Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella
University of Adelaide
School of Economics

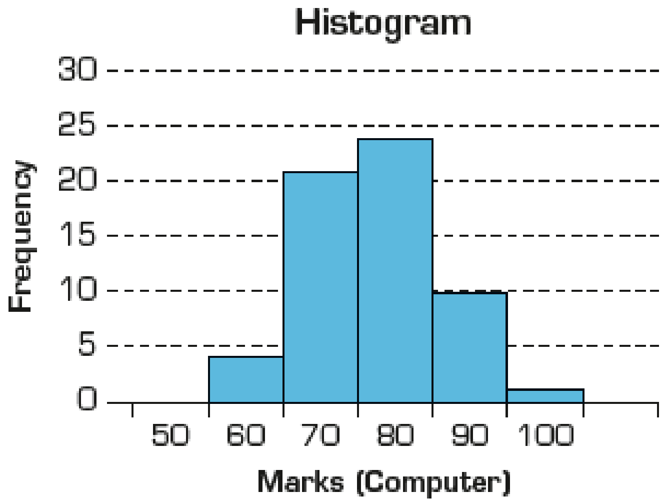
CHAPTER 5

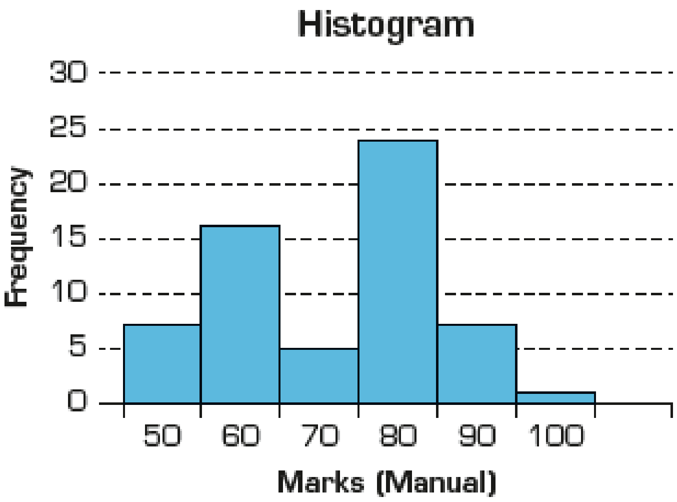
Numerical descriptive measures

Two classes of measures

1. Measures of **central location**
2. Measures of **variability**

Remember from last week:





Measures of central location

Three main types of **measures of central location** are:

- Arithmetic mean (or average)
- Median
- Mode

Arithmetic Mean (or Average)

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

$$\text{Mean} = \frac{\text{Sum of measurements}}{\text{Number of measurements}}$$

Sample mean

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

Sample size

Population mean

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

Population size

Example 1

Find the mean of a *sample* of six measurements

1, 3, 5, 2, 4, 3

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{\boxed{1} + \boxed{3} + \boxed{5} + \boxed{2} + \boxed{4} + \boxed{3}}{6} = \boxed{3.0}$$

Example 2

When many of the measurements have the same value, the measurement can be summarised in a frequency table. Suppose the numbers of children in a sample of 20 families were recorded below. Calculate the average number of children in a family.

NUMBER OF CHILDREN	0	1	2	3	4
NUMBER OF FAMILIES	3	4	7	2	4

20 families

Solution:

Average number of children in a family is

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{x_1 + x_2 \ldots + x_{20}}{20} = \frac{3(0)+4(1)+7(2)+2(3)+4(4)}{20} = 2.0$$

The arithmetic mean…

The average or the arithmetic mean is appropriate for describing measurement data, e.g. heights of people, marks of student exams, etc.

The mean is seriously affected by extreme values called ‘outliers’. E.g. as soon as a billionaire moves into a neighborhood, the average household income for the neighbourhood increases beyond what it was previously!

Median

Another most commonly used measure of central location is the median.

The **median** of a set of measurements is the value that falls in the middle when the measurements are arranged in order of magnitude.

There is a unique median for each data set

Example 3

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
Median = **8**

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.

Impact of an outlier on the Mean and Median

Example 4

Seven employee salaries were recorded (in ‘000s):

42, 45, 40, 46, 44, 40, 43.

(a) Find the median salary.

(b) Suppose the director’s salary of $200 000 was added to the group recorded before. Find the median salary.

(c) Compare the mean and the median values for the data in parts a and b.

Example 4: Solution

a. Odd number of observations

First, sort the salaries.

Then, locate the value in the middle.

40,40,42,**43**,44,45,46

b. Even number of observations

First, sort the salaries.

Then, locate the values in the middle.

There are two middle values!

40,40,42,**43**,**44**,45,46,200

40,40,42,**43**, **44**,45,46,200

40,40,42,43, **43.5**, 44,45,46,200

Example 4: Solution…

c) For the data in (a) and (b),

(a) Without the outlier

$$\text{Median}_{(a)} = 43.0$$

$$\text{Mean}_{(a)} = \frac{42+45+\ldots+43}{7} = \frac{300}{7} = 42.8$$

(b) With the outlier

$$\text{Median}_{(b)} = 43.5$$

$$\text{Mean}_{(b)} = \frac{500}{8} = 62.5$$

As can be seen, the median did not change that much (43 vs 43.5), even with the outlier (200). However, the mean has changed from 42.8 to 62.5.

Mean is affected by the outlier, whereas the median is not.

Mode

Another commonly used measure of central location is the mode.

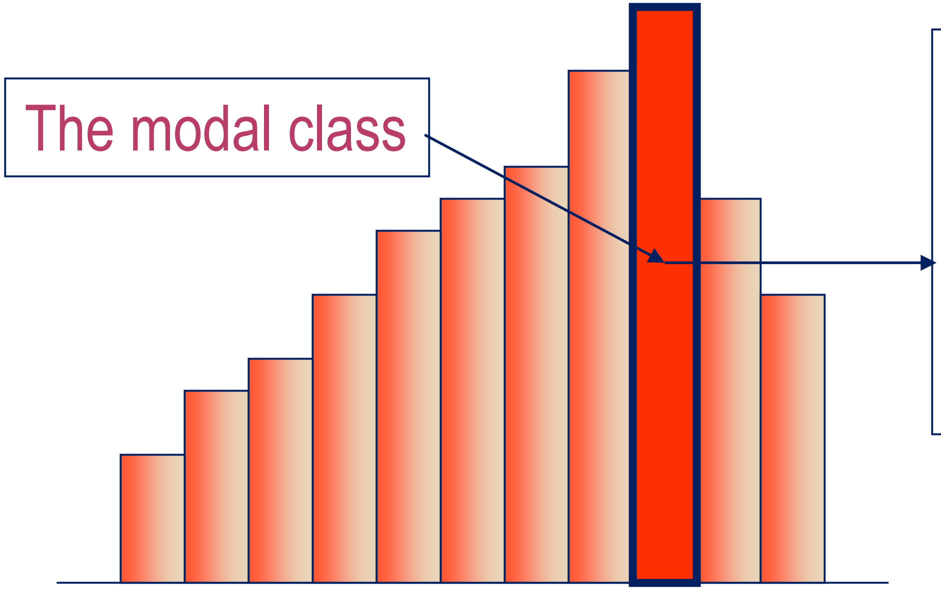
The **mode** of a set of observations is the value that occurs most frequently.

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

Mode is 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.

Mode



The modal class

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

Sample and population modes are computed the same way.

Example 5

(Example 5.4, page 134)

XM05-04 The manager of a menswear store observed the waist size (in centimeters) of trousers sold yesterday: 77, 85, 90, 85, 82, 70, 85, 75, 85, 80, 77, 100, 85, 70. Suggest a suitable size of trousers to be ordered more with the next order.

Solution:

The mode, the size with the highest sales, for this data set, is 85 cm.

Mean = 81.9

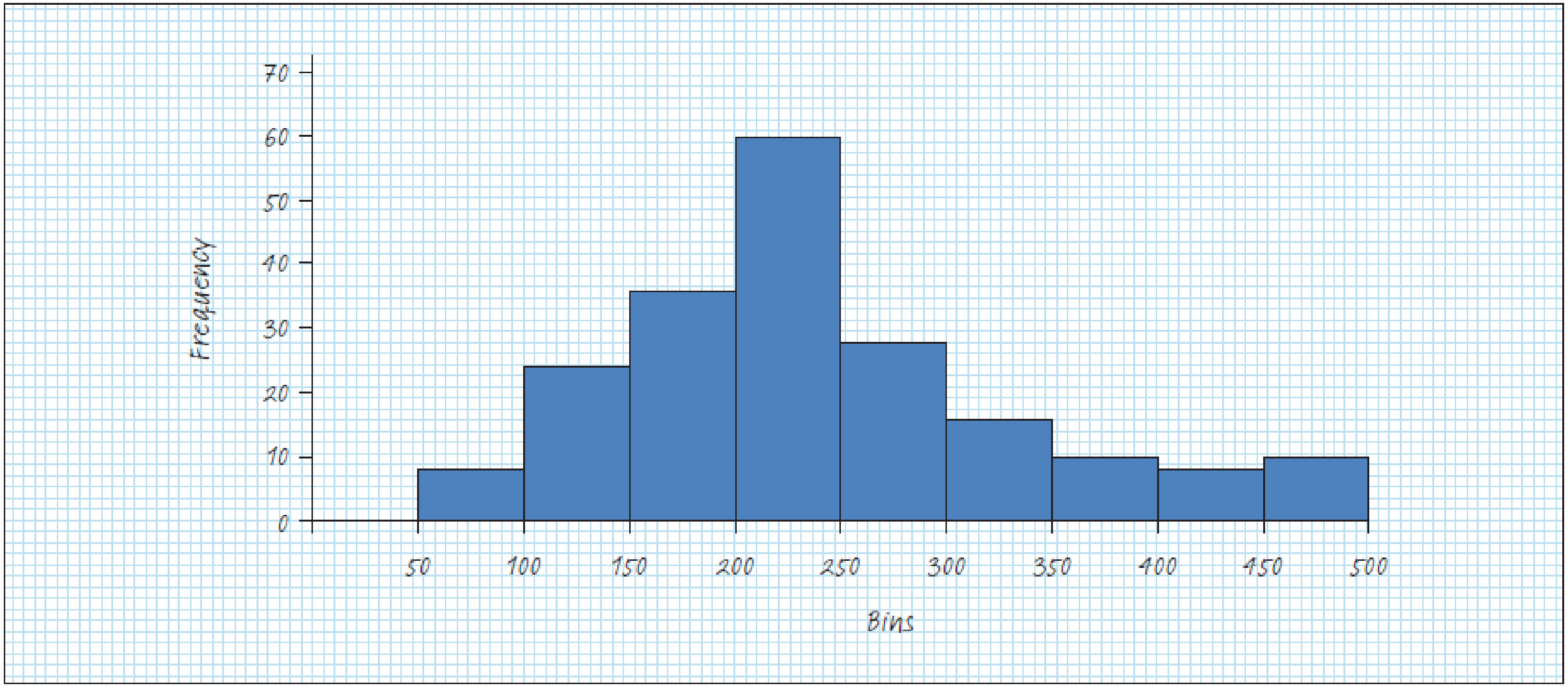
Median = 83.5

This information seems valuable (for example, for the design of a new display in the store), much more than 'the median is 83.5 cm'.

Relationship between Mean, Median and Mode

Preliminary note: it's convenient to approximate a **histogram** (or **distribution**) with a smooth line.

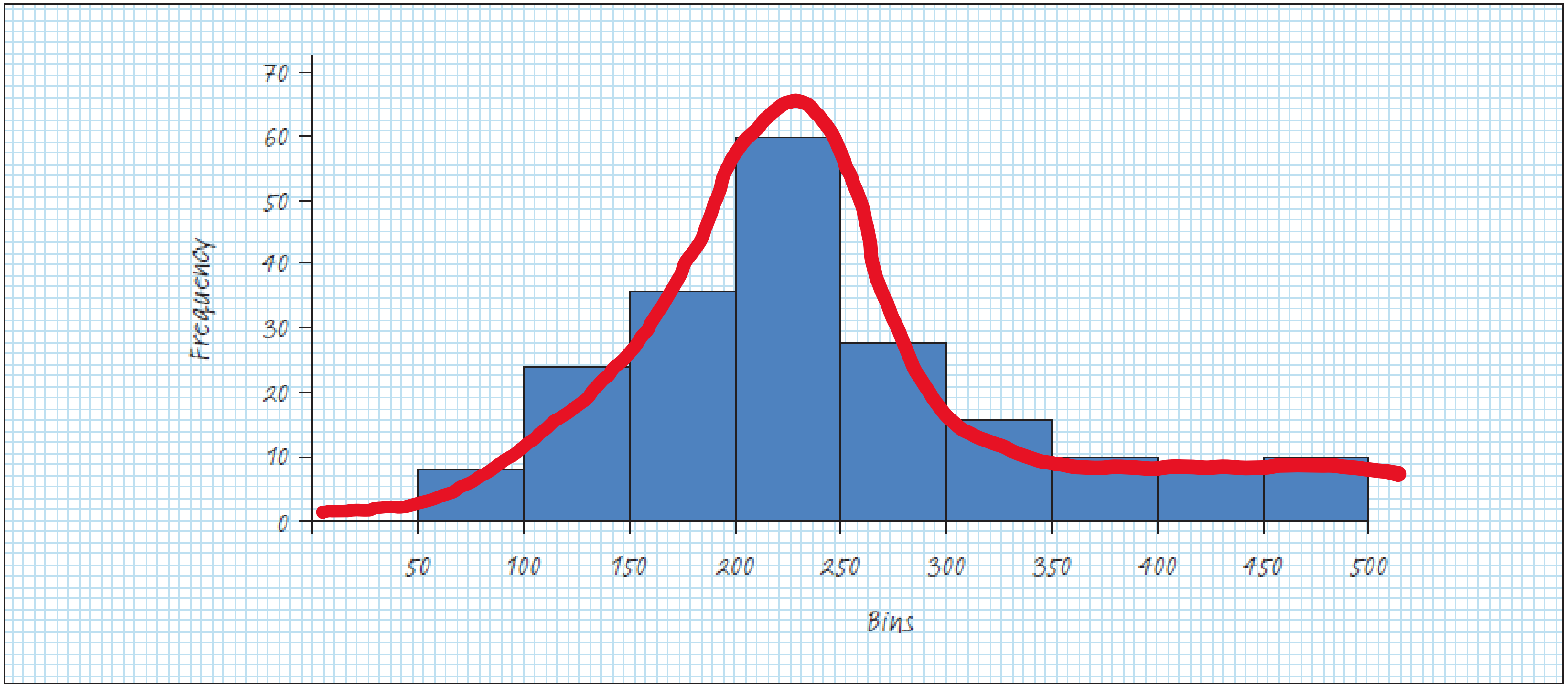
Figure 4.1 Histogram of electricity bills of 200 new Brisbane customers



Relationship between Mean, Median and Mode

Preliminary note: it's convenient to approximate a **histogram** (or **distribution**) with a smooth line.

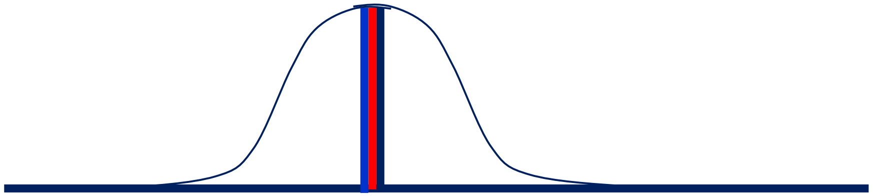
Figure 4.1 Histogram of electricity bills of 200 new Brisbane customers



Relationship between Mean, Median and Mode

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

A symmetric distribution



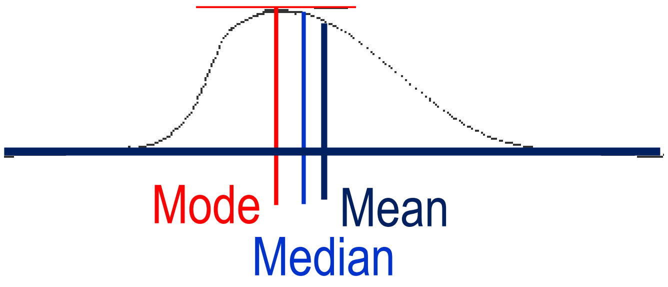
Mean=Median=Mode

If the distribution is symmetrical, then
Mean = Median = Mode.

Relationship between Mean, Median and Mode

If a distribution is not symmetrical, and skewed to the right (positively skewed), the three measures differ.

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



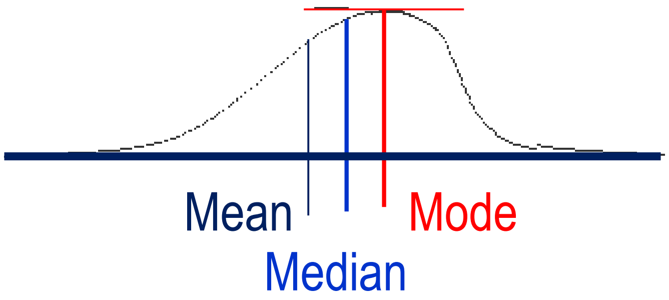
If the distribution is positively skewed, then

Mean > Median > Mode.

Relationship between Mean, Median and Mode

If a distribution is not symmetrical, and skewed to the left (negatively skewed), the three measures differ.

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



If the distribution is negatively skewed, then

Mean < Median < Mode.

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 (for example, if there are outliers in the dataset).

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 the following example.

The number of hours of Internet use in the previous month among 10 primary school children were 13, 11, 12, 10, 13, 14, 11, 7, 9, 10.

The mean was 11.0 and the median was 8.5.

Now suppose that the child who reported 14 hours actually reported 114 hours (obviously an Internet addict). The data now is 13, 11, 12, 10, 13, 114, 11, 7, 9, 10.

The new mean is 21.0 and the median is 8.5.

The median is not affected much by this outlier, but the mean is.

Mean, Median, Mode: Which is best?...

This value is only exceeded by only one of the ten observations in the sample, making this statistic (mean) 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 and Mode for Ordinal and 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 numerical (or interval) data.

Compute the median to

Describe the central location of a single set of numerical or ordinal (ranked) data.

Compute the mode to

Describe a single set of nominal (or categorical) data.

Measures of variability

Measures of central location fail to tell the whole story about the distribution.

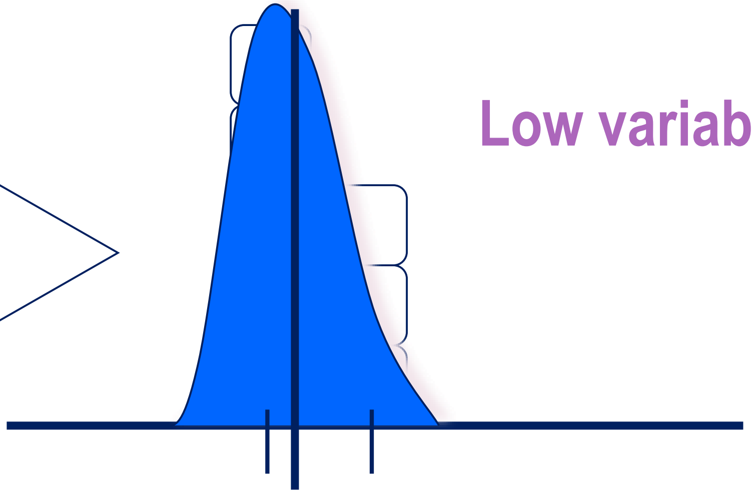
A question of interest still remains unanswered:

How typical is the average value of all the measurements in the data set?

or

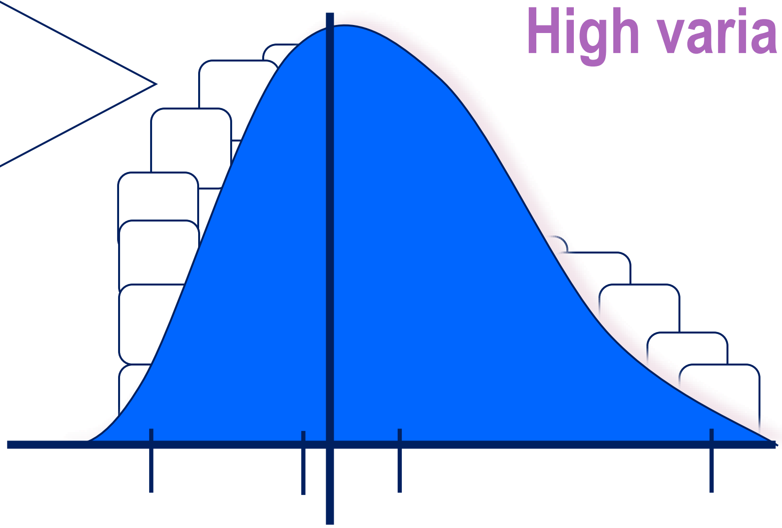
How spread out are the measurements around the average value?

Observe Two Hypothetical Data Sets



Low variability data set

The average value provides a good representation of the values in the data set.



High variability data set

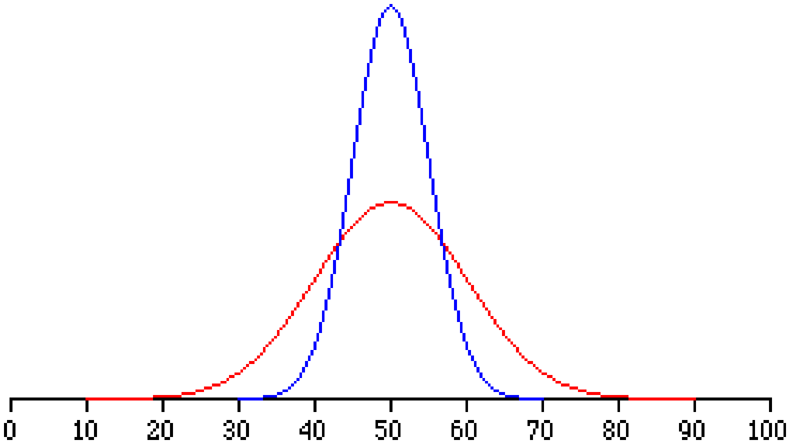
The same average value does not provide as good presentation of the values in the data set as before.

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, very easy to calculated as:

Range = Largest observation – Smallest observation

E.g.

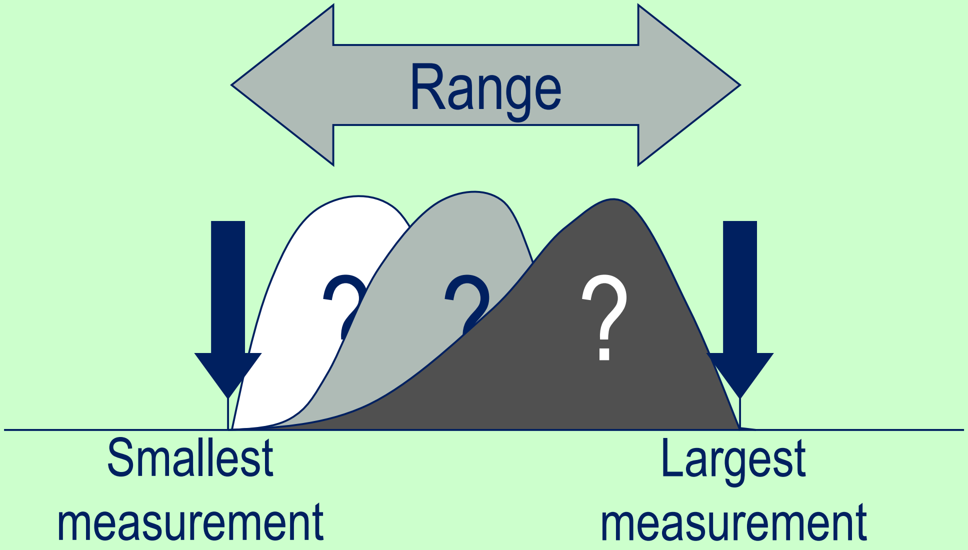
Data: {4, 4, 4, 4, 50}	Range = 46
Data: {4, 8, 15, 24, 39, 50}	Range = 46

The range is the same in both cases, but the data sets have very different distributions…

Range…

But how do all the measurements spread out?

The range cannot assist in answering this question.



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).

(this is a convention, of course, not a law of nature…)

Variance…

This measure of dispersion reflects the values of *all* the measurements.

- The variance of *a population* of N measurements x_1, x_2, …, x_N having a mean μ is defined as

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

Variance…

- The variance of a sample of n measurements x_1, x_2, …, x_n having a mean $\overline{X}$ is defined as

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

Two good questions:

1. Why n-1 and not n ?!? This way the sample variance is a good estimate of the population variance (more later)
2. Why not using just the sum of deviations?

Consider two small populations:

Population A: 8, 9, 10, 11, 12

Population B: 4, 7, 10, 13, 16

Let us start by calculating the sum of deviations

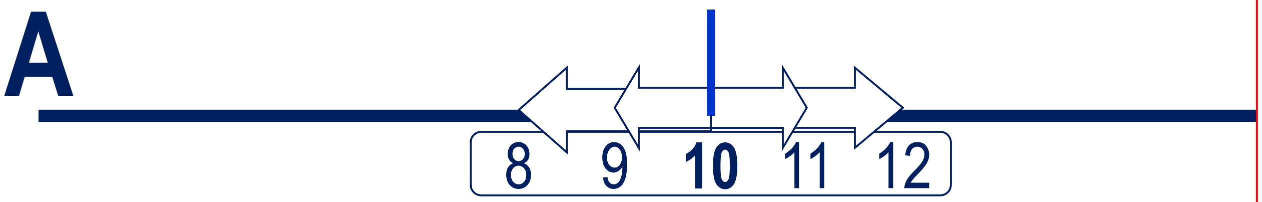
9–10= –1

11–10= +1

8–10= –2

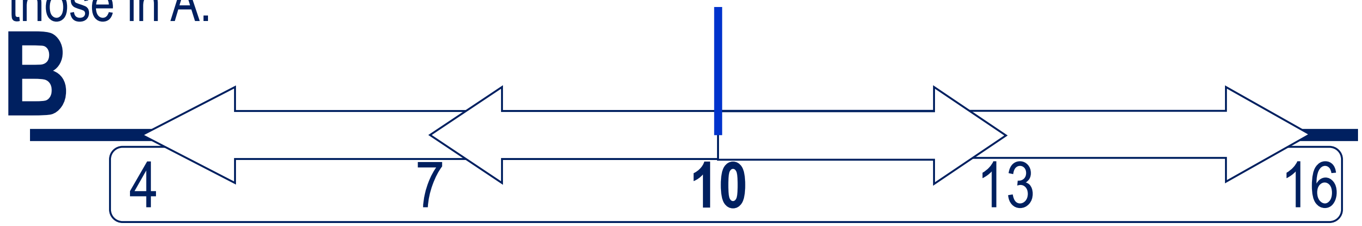
12–10= +2

Sum = 0



… but measurements in B are much more dispersed then those in A.

The mean of both populations is 10...



4–10 = –6

16–10 = +6

7–10 = –3

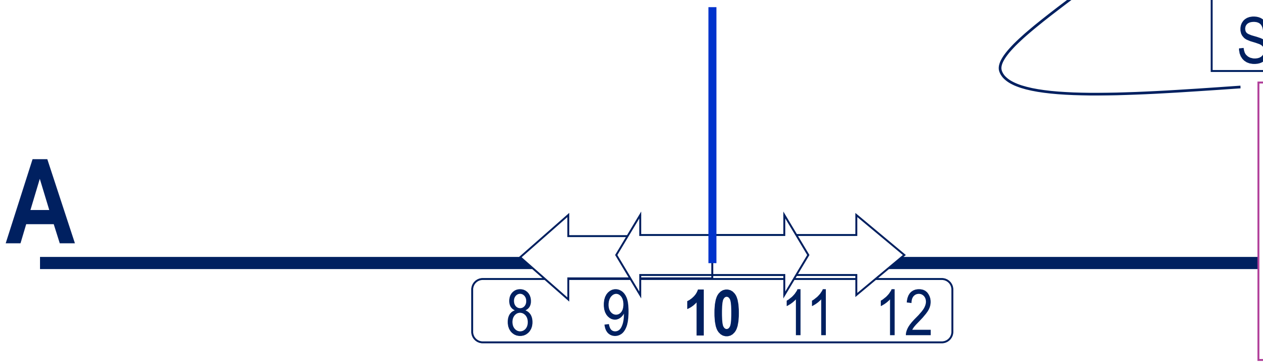
13–10 = +3

Sum = 0

The sum of deviations is zero in both cases, therefore another measure is needed.

Thus, a measure of dispersion is needed that agrees with this observation.

The sum of *squared* deviations is used in calculating the variance. See example next.



9–10= –1

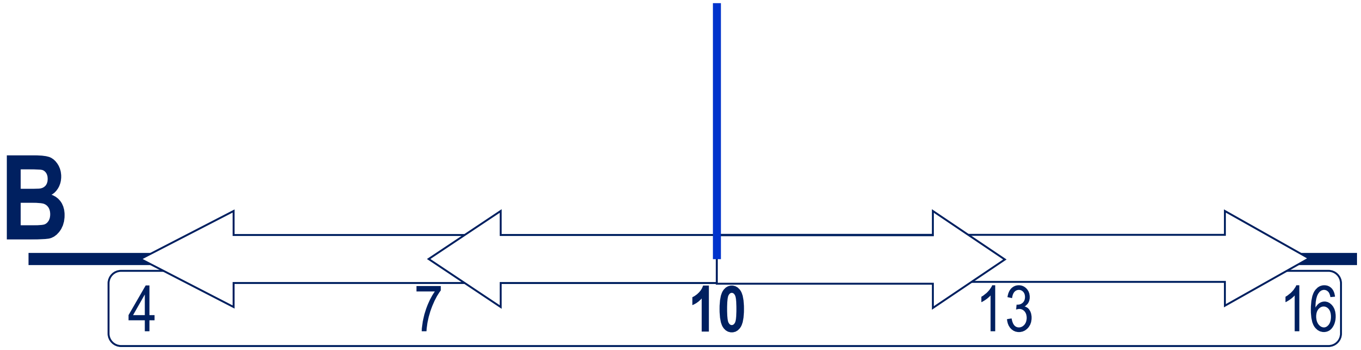
11–10= +1

8–10= –2

12–10= +2

Sum = 0

The sum of deviations is zero in both cases, therefore another measure is needed.



4–10 = – 6

16–10 = +6

7–10 = –3

13–10 = +3

Sum = 0

Variance…

Let us calculate the variance of the two populations.

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

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

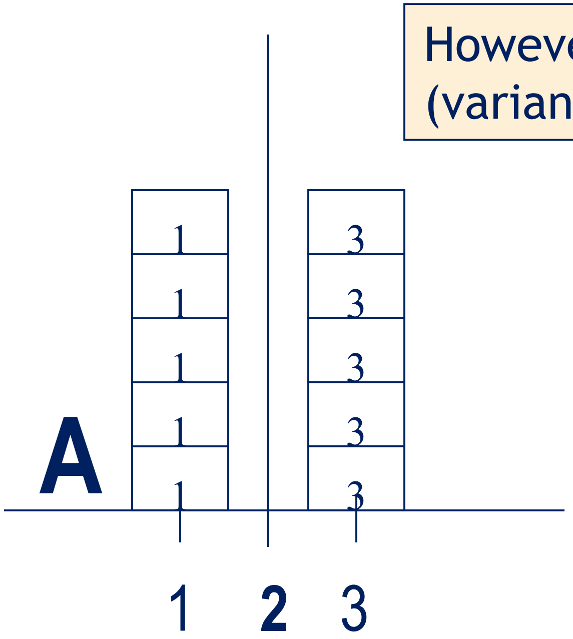
Why is the variance defined as the average squared deviation? Why not use the sum of squared deviations as a measure of dispersion instead?

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

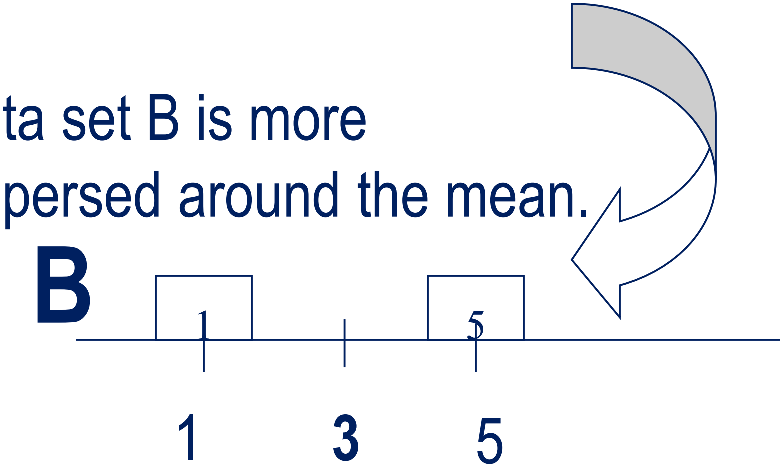
Which data set has a larger dispersion?

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

However, when calculated on a ‘per observation’ basis (variance), the data set dispersions are properly ranked.



Data set B is more dispersed around the mean.



$$\text{Sum}_A = \underbrace{(1-2)^2 + \ldots + (1-2)^2}_{\text{5 times}} + \underbrace{(3-2)^2 + \ldots + (3-2)^2}_{\text{5 times}} = 10$$

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

$$\sigma_A^2 = \text{Sum}_A/N = 10/10 = 1$$

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

Example

The following sample consists of the number of jobs six students applied for: 17, 15, 23, 7, 9, 13. Finds its mean and variance.

Solution:

Sample Mean

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

…as opposed to μ or σ^2

Example: Solution…

Sample Variance

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

Standard deviation

The ***standard deviation*** of a set of measurements is the square root of the variance of the measurements.

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

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

Example

(Example 5.8, page 149)

XM05-08 Rates of return over the past 10 years for two unit trusts are shown below. Which one has a higher level of risk?

Trust A: 12.3, –2.2, 24.9, 1.3, 37.6, 46.9, 28.4, 9.2, 7.1, 34.5
Trust B: 15.1, 0.2, 9.4, 15.2, 30.8, 28.3, 21.2, 13.7, 1.7, 14.4

Example 8: Solution

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

Trust A		***Trust B***	
Mean	20	Mean	15
Standard Error	5.295	Standard Error	3.152
Median	18.6	Median	14.75
Mode	#N/A	Mode	#N/A
Standard Deviation	16.743	Standard Deviation	9.969
Sample Variance	280.340	Sample Variance	99.373
Kurtosis	-1.342	Kurtosis	-0.464
Skewness	0.217	Skewness	0.107
Range	49.1	Range	30.6
Minimum	-2.2	Minimum	0.2
Maximum	46.9	Maximum	30.8
Sum	200	Sum	150
Count	10	Count	10

Even though Trust A has a higher average return, it should be considered riskier because its standard deviation is larger.

Interpreting Standard Deviation

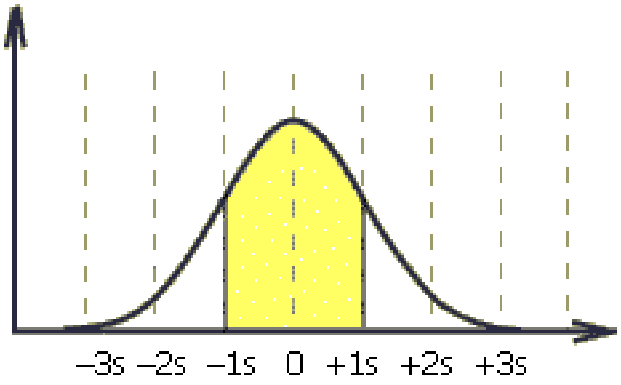
The standard deviation can be used to compare the variability of several distributions and make a statement about the general shape of a distribution.

If the histogram is **bell shaped**, then we can use the ***Empirical Rule***, which states:

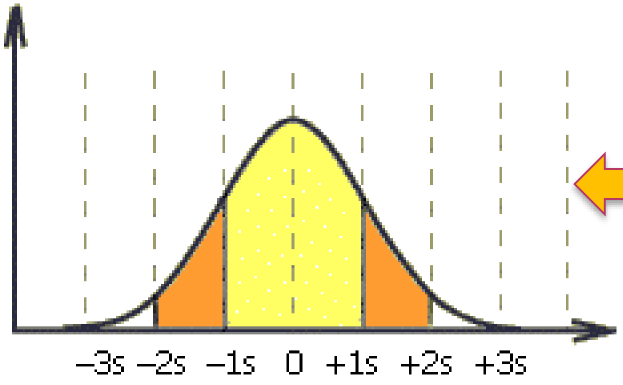
1) Approximately 68% of all observations fall within one standard deviation of the mean.
2) Approximately 95% of all observations fall within two standard deviations of the mean.
3) Approximately 99.7% of all observations fall within three standard deviations of the mean.

Empirical rule…

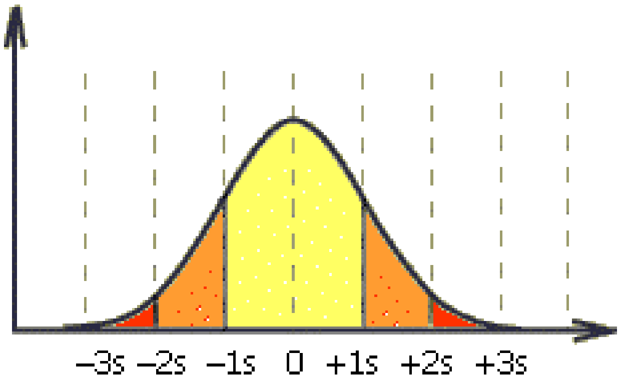
Approximately **68%** of all observations fall within **one** standard deviation of the mean.



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.



Example

A statistician wants to describe the way returns on investment are distributed.

The mean return = 10%

The standard deviation of the return = 3%

The histogram is bell-shaped.

How can the statistician use the mean and the standard deviation to describe the distribution?

Example: Solution

The empirical rule can be applied (bell-shaped histogram).

Describing the return distribution:

Approximately 68% of the returns lie between 7% and 13%
[10 – 1(3), 10 + 1(3)]

Approximately 95% of the returns lie between 4% and 16%
[10 – 2(3), 10 + 2(3)]

Approximately 99.7% of the returns lie between 1% and 19%
[10 – 3(3), 10 + 3(3)]

Example

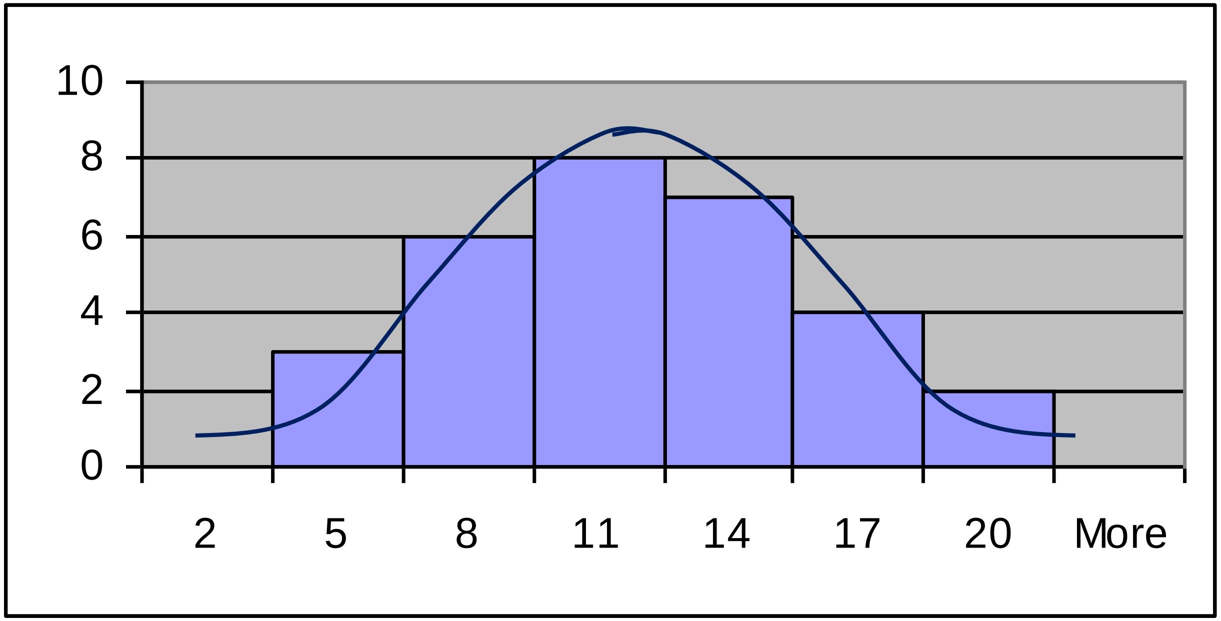
(Example 5.10, page 152)

XM05-10 The duration of 30 telephone calls (in minutes) are shown below. Check the empirical rule for this set of measurements.

11.8	3.6	16.6	13.5	4.8	8.3
8.9	9.1	7.7	2.3	12.1	6.1
10.2	8	11.4	6.8	9.6	19.5
15.3	12.3	8.5	15.9	18.7	11.7
6.2	11.2	10.4	7.2	5.5	14.5

Example: Solution

1. First check if the histogram has an approximate shape:



Example: Solution…

2. Calculate the mean and the standard deviation:
   mean = 10.26; standard deviation = 4.29.

3. Calculate the intervals:

$$(\bar{x} - s, \bar{x} + s) = (10.26 - 4.29, 10.26 + 4.29) = (5.97, 14.55)$$

$$(\bar{x} - 2s, \bar{x} + 2s) = (1.68, 18.84)$$

$$(\bar{x} - 3s, \bar{x} + 3s) = (-2.61, 23.13)$$

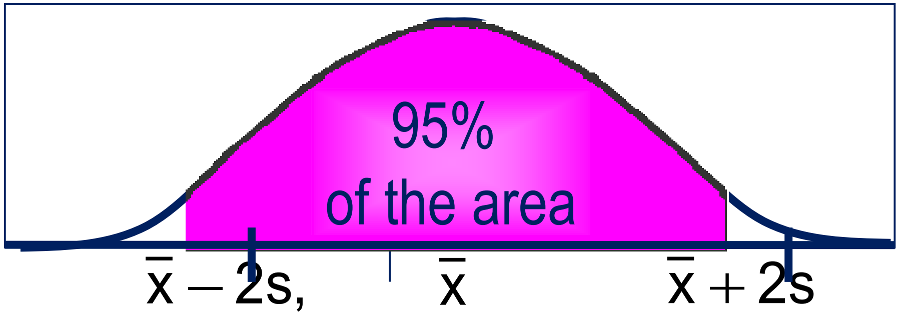
k			Interval	Empirical rule	Actual percentage
1	$(\bar{x} - s, \bar{x} + s)$	=	[5.97, 14.55]	68%	70%
2	$(\bar{x} - 2s, \bar{x} + 2s)$	=	[1.68, 18.84]	95%	96.7%
3	$(\bar{x} - 3s, \bar{x} + 3s)$	=	[–2.61, 23.13]	100%	100%

Approximate standard deviation

By the empirical rule, approximately 95% of the area under a mound-shaped histogram lies within $(\bar{x}-2s, \bar{x}+2s)$

Therefore, range can be approximated by 4s. In other words,

$$s \cong \frac{\text{Range}}{4}$$



For Example 8, for Trust B returns, the range is 30.8 - 0.2 = 30.6 percent.

$$s \cong \frac{30.6}{4} = 7.51 \text{ percent}$$

Actual standard deviation of Trust B returns is 9.97%

Interpreting Standard Deviation

Suppose that the mean and standard deviation of last year's mid-semester exam 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 89% 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 measurements is the standard deviation divided by the mean value.

$$\text{Sample coefficient of variation: } \text{cv} = \frac{s}{\bar{x}}$$

$$\text{Population coefficient of variation: } \text{CV} = \frac{\sigma}{\mu}$$

Coefficient of Variation

$$\text{Sample coefficient of variation: } \text{cv} = \frac{s}{\bar{x}}$$

$$\text{Population coefficient of variation: } \text{CV} = \frac{\sigma}{\mu}$$

This coefficient provides a proportionate measure of variation.

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

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.

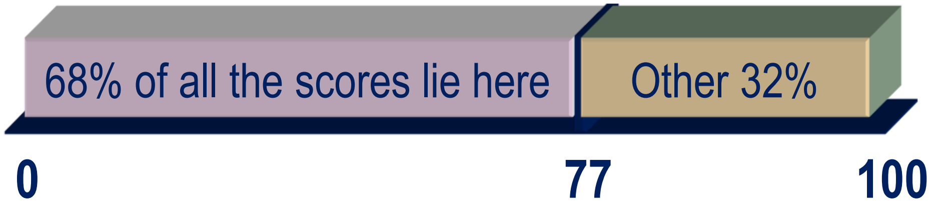
Suppose you scored in the 60th percentile on your final exam, that means 60% of the other students' scores were *below* yours, while 40% of scores were *above* yours.

Percentiles

The p^{th} percentile of a set of measurements is the value for which

- at most p% of the measurements are less than that value
- at most (100-p)% of all the measurements are greater than that value.

For example, suppose 77 is the 68th percentile of a statistics exam score. Then



Quartiles

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

- First (lower) quartile, **Q_1** = 25th percentile (p_{25})
- Second (middle) quartile, **Q_2** = 50th percentile (p_{50}) (which is also the median)
- Third (upper) quartile, **Q_3** = 75th percentile (p_{75})

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

For example, 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 and 20% scored **higher** than you in the exam. It is about your position relative to others, not the actual mark.

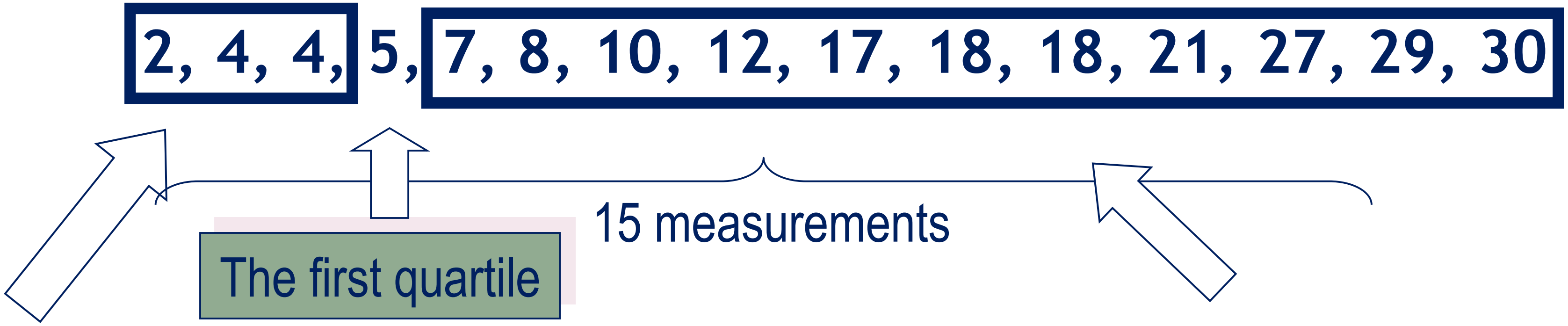
Example

Find the quartiles of the following set of measurements

7, 18, 12, 17, 29, 18, 4, 27, 30, 2, 4, 10, 21, 5, 8

Example: Solution

First sort the measurements



2, 4, 4, 5, 7, 8, 10, 12, 17, 18, 18, 21, 27, 29, 30

15 measurements

The first quartile

At most (0.25)(15) = 3.75 measurements should appear below the first quartile. Check the first 3 measurements on the left hand side.

At most (0.75)(15)=11.25 measurements should appear above the first quartile. Check 11 measurements on the right hand side.

If the number of measurements is even, two measurements will remain unchecked. In this case choose the midpoint between these two measurements.

Location of Percentiles

Find the location of any percentile using the formula

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

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

Think of easy cases, e.g., the median, to make sense of this formula. You will realise it’s very intuitive!

Example

Calculate the 25th, 50th, and 75th percentile of the data:

0, 7, 12, 5, 33, 14, 8, 0, 9, 22

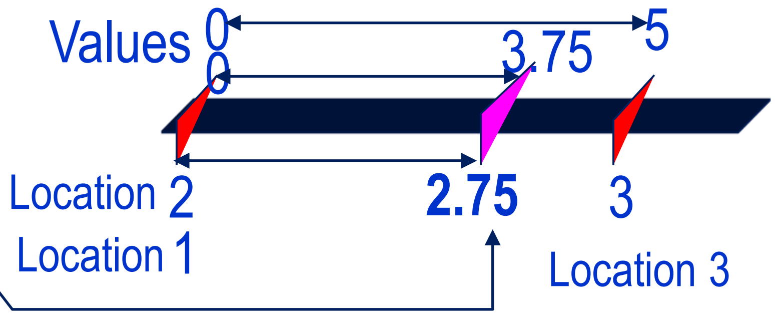
Example: Solution

After sorting the data we have

0, 0, 5, 7, 8, 9, 12, 14, 22, 33.

Location (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

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



The 2.75th location translates to the value

$p_{25} = 0 + (.75)(5 – 0) = 3.75$

2nd observation

3rd observation

2nd observation

Example 12: Solution…

$$L_{50} = (10 + 1)\frac{50}{100} = 5.5$$

The 50th percentile is halfway between the fifth and sixth observations (in the middle between 8 and 9), that is 8.5. That is,

$$p_{50} = \underset{\uparrow}{8} + (0.5)(\underset{\uparrow}{9} - \underset{\uparrow}{8}) = 8.5$$

5th observation

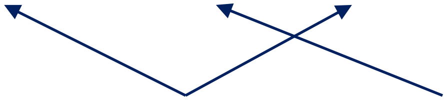
6th observation

Example 12: Solution…

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

The 75th percentile is one quarter of the distance between the eighth and ninth observation. That is

$$p_{75} = 14 + .25(22 - 14) = 16.$$

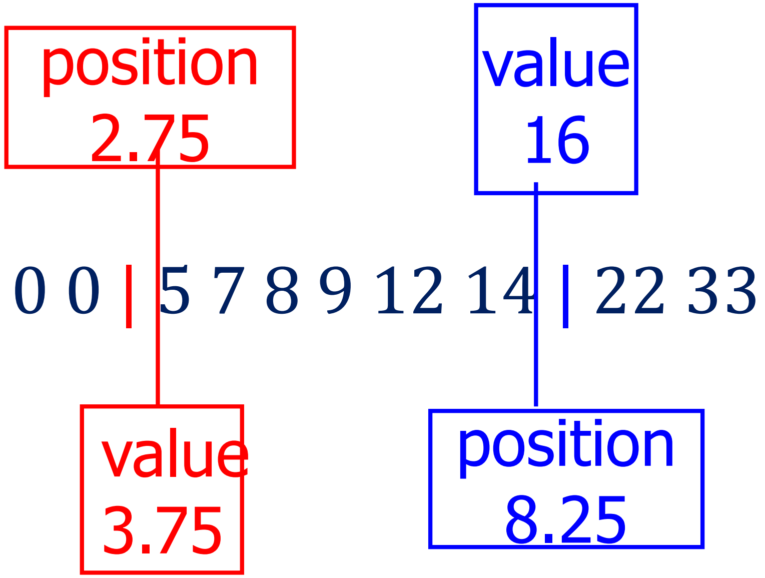


8th observation

9th observation

Location of Percentiles…

Please remember…



position 2.75

value 16

0 0 | 5 7 8 9 12 14 | 22 33

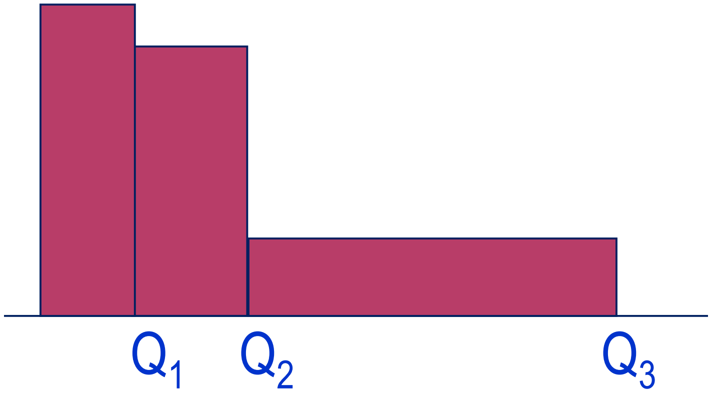
value 3.75

position 8.25

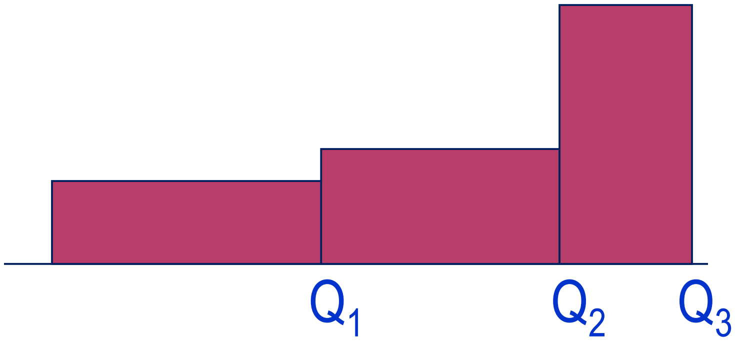
L_p determines the **position** in the data set where the percentile value lies, not the value of the percentile itself.

Quartiles and Variability

Quartiles can provide an idea about the shape of a histogram.



Positively skewed histogram



Negatively skewed histogram

Interquartile Range…

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

Interquartile Range (IQR) = $Q_3 - Q_1$

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.