"As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality."
- Albert Einstein (1879-1955)

Statistics:

"The science of producing unreliable facts from reliable figures."
- Evan Esar

"Facts are stubborn, but statistics are more pliable."
- Mark Twain

"I can prove anything by statistics except the truth."
- George Canning

Two main branches:

Descriptive Statistics:

Trying to describe a data set with a few or several numbers.

Predictive or Inferential Statistics:

Trying to reach conclusions beyond an original data set.

OPENING PROBLEM

A farmer is investigating the effect of a new organic fertiliser on his crops of peas. He has divided a small garden into two equal plots and planted many peas in each. Both plots have been treated the same except the fertiliser have been used on one but not the other.

A random sample of 150 pods is harvested from each plot at the same time, and the number of peas in each pod is counted. The results are:

© iStockphoto

Without fertiliser

4 6 5 6 5 6 4 6 4 9 5 3 6 8 5 4 6 8 6 5 6 7 4 6 5 2 8 6 5 6 5 5 5 4 4 4 6 7 5 6 7 5 5 6
4 8 5 3 7 5 3 6 4 7 5 6 5 7 5 7 6 7 5 4 7 5 5 5 6 6 5 6 7 5 8 6 8 6 7 6 6 3 7 6 8 3 3 4
4 7 6 5 6 4 5 7 3 7 7 6 7 7 4 6 6 5 6 7 6 3 4 6 6 3 7 6 7 6 8 6 6 6 6 4 7 6 6 5 3 8 6 7
6 8 6 7 6 6 6 8 4 4 8 6 6 2 6 5 7 3

With fertiliser

6 7 7 4 9 5 5 5 8 9 8 9 7 7 5 8 7 6 6 7 9 7 7 7 8 9 3 7 4 8 5 10 8 6 7 6 7 5 6
8 7 9 4 4 9 6 8 5 8 7 7 4 7 8 10 6 10 7 7 7 9 7 7 8 6 8 6 8 7 4 8 6 8 7 3 8 7 6
9 7 6 9 7 6 8 3 9 5 7 6 8 7 9 7 8 4 8 7 7 7 6 6 8 6 3 8 5 8 7 6 7 4 9 6 6 6 8 4
7 8 9 7 7 4 7 5 7 4 7 6 4 6 7 7 6 7 8 7 6 6 7 8 6 7 10 5 13 4 7 11

Things to think about:

- Can you state clearly the problem that the farmer wants to solve?
- How has the farmer tried to make a fair comparison?
- How could the farmer make sure that his selection was at random?
- What is the best way of organising this data?
- What are suitable methods of displaying the data?
- Are there any abnormally high or low results and how should they be treated?
- How can we best describe the most typical pod size?
- How can we best describe the spread of possible pod sizes?
- Can the farmer make a reasonable conclusion from his investigation?

A KEY STATISTICAL CONCEPTS

	A collection of individuals about which we want to draw conclusions.
	The collection of information from the **whole population**.
	A subset of the population. It is important to choose a sample at **random** to avoid **bias** in the results.
	The collection of information from a **sample**.
• **Data** (singular **datum**)	Information about individuals in a population.
	A numerical quantity measuring some aspect of a population.
	A quantity calculated from data gathered from a sample. It is usually used to estimate a population parameter.
• **Distribution**	The pattern of variation of data. The distribution may be described as:

symmetrical positively skewed negatively skewed

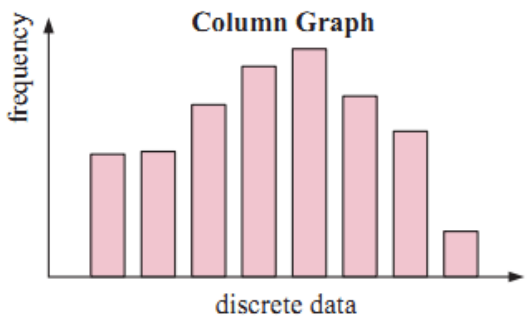
• **Outliers**	Data values that are either much larger or much smaller than the general body of data. They should be included in an analysis *unless* they are the result of human or other error.

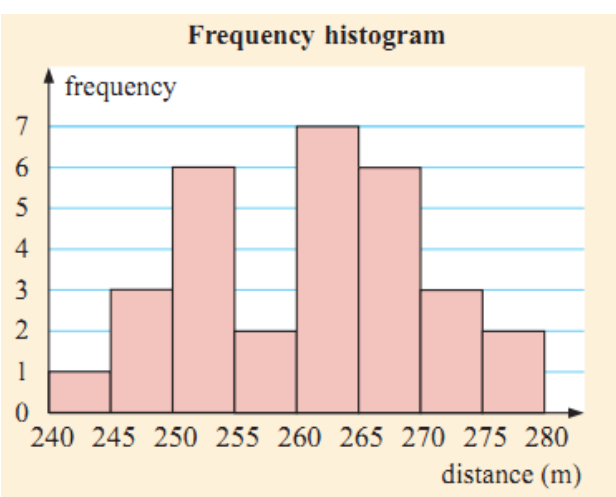
A **discrete numerical variable** takes exact number values and is often a result of **counting**.

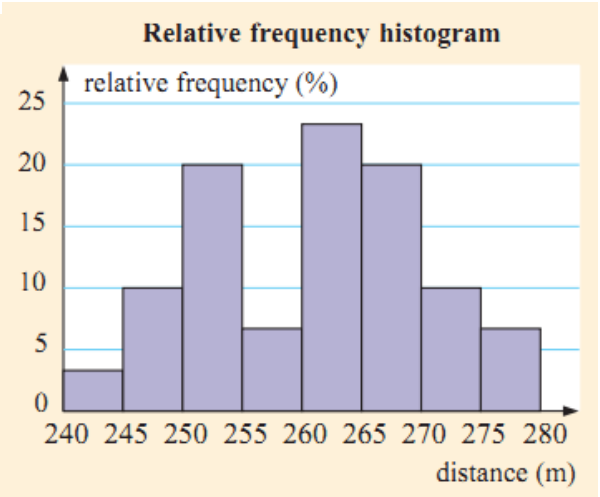
A **continuous variable** takes numerical values within a certain continuous range. It is usually a result of **measuring**.

Ways to display data:

Graphs:







Tables

Length (cm)	*Frequency*
$3 \leqslant l < 4$	3
$4 \leqslant l < 5$	6
$5 \leqslant l < 6$	5
$6 \leqslant l < 7$	4
$7 \leqslant l < 8$	2

Stem & Leaf Plot

Stem	*Leaf*
3	1 6 7
4	2 2 4 6 7 9
5	0 4 4 6 8
6	0 0 6 7
7	2 3

Scale: 3 | 1 means 3.1 cm

Others to come...
HW 14A #1-4

B MEASURING THE CENTRE OF DATA

...or in the US....***Measures of Central Tendency***

Average is a ***generic*** term to mean one number that describes a set of numbers.

Mean:

There is more than one type of mean. Most common is the ***arithmetic mean***.

Mean of a *sample*: $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ Mean of a *population*: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Median:

The middle number in an ordered series of numbers (for *n* odd)
If *n* is even, find the arithmetic mean of the middle two numbers.

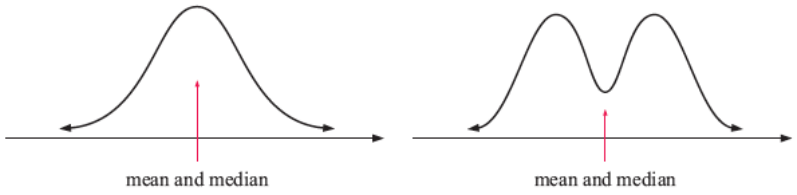
Mode:

The number that occurs most often.
There can be multiple modes.

Which is best? Mean, Median or Mode:

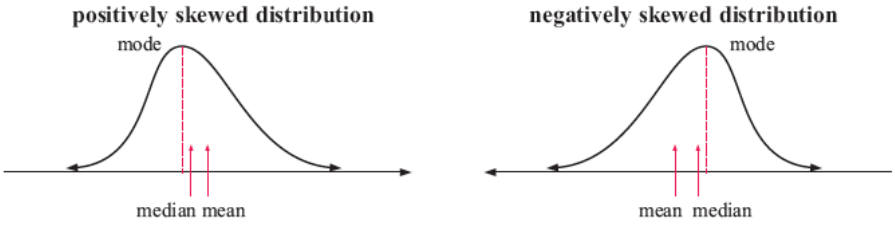
Consider 3,3,3,3,3,10000
1,1,1,1,5,5,5,800
1,1,1,1,2,12,13,13,13

For distributions that are **symmetric**, the mean or median will be approximately equal.



If the data set has symmetry, both the mean and the median should accurately measure the centre of the distribution.

If the data set is not symmetric, it may be positively or negatively skewed:



Notice that the mean and median are clearly different for these skewed distributions.

SL: Opening problem
14A #3, 4
14B.1 #3,5,8,9,11,12,14

Calculations from data summaries

Consider the table below. How can you calculate the mean (efficiently):

Data value (x)	*Frequency* (f)	*Product* (fx)
3	1	$1 \times 3 = 3$
4	1	$1 \times 4 = 4$
5	3	$3 \times 5 = 15$
6	7	$7 \times 6 = 42$
7	15	$15 \times 7 = 105$
8	8	$8 \times 8 = 64$
9	5	$5 \times 9 = 45$
Total	$\sum f = 40$	$\sum fx = 278$

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$ where k is the number of *different* data values.

This formula is often abbreviated as $\overline{x} = \frac{\sum fx}{\sum f}$.

Finding the median from a summary of data.
To find the center number(s) it would help to have the ***cumulative frequency***

Data Value	*Frequency*	*Cumulative Frequency*
3	1	1 ← one number is 3
4	1	2 ← two numbers are 4 or less
5	3	5 ← five numbers are 5 or less
6	7	12 ← 12 numbers are 6 or less
7	15	27 ← 27 numbers are 7 or less
8	8	35 ← 35 numbers are 8 or less
9	5	40 ← all numbers are 9 or less
Total	40	

Working with data in "classes"

Skate purchases					
Price range	50-75	76-85	86-95	96-105	106-120
# of purchases	7	15	11	7	4

What was the mean price of skates purchased?

Use the midpoint of a range when data is grouped in classes.

C MEASURING THE SPREAD OF DATA

We'll explore four measures - the range, the interquartile range, the variance and the standard deviation.

1	3	3	4	6	6	7	8	10	12	12	14	15	15	16	18

Range: The total spread between the extremes.

Largest - smallest = 18 - 1 = 17

Not particularly meaningful in many cases

Quartiles: Break the data into four groups:

Begin by finding the median

1	3	3	4	6	6	7	8	10	12	12	14	15	15	16	18

Lower Quartile = 5

Median = 9

Upper Quartile = 14.5

The median of the lower half of the data is called the ***Lower Quartile***

25% of the data is ≤ the lower quartile

Also known as ***Q1*** or the 25th ***percentile***

The median of the upper half of the data is called the ***Upper Quartile***

75% of the data is ≤ the upper quartile

Also known as ***Q3*** or the 75th ***percentile***

The difference between the upper and lower quartiles is called the ***Interquartile Range***

The middle 50% of the data lie within the interquartile range

Also known as the ***IQR = Q3 - Q1***

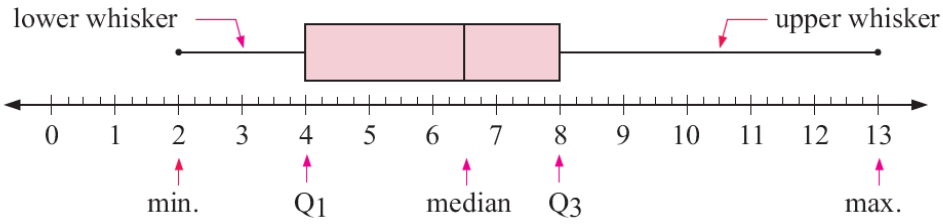
HW: 14C.1: 2, 4, 5

Box & Whisker Plots

Display 5 numbers effectively:

Minimum Q1 Median Q3 Maximum

(aka) ***five number summary***



Box & Whisker Plot Features:

1) Half the data are in the "box"
2) Half the data are in the whiskers, 25% low, 25% high

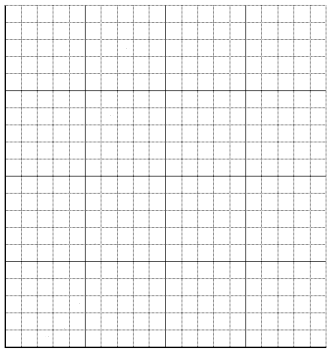
Your calculator will do them!

SL:	**14B.2**	**#2,4,5,7,8,10**
	14B.3	**#2, 3**
SL:	**14C.1**	**#2, 4, 5**
	14C.2	**#2, 3, 6**

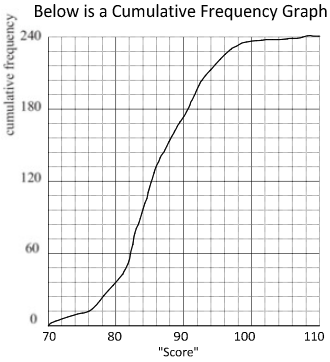
D CUMULATIVE FREQUENCY GRAPHS

We can graph Cumulative Frequency to visualize it. Remember the table:

Data Value	Frequency	Cumulative Frequency
3	1	
4	1	
5	3	
6	7	
7	15	
8	8	
9	5	
Total	40	

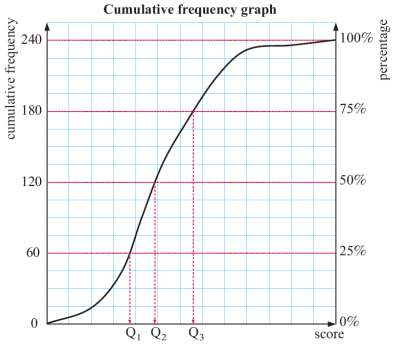


Below is a Cumulative Frequency Graph



What questions can we answer from it?

Can we find the 5 number summary from it?



Percentile:

Lake Wobegon: "...where all the kids are above average"

What does it mean to be in the 75th ***percentile*** in height for HS students?

75% of students are shorter than you!

Q1 is the 25th percentile
Q2 is the 50th percentile
Q3 is the 75th percentile

Technology and statistics - the perfect combination!

Excel:	Easy entry, powerful capabilities, can see/graph/save data
Calculators:	Increasingly powerful, portable, TI-nSpire!
Software:	Readily available, some web based, very sophisticated niche markets (Geo-statistics)

Consider the data set:

5 2 3 3 6 4 5 3 7 5 7 1 8 9 5

One dimensional statistics - enter in L_1

Calculate the 5 number summary

Box & Whisker

Histogram/column graph

SL: 14D #1-7 odd
14E #1-3

F VARIANCE AND STANDARD DEVIATION

The 5 number summary:

+ Easy to calculate
+ A lot of info for not much calculation
- Information can be hidden within a quartile

Consider calculating the distance of each data point from the mean $x - \bar{x}$

This is sometimes referred to as the "***error***" and would be a good measure of variation if we sum them. But the sum would always be zero (can you prove that?)

But if we square these differences they will sum OK. To correct for the number of values we can divide by that number and define:

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

Some characteristics of the variance:

The differences being summed are a measure of how far each value deviates from the mean.

By squaring them, we make them all positive and weight larger errors more.

Dividing by *n* finds how far *on average* the data is from the mean.

If the sum in the numerator is small, most of the values are close to the mean.

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

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

Some characteristics of the standard deviation:

The units are the same as for the data (handy)

A few large "errors" can change it significantly since the difference from the mean is squared.

Therefore is is called a ***non-resistant*** measure of spread.

It's good for symmetric data. IQR and percentiles are better for skewed data.

Populations vs. Samples

Notation for these measures is different to help distinguish:

	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s_n^2
Standard Deviation	σ	s_n

We draw inferences from ***sample statistics*** to ***estimate population parameters***.

When a sample of size n is used to draw inference about a population:

- the mean of the sample $\bar{x}$ is an unbiased estimate of μ
- $s_{n-1}^2 = \left(\frac{n}{n-1}\right) s_n^2$ is an unbiased estimate of the variance σ^2.

Note: Even if s_{n-1}^2 is an unbiased estimate of σ^2, this does not imply that s_{n-1} is an unbiased estimate of σ.

$$s_{n-1}{}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \cdot \left(\frac{n}{n-1}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Why?: Applet

Why?: Video Proof

Why?: The values in the sample will generally be closer to the sample mean than they are to the population mean so the sum on top is a bit smaller than it would be if we know the population mean. To correct for that, we use *n* - 1 in the denominator.

This is known as ***Bessel's Correction***.
For more see: http://en.wikipedia.org/wiki/Bessel%27s_correction

An ***unbiased*** estimate of the population ***variance*** uses *n* - 1 in the denominator! Know which one you're using!

Important: On your calculator, the value S_x represents the unbiased standard deviation of a population assuming that the values in the list represent only a sample of that population. In other words, it uses *n*-1 in the denominator. The value calculated in σ_x uses *n* in the denominator assuming that your list is the entire population.

Standard Deviation for Summarized (grouped) data

Value (x)	Frequency (f)	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
1	2	-2	4	8
2	5	-1	1	5
3	6	0	0	0
4	5	1	1	5
5	2	2	4	8
Total	20			26

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{60}{20} = 3$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i}} \text{ or } \sqrt{\frac{\sum f(x_i - \bar{x})^2}{\sum f}} = \sqrt{\frac{26}{20}} = \sqrt{\frac{13}{10}} \approx 1.14$$

SL: **14F.1** **#1, 3, 5-9**
14F.2 **#2**
14F.3 **#3, 4, 5**

Simplifying the calculation of variance

Notice that we can rewrite the calculation of a sample variance as shown below. This generally makes the calculation simpler.

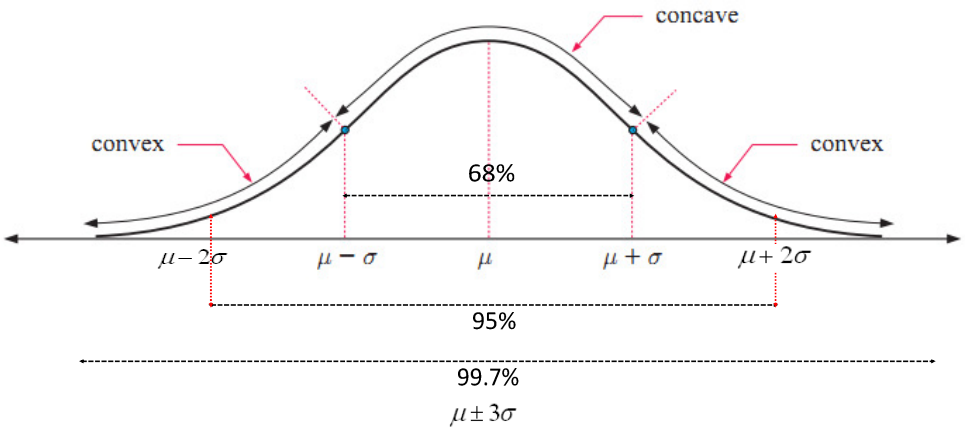
$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$	Start with the definition and expand the square.
$s_n^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{\sum_{i=1}^{n} 2x_i\bar{x}}{n} + \frac{\sum_{i=1}^{n} \bar{x}^2}{n}$	Summation distributes over terms (could you prove that?)
$s_n^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{2\bar{x}\sum_{i=1}^{n} x_i}{n} + \frac{\bar{x}^2\sum_{i=1}^{n} 1}{n}$	Constants can be factored out of a summation. 2 and $\bar{x}$ are both constants. Why do we need to leave $\sum_{i=1}^{n} 1$ in the third numerator and what does it mean?
$s_n^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - 2\bar{x}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) + \frac{n\bar{x}^2}{n}$	Rewrite more simply
$s_n^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2$	The factor in parentheses is just the mean.
$s_n^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$	Combine terms for a simpler form.

Standard Deviation and the Normal Distribution

The classic symmetric, bell shaped distribution is known as the ***Normal Distribution***

bell-shaped
distribution

More in Chapter 29
Connections between standard deviation and the normal distribution:



The first σ from the mean represents the inflection points of the curve.
About 68% of the data fall within 1σ of the mean.
About 95% of the data fall within 2σ of the mean.
About 99.7% of the data fall within 3σ of the mean.

SL: 14G #1, 3, 4