population mean

The variance of a **population** is: $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

population size

sample mean

The variance of a **sample** is: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Note: The denominator is sample size (n) minus one !

Variance...

As you can see, you have to calculate the sample mean (x-bar) in order to calculate the sample variance.

Alternatively, there is a short-cut formulation to calculate sample variance directly from the data without the intermediate step of calculating the mean. Its given by:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Application...

Example 4.7. The following sample consists of the number of jobs six students applied for: 17, 15, 23, 7, 9, 13.

Finds its mean and variance.

What are we looking to calculate?

The following **sample** consists of the number of jobs six students applied for: 17, 15, 23, 7, 9, 13.

Finds its **mean** and **variance**.

$\bar{x}$ s^2

…as opposed to μ or σ^2

Sample Mean & Variance...

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14 \; jobs$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots(13-14)^2\right] = 33.2$$

Sample Variance (shortcut method)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2 + 15^2 + \ldots + 13^2\right) - \frac{\left(17+15+\ldots+13\right)^2}{6}\right] = 33.2$$

Standard Deviation...

The standard deviation is simply the square root of the variance, thus:

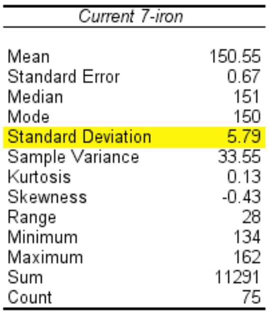
Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

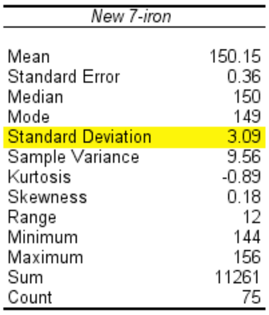
Standard Deviation...

Consider Example 4.8 [Xm04-08]where a golf club manufacturer has designed a new club and wants to determine if it is hit more consistently (i.e. with less variability) than with an old club.

Using Data > Data Analysis > Descriptive Statistics in Excel, we produce the following tables for interpretation…



Current 7-iron	
Mean	150.55
Standard Error	0.67
Median	151
Mode	150
Standard Deviation	5.79
Sample Variance	33.55
Kurtosis	0.13
Skewness	-0.43
Range	28
Minimum	134
Maximum	162
Sum	11291
Count	75



New 7-iron	
Mean	150.15
Standard Error	0.36
Median	150
Mode	149
Standard Deviation	3.09
Sample Variance	9.56
Kurtosis	-0.89
Skewness	0.18
Range	12
Minimum	144
Maximum	156
Sum	11261
Count	75

You get more consistent distance with the new club.

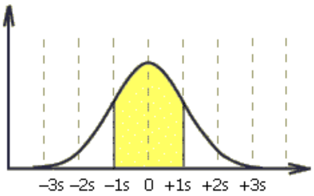
Interpreting Standard Deviation...

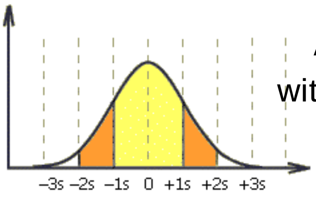
The standard deviation can be used to compare the variability of several distributions and make a statement about the general shape of a distribution. If the histogram is **bell shaped**, we can use the ***Empirical Rule***, which states:

1) Approximately 68% of all observations fall within one standard deviation of the mean.
2) Approximately 95% of all observations fall within two standard deviations of the mean.
3) Approximately 99.7% of all observations fall within three standard deviations of the mean.

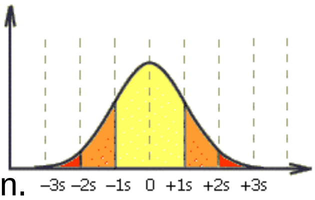
The Empirical Rule...

Approximately 68% of all observations fall within **one** standard deviation of the mean.





Approximately 95% of all observations fall within **two** standard deviations of the mean.



Approximately 99.7% of all observations fall within **three** standard deviations of the mean.

Chebysheff's Theorem…

A more general interpretation of the standard deviation is derived from ***Chebysheff's Theorem***, which applies to all shapes of histograms (not just bell shaped).

The proportion of observations in any sample that lie within **k** standard deviations of the mean is *at least:*

$$1-\frac{1}{k^2}\quad for\quad k>1$$

For k=2 (say), the theorem states that *at least* 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

Interpreting Standard Deviation

Suppose that the mean and standard deviation of last year's midterm test marks are 70 and 5, respectively. If the histogram is bell-shaped then we know that approximately 68% of the marks fell between 65 and 75, approximately 95% of the marks fell between 60 and 80, and approximately 99.7% of the marks fell between 55 and 85.

If the histogram is not at all bell-shaped we can say that at least 75% of the marks fell between 60 and 80, and at least 88.9% of the marks fell between 55 and 85. (We can use other values of k.)

Coefficient of Variation...

The ***coefficient of variation*** of a set of observations is the standard deviation of the observations divided by their mean, that is:

Population coefficient of variation = CV = $\dfrac{\sigma}{\mu}$

Sample coefficient of variation = cv = $\dfrac{s}{\bar{x}}$

Coefficient of Variation...

This coefficient provides a ***proportionate*** measure of variation, e.g.

A standard deviation of 10 may be perceived as large when the mean value is 100, but only moderately large when the mean value is 500.

Measures of Relative Standing & Box Plots

Measures of relative standing are designed to provide information about the ***position*** of particular values ***relative*** to the entire data set.

Percentile: the P^{th} percentile is the value for which P percent are *less than* that value and (100-P)% are greater than that value.

Suppose you scored in the 60th percentile on the GMAT, that means 60% of the other scores were *below* yours, while 40% of scores were *above* yours.

Quartiles...

We have special names for the 25th, 50th, and 75th percentiles, namely ***quartiles***.

The first or lower quartile is labeled Q_1 = 25th percentile.

The second quartile, Q_2 = 50th percentile (which is also the median).

The third or upper quartile, Q_3 = 75th percentile.

We can also convert percentiles into quintiles (fifths) and deciles (tenths).

Commonly Used Percentiles…

First (lower) decile	= 10th percentile
First (lower) quartile, Q_1,	= 25th percentile
Second (middle)quartile,Q_2,	= 50th percentile
Third quartile, Q_3,	= 75th percentile
Ninth (upper) decile	= 90th percentile

Note: If your exam mark places you in the 80th percentile, that doesn't mean you scored 80% on the exam – it means that 80% of your peers scored **lower** than you on the exam; It is about your position relative to others.

Location of Percentiles…

The following formula allows us to approximate the location of any percentile:

$$L_P = (n+1)\frac{P}{100}$$

where L_P is the location of the P^{th} percentile

Location of Percentiles…

Recall the data from Example 4.1:

0 0 5 7 8 9 12 14 22 33

Where is the location of the 25th percentile? That is, at which point are 25% of the values lower and 75% of the values higher?

0 0 5 7 8 9 12 14 22 33

$$L_{25} = (10+1)(25/100) = \boxed{2.75}$$

The 25th percentile is three-quarters of the distance between the second (which is 0) and the third (which is 5) observations. Three-quarters of the distance is: (.75)(5 – 0) = 3.75

Because the second observation is 0, the 25th percentile is 0 + 3.75 = **3.75**

Location of Percentiles…

What about the upper quartile?

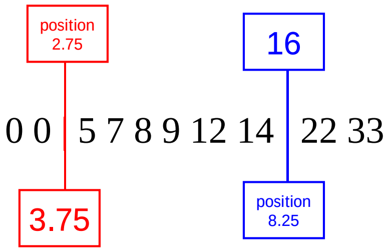
$$L_{75} = (10+1)(75/100) = \boxed{8.25}$$

0 0 5 7 8 9 12 14 22 33

It is located one-quarter of the distance between the eighth and the ninth observations, which are 14 and 22, respectively. One-quarter of the distance is: (.25)(22 - 14) = 2, which means the 75th percentile is at: 14 + 2 = **16**

Location of Percentiles…

Please remember…



L_p determines the **position** in the data set where the percentile value lies, not the value of the percentile itself.

Interquartile Range…

The quartiles can be used to create another measure of variability, the ***interquartile range***, which is defined as follows:

$$\text{Interquartile Range} = Q_3 - Q_1$$

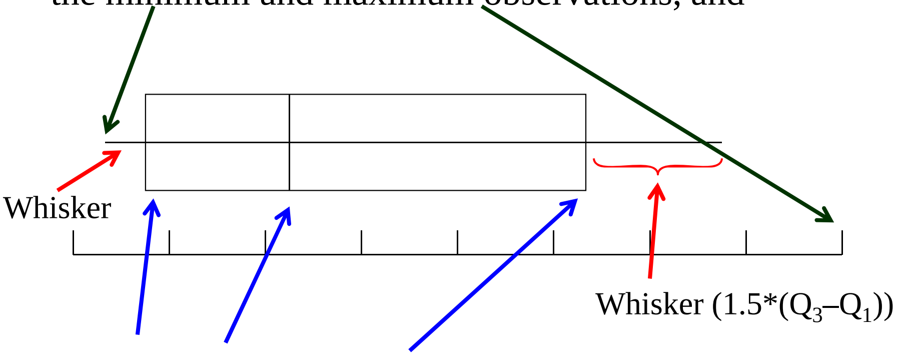
The interquartile range measures the spread of the middle 50% of the observations.

Large values of this statistic mean that the 1st and 3rd quartiles are far apart indicating a high level of variability.

Box Plots...

The ***box plot*** is a technique that graphs **five** statistics:

- the minimum and maximum observations, and



- the first, second, and third quartiles.

The lines extending to the left and right are called whiskers. Any points that lie outside the whiskers are called outliers. The whiskers extend outward to the smaller of 1.5 times the interquartile range or to the most extreme point that is not an outlier.

Example 4.15

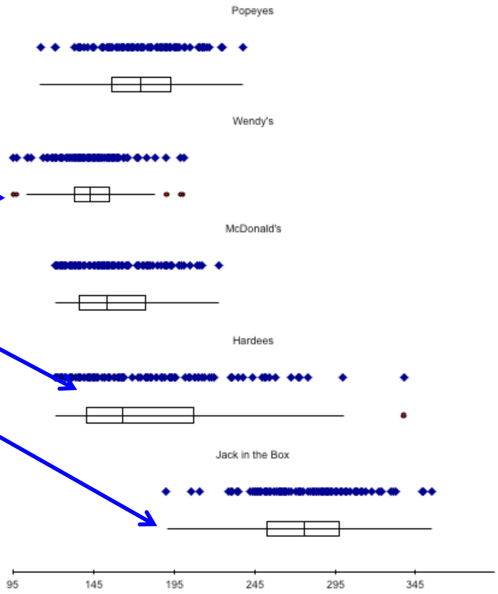
A large number of fast-food restaurants with drive-through windows offering drivers and their passengers the advantages of quick service. To measure how good the service is, an organization called QSR planned a study wherein the amount of time taken by a sample of drive-through customers at each of five restaurants was recorded. Compare the five sets of data using a box plot and interpret the results.

Box Plots...

These box plots are based on data in Xm04-15.

Wendy's service time is shortest and least variable.

Hardee's has the greatest variability, while Jack-in-the-Box has the longest service times.



Measures of Linear Relationship...

We now present three numerical measures of linear relationship that provide information as to the **strength & direction** of a linear relationship between two variables (if one exists).

They are the ***covariance,*** the ***coefficient of correlation,*** and the ***coefficient of determination***.

Covariance...

population mean of variable X, variable Y

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

sample mean of variable X, variable Y

$$\text{Sample covariance} = \mathrm{s}_{xy} = \frac{\sum_{i=1}^{n}(\mathrm{x_i} - \bar{x})(y_i - \bar{y})}{\mathrm{n - 1}}$$

Note: divisor is n-1, not n as you may expect.

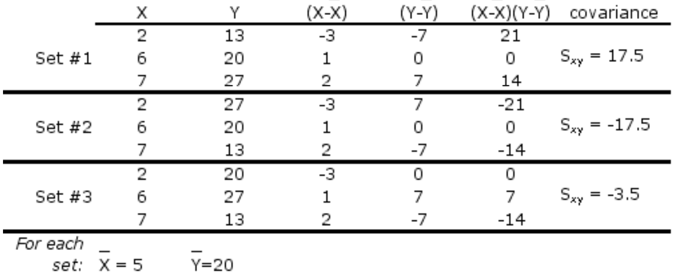
Covariance...

In much the same way there was a “shortcut” for calculating sample variance without having to calculate the sample mean, there is also a shortcut for calculating sample covariance without having to first calculate the means:

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

Covariance Illustrated...

Consider the following three sets of data (textbook §4.5)…



	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	covariance
	2	13	-3	-7	21	
Set #1	6	20	1	0	0	S_{xy} = 17.5
	7	27	2	7	14	
	2	27	-3	7	-21	
Set #2	6	20	1	0	0	S_{xy} = -17.5
	7	13	2	-7	-14	
	2	20	-3	0	0	
Set #3	6	27	1	7	7	S_{xy} = -3.5
	7	13	2	-7	-14	

For each set: $\bar{X} = 5$ $\bar{Y}=20$

In each set, the values of X are the same, and the value for Y are the same; the only thing that's changed is the order of the Y's.

In set #1, as X increases so does Y; S_{xy} is large & positive

In set #2, as X increases, Y decreases; S_{xy} is large & negative

In set #3, as X increases, Y doesn't move in any particular way; S_{xy} is "small"

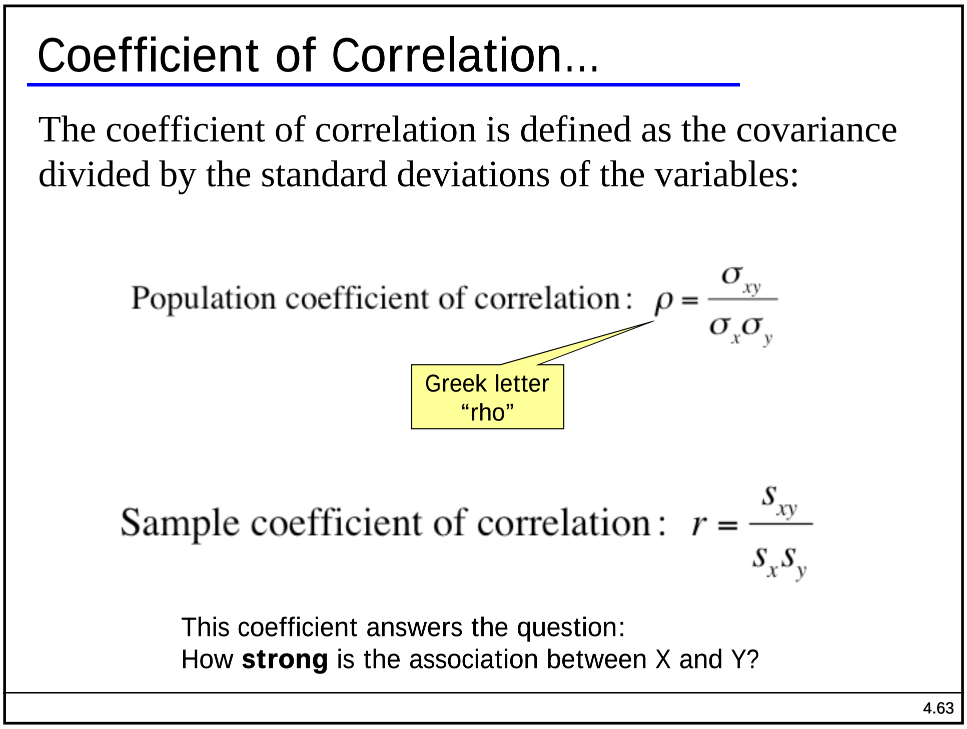
Covariance... (Generally speaking)

When two variables move in the ***same direction*** (both increase or both decrease), the covariance will be a ***large positive number***.

When two variables move in ***opposite directions***, the covariance is a ***large negative number***.

When there is ***no particular pattern***, the covariance is a ***small number***.

However, it is often difficult to determine whether a particular covariance is large or small. The next parameter/statistic addresses this problem.



Coefficient of Correlation…

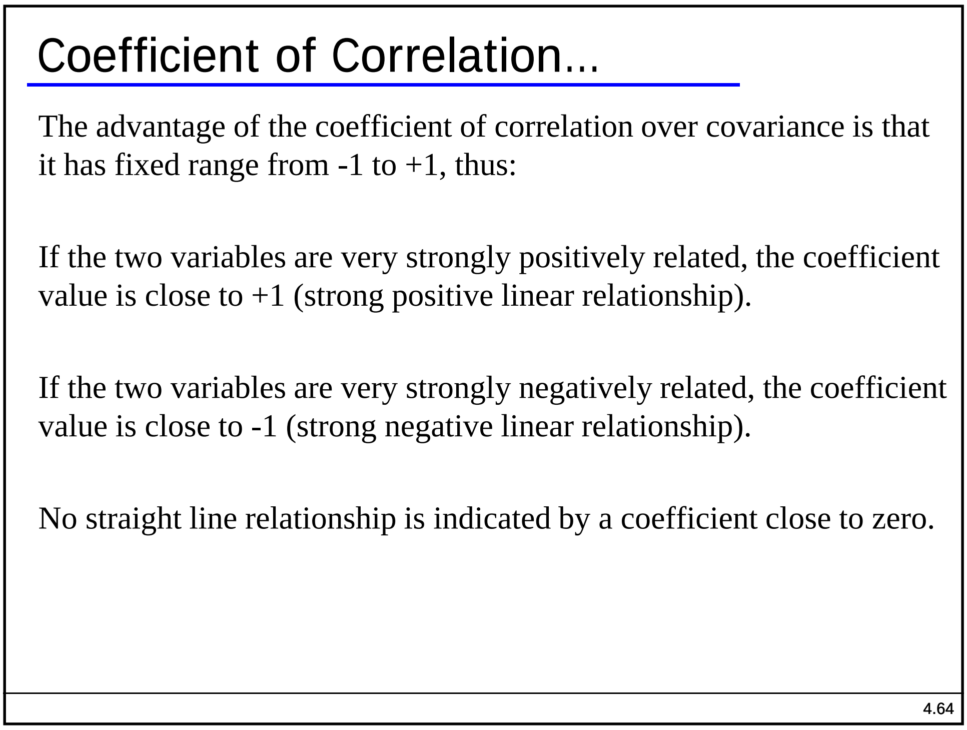
The coefficient of correlation is defined as the covariance divided by the standard deviations of the variables:

$$\text{Population coefficient of correlation: } \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Greek letter "rho"

$$\text{Sample coefficient of correlation: } r = \frac{s_{xy}}{s_x s_y}$$

This coefficient answers the question:
How **strong** is the association between X and Y?



Coefficient of Correlation…

The advantage of the coefficient of correlation over covariance is that it has fixed range from -1 to +1, thus:

If the two variables are very strongly positively related, the coefficient value is close to +1 (strong positive linear relationship).

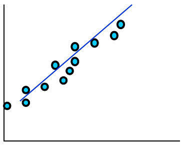
If the two variables are very strongly negatively related, the coefficient value is close to -1 (strong negative linear relationship).

No straight line relationship is indicated by a coefficient close to zero.

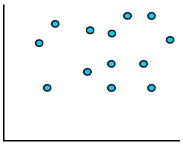
Coefficient of Correlation…

ρ or r =

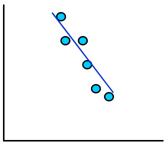
+1 Strong positive linear relationship



0 No linear relationship



-1 Strong negative linear relationship



Example 4.16

Calculate the coefficient of correlation for the three sets of data above.

Example 4.16

Because we've already calculated the covariances we only need compute the standard deviations of X and Y.

$$\bar{x} = \frac{2+6+7}{3} = 5.0$$

$$\bar{y} = \frac{13+20+27}{3} = 20.0$$

$$s_x^2 = \frac{(2-5)^2 + (6-5)^2 + (7-5)^2}{3-1} = \frac{9+1+4}{2} = 7.0$$

$$s_y^2 = \frac{(13-20)^2 + (20-20)^2 + (27-20)^2}{3-1} = \frac{49+01+49}{2} = 49.0$$

Example 4.16

The standard deviations are

$$s_x = \sqrt{7.0} = 2.65$$

$$s_y = \sqrt{49.0} = 7.00$$

Example 4.16

The coefficients of correlation are

Set 1: $r = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{(2.65)(7.0)} = .943$

Set 2: $r = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{(2.65)(7.0)} = -.943$

Set 3: $r = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{(2.65)(7.0)} = -.189$

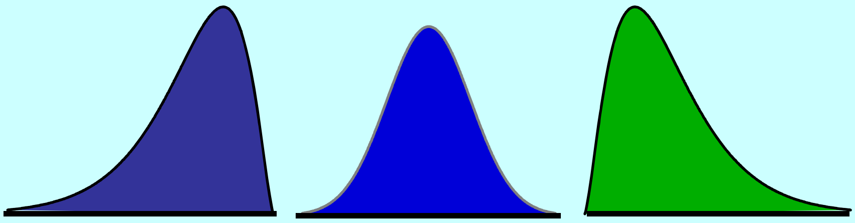
Parameters and Statistics

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	S^2
Standard Deviation	σ	S
Coefficient of Variation	CV	cv
Covariance	σ_{xy}	S_{xy}
Coefficient of Correlation	ρ	r

Measures of Shape

- **Skewness**
  - –Absence of symmetry
  - –Extreme values in one side of a distribution
- **Kurtosis**
  - –Peakedness of a distribution
  - –Leptokurtic: high and thin
  - –Mesokurtic: normal shape
  - –Platykurtic: flat and spread out

Skewness

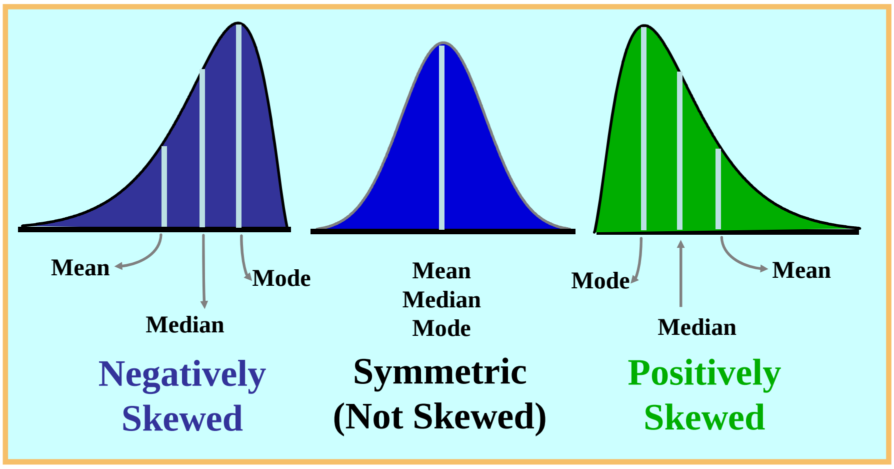


Negatively Skewed

Symmetric (Not Skewed)

Positively Skewed

Skewness



Coefficient of Skewness

- Summary measure for skewness

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

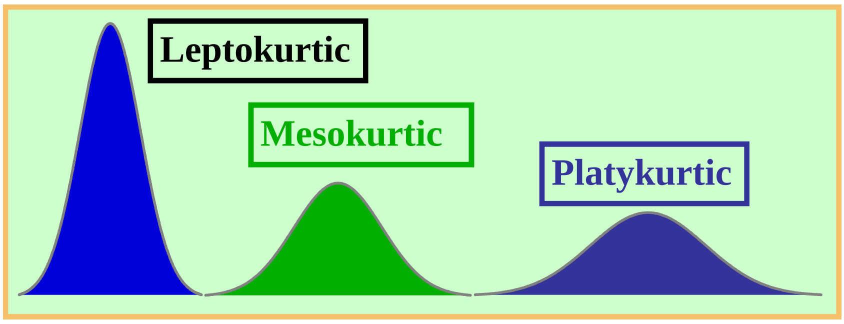
- If S < 0, the distribution is <u>negatively skewed</u> (skewed to the left).
- If S = 0, the distribution is <u>symmetric</u> (not skewed).
- If S > 0, the distribution is <u>positively skewed</u> (skewed to the right).

Coefficient of Skewness

$\mu_1 = 23$ $M_{d1} = 26$ $\sigma_1 = 12.3$ $S_1 = \frac{3(\mu_1 - M_{d1})}{\sigma_1}$ $= \frac{3(23-26)}{12.3}$ $= -0.73$	$\mu_2 = 26$ $M_{d2} = 26$ $\sigma_2 = 12.3$ $S_2 = \frac{3(\mu_2 - M_{d2})}{\sigma_2}$ $= \frac{3(26-26)}{12.3}$ $= 0$	$\mu_3 = 29$ $M_{d3} = 26$ $\sigma_3 = 12.3$ $S_3 = \frac{3(\mu_3 - M_{d3})}{\sigma_3}$ $= \frac{3(29-26)}{12.3}$ $= +0.73$

Kurtosis

- Peakedness of a distribution
  - –Leptokurtic: high and thin
  - –Mesokurtic: normal in shape
  - –Platykurtic: flat and spread out


Leptokurtic
Mesokurtic
Platykurtic