"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:

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

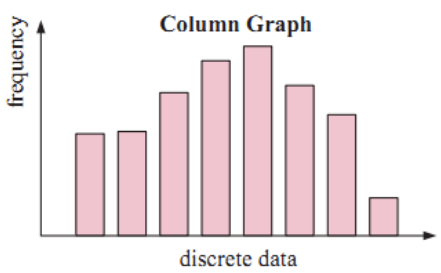
• **Outliers**	Data values that are either much larger or much smaller than the general body of data. They should be included in an analysis *unless* they are the result of human or other error.

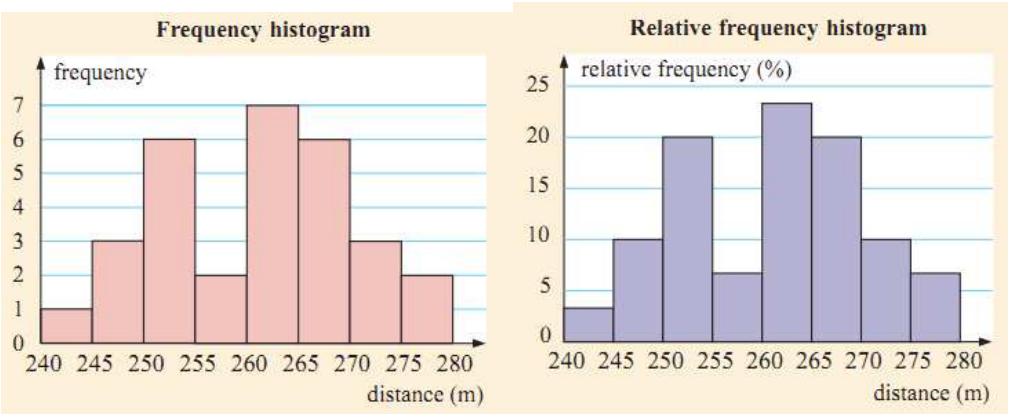
A **discrete numerical variable** takes exact number values and is often a result of **counting**.

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

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

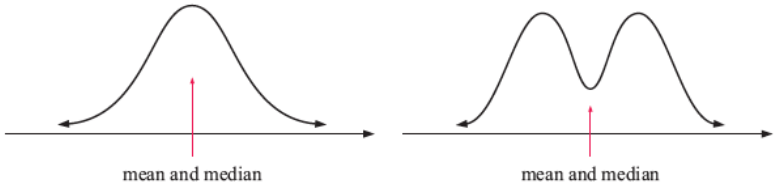
Mode:

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

Which is best? Mean, Median or Mode:

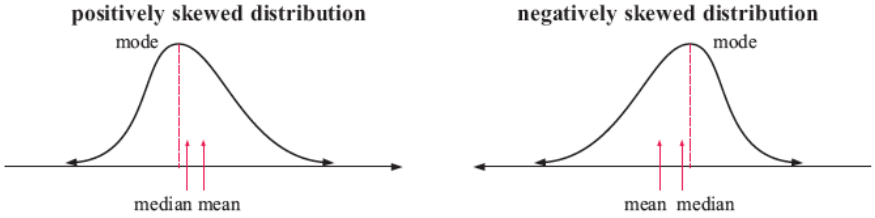
Consider 3,3,3,3,3,10000
1,1,1,1,5,5,5,800
1,1,1,1,2,12,13,13,13

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.

14A: Opener, #3,4 (Vocabulary, histograms, frequency tables, stem and leaf)
14B.1: #3,5,8,9,11,12,14,15,16 (Mean, median, mode, skew, outlier, bimodal)

Present: Questions from opener
14A: #3, Column graph vs Freq Histogram, #4
14B.1: 5,8,9,11,12,14,15,16

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

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***

Important Note: When there are an odd number of numbers, ***do not*** include the median in the upper "half" of the data values. See the next example.

For the data set: 7, 3, 1, 7, 6, 9, 3, 8, 5, 8, 6, 3, 7, 1, 9

a median **b** lower quartile **c** upper quartile **d** interquartile range

The ordered data set is:

~~1, 1, 3, 3, 3, 5, 6,~~ 6, ~~7, 7, 7, 8, 8, 9, 9~~ (15 of them)

a As $n = 15$, $\frac{n+1}{2} = 8$ $\therefore$ the median = 8th data value = 6

b/c As the median is a data value we now ignore it and split the remaining data into two:

lower: 1 1 3 3 3 5 6 upper: 7 7 7 8 8 9 9

Q_1 = median of lower half = 3
Q_3 = median of upper half = 8

d IQR $= Q_3 - Q_1 = 8 - 3 = 5$

For the data set: 6, 4, 9, 15, 5, 13, 7, 12, 8, 10, 4, 1, 13, 1, 6, 4, 5, 2, 8, 2 find:

a the median **b** Q_1 **c** Q_3 **d** the interquartile range

The ordered data set is:

~~1 1 2 2 4 4 4 5 5~~ 6 6 ~~7 8 8 9 10 12 13 13 15~~ (20 of them)

a As $n = 20$, $\frac{n+1}{2} = 10.5$

$\therefore$ median $= \frac{\text{10th value} + \text{11th value}}{2} = \frac{6+6}{2} = 6$

b/c As we have an even number of data values, we split the data into two:

lower: 1 1 2 2 4 4 4 5 5 6 upper: 6 7 8 8 9 10 12 13 13 15

$\therefore$ $Q_1 = \frac{4+4}{2} = 4$, $Q_3 = \frac{9+10}{2} = 9.5$

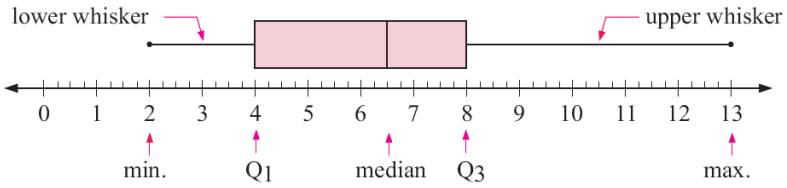
d IQR $= Q_3 - Q_1$
$= 9.5 - 4$
$= 5.5$

Some computer packages calculate quartiles differently.

Your TI-83/84 is fine.

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

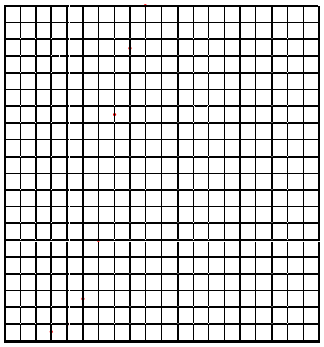
Your calculator will do them!

14B.2: #2,4,5,7,8,10 (Cumulative frequency, summarized data)
14B.3: #2,3 (Use midpoint of range for data in ranges (classes))
14C.1: #2,4,5 (Range, quartiles, IQR)
14C.2: #2,3,6 (Box & Whisker)

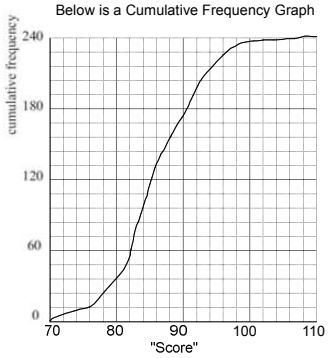
D CUMULATIVE FREQUENCY GRAPHS

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

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	

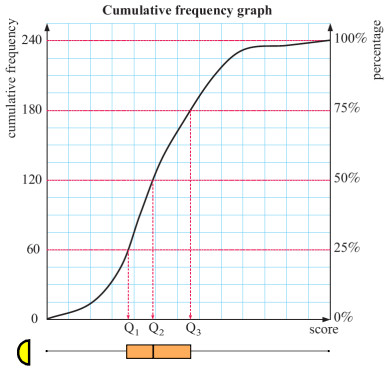


Below is a Cumulative Frequency Graph



What questions can we answer from it?

Can we find the 5 number summary from it?



Understand this well...it's something IB stresses....

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

In practice: Minimize the number of graph "reads" - maximize the number of calculations to ensure that totals match known information. (This also shows work for points!)

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

TI-84 Plus Steps

1. Enter data values in L_1 : [STAT]/[EDIT]
2. Enter frequency counts in L_2 if applicable: [STAT]/[EDIT]
3. Calculate one variable statistics
   > [STAT]/[CALC]/[1-VAR-STATS]
   > Enter L_1 (and L_2 if frequencies are used)
4. Calculates:
   > Mean, Σx, Σx^2, n, and 5 number summary
5. Graph as histogram or box and whisker plot

14D: #1-7 Odd (Cumulative frequency graphs, percentiles)
14E: #1,2,3 (Using technology)
QB: #1,2,4,5 (IB Practice)

Present: 14D: #3, 5 & 14E: #2, QB #1,2,4,6

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_n uses *n* in the denominator assuming that your list is the entire population.

14F.1: #1,3,5-9 (Variance, std. dev. #1 by hand, calculator for others.)
14F.2: #2 (Sample vs. population, statistics vs. parameters. σ_n vs. S_n.)
QB: #7,9,10,12,14 (not e ii) (IB Practice)

Present: 14F.1#3,6,8 QB #9,10,12,14

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

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

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

Again, your calculator will do this for you if you enter a value list and a frequency list.

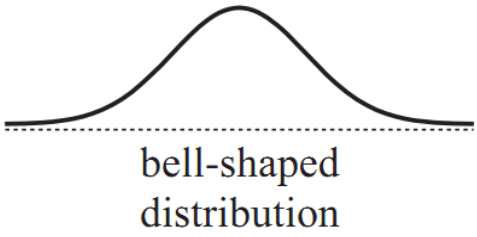
Simplifying the calculation of variance

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

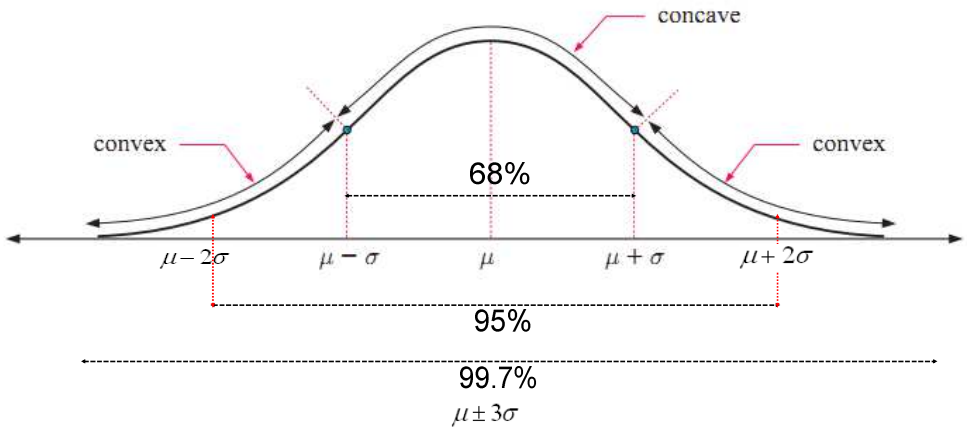
$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$	Start with the definition and expand the square.
$s_n^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{\sum_{i=1}^{n} 2x_i\bar{x}}{n} + \frac{\sum_{i=1}^{n} \bar{x}^2}{n}$	Summation distributes over 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*



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.

14F.3: #3,4,5 (Grouped data)
14G: #1,3,4 S.D. and the Normal Distribution
QB: #3,6,7,9,13 (IB Practice)

Present QB #3,5,7,9

10 Bivariate analysis

Investigation – leaning tower of Pisa

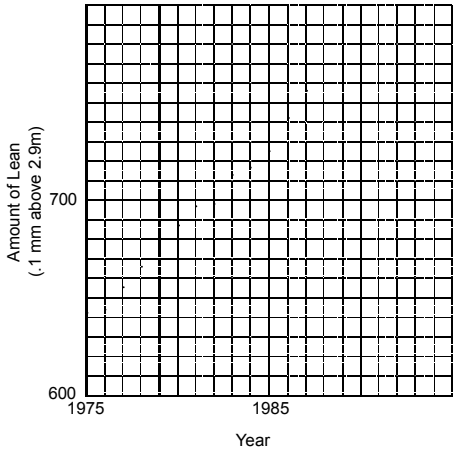
The bell tower of Pisa cathedral was built in 1178 and soon began leaning to one side – hence its name. The measurements below show the lean in tenths of a millimetre beyond 2.9 metres. So in 1975 the tower was leaning 2.9642 metres from the vertical.

Year	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
Lean	642	644	656	667	673	688	696	698	713	717	725	742	757

Does it look like the lean of the tower is increasing with time?
If so, how fast is the lean increasing with time?
Is there evidence that the lean changes significantly with time?
Is there an approximate formula for calculating the lean?
Can you predict the lean in the future?

➔ Bivariate analysis is concerned with the relationships between pairs of variables (x, y) in a data set.

The first tool, is to try to visualize the data with a ***scatter plot***.

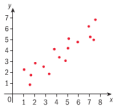


Some conventions:

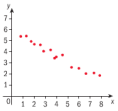
1. The ***dependent*** variable is the one that *depends on* the other one. It should be plotted on the *y*-axis (vertical)
2. The ***independent*** variable is the one that *controls* the other one. It should be plotted on the *x*-axis (horizontal)

The relationship between the variables is called ***correlation***.

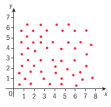
➔ A general upward trend in the pattern of dots shows **positive** correlation.



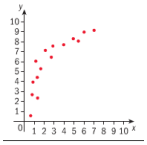
➔ Scattered points with no trend may indicate correlation close to **zero**.



➔ A general downward trend in the pattern of dots shows **negative** correlation.



Correlation does not imply a linear relationship.
The data to the right is positively correlated, but not linear.

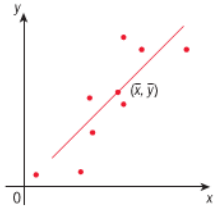


Correlation also does not imply ***causation***. For children younger than 15, their height is positively correlated with the number of words in their vocabulary but height alone clearly does not ***cause*** students to have a larger vocabulary. Be careful about confusing these.

Much of practical mathematics is about ***modeling*** data with a smooth function that approximates the behavior of the data. For data that appears to have a linear relationship, we can find a ***line of best fit***. Such a line is also called a ***regression line***.

To begin, we can estimate such a line by finding the ***mean point***. It is calculated as the mean of the *x* values and the mean of the *y* values and is written as $(\bar{x},\bar{y})=\left(\frac{\sum x}{n},\frac{\sum y}{n}\right)$

It makes sense that a line through the data should go through this point:



But we need to find a slope for the line. One way, is to just estimate it so that the data has approximately the same amount of "weight" above and below the line.

Once we have the equation of the line, we can predict values for data that we did not collect.

- Predicting values ***outside the data set*** is called ***extrapolation***. Be careful about this as the line may not have any meaning beyond certain values.
- Predicting values ***between values in the data set*** is called ***interpolation***.

In either case, we can evaluate the equation of the line at any point to predict a result.

Miss Lincy's 10 students' scores, out of 100, for their classwork and final exam are shown below.

Student	Ed	Craig	Uma	Phil	Jenny	James	Ron	Bill	Caroline	Steve
Classwork	95	66	88	75	90	82	50	45	80	84
Final	95	59	85	77	92	70	40	50	Abs	80

Caroline was absent for the final. Do not include her grades in finding the mean point.

a Find the mean classwork score.
b Find the mean final exam score.
c Construct a scatter diagram and draw a line of best fit through your mean point.
d Find the equation of the regression line.
e Use the equation of the regression line to estimate Caroline's score for the final exam.

Given the line, it is important to be able to interpret it:

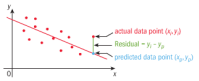
A biologist wants to study the relationship between the number of trees *x* per hectare and the number of birds *y* per hectare.
She calculates the equation of the regression line to be $y = 8 + 5.4x$.
State the gradient and the *y*-intercept and interpret them.

Answer

The slope is 5.4. This means that for every additional tree, you can expect an average of 5.4 additional birds per hectare.
The *y*-intercept is 8, which means that an area with no trees averages 8 birds per hectare.

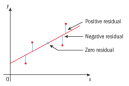
Let's get more precise about finding the ***best*** fit line (or, more generally, the best fit curve).

Consider the following. Each data point differs from the given line of best fit by some amount.



The ***residual*** or ***error*** is the vertical distance between a data point and the graph of a regression equation.

Notice that residuals can be positive or negative (or zero!)



So now we can see that we can put a number on the quality of the fit by summing up the sizes of the residuals. A regression equation that results in a small total residual will represent a better fit.

Can we just add the residuals? NO! Because some are positive and some are negative, they would cancel if we did that. So instead we sum the ***squares of the residuals***. This has two effects:

- All the residuals contribute a positive amount to the total.
- Larger residuals are weighted more heavily that smaller ones (since they are squared)

The result is a value called the total squared error.

Using summation notation, the total squared error is given by $\sum_{i=1}^{n}\left[x_i - f(x_i)\right]^2$

where $f(x)$ is the regression line in question. One way to define the "best" fit regression equation is to find the one that has the least total squared error. That is:

Least Squares Regression

The function $f(x)$ that results in the least total squared error is called the ***least squares regression function***. For linear equations we sometimes call this the ***least square line***.

For a linear fit, the slope of the least squares regression line can be calculated from the data values using a rather complex formula:

Slope of the Least Squares Regression Line

The ***slope*** of the ***least squares regression line*** for a set of points $\{x_i, y_i\}_{i=1}^{n}$ is given by:

$$m = \frac{s_{xy}}{(s_x)^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

This looks complex and a thorough development is beyond this course. Suffice it to say that S_x is related to the standard deviation of x and S_{xy} is a similar formula that uses both x and y.

But back to the line of best fit. We now have a formula for its slope and we previously found the mean point so we can use point slope form to create...

Equation of the Least Squares Regression Line

The ***equation*** of the ***least squares regression line*** for a set of points $\{x_i, y_i\}_{i=1}^{n}$ is given by:

$$y - \bar{y} = m(x - \bar{x})$$

Where $(\bar{x}, \bar{y}) = \left(\frac{\sum x}{n}, \frac{\sum y}{n}\right)$ and $m = \frac{s_{xy}}{(s_x)^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$

Just for the heck of it, let's see this in its full glory in slope intercept form:

$$y = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} x + \frac{\sum y}{n} - \left(\frac{\sum x}{n}\right)\frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

The good news, is that all of this can be calculated on your TI-84:

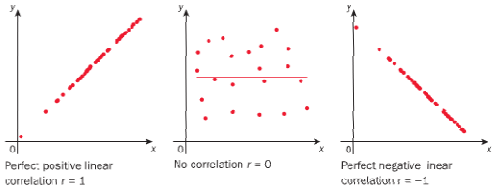
Finding Best Fit Curves on TI-84

1. Enter x values in L_1 and y values in L_2.
2. Set CATALOG/DiagOn if you want ***r*** values
3. Use STAT/CALC and select the kind of curve you desire:
   - > LinReg (ax + b) - Linear
   - > QuadReg - Quadratic
   - > CubicReg - Cubic
   - > QuartReg - Quartic
   - > LinReg (a + bx) - Linear
   - > LnReg - Logarithmic
   - > ExpReg - Exponential ($y = ab^x$)
   - > PwrReg - Power function ($y = ax^b$)
   - > Logistic - Logistic (growth that levels out)
   - > SinReg - Siunsoidal
4. The ***parameters*** of the best fit curve will be stored in VARS/STATISTICS/EQ
   - > RegEQ has the equation itself which you can paste into a Y= function to graph it.

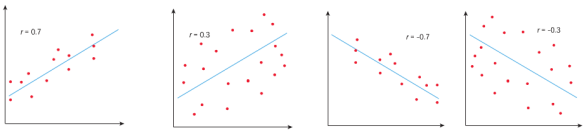
We now understand how to find the equation of best fit - for lines or other functions. Understand what the calculator is doing, but let it do the work for you.

Measure of correlation - the ***Pearson product-moment correlation coefficient***

The ***r*** value that the calculator provides is a measure of the correlation that describes the strength of ***linear*** dependence between two variables. The number ranges between -1 and one as follows:



Notice that ***r*** is not the slope of the line, although the sign of ***r*** is the same as the sign of the slope. ***r*** is a measure of ***how scattered*** the data are from a straight line.



The formula for calculating ***r*** is similar to that of the slope of the best fit line:

Pearson's Correlation Coefficient

The ***Pearson's correlation coefficient*** for a set of points $\{x_i, y_i\}_{i=1}^{n}$ is given by:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

Again, you will generally calculate this on a GDC. Here is a summary for interpretation:

r-value	Correlation
$0 < \lvert r \rvert \le 0.25$	Very weak
$0.25 < \lvert r \rvert \le 0.5$	Weak
$0.5 < \lvert r \rvert \le 0.75$	Moderate
$0.75 < \lvert r \rvert \le 1$	Strong

Try one on your calculator:

Sue wants to determine the strength of the correlation between the number of spoons of plant food she uses and the extra number of orchids grown from a plant. Use Pearson's correlation coefficient formula to interpret the relationship.

Plant	Spoons of plant food *x*	Increase in the number of orchids *y*
A	1	2
B	2	3
C	3	8
D	4	7

Plant	x	y	xy	x^2	y^2
A	1	2	2	1	4
B	2	3	6	4	9
C	3	8	24	9	64
D	4	7	28	16	49
Total	10	20	60	30	126

n=4

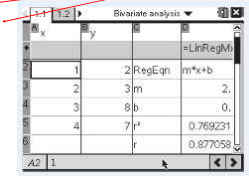
Σx

Σy

Σxy

Σy^2

Σx^2



Two approaches: LinReg or 2-Var Stats. Only LinReg does *r*.

HW: Handout: Regression and Correlation