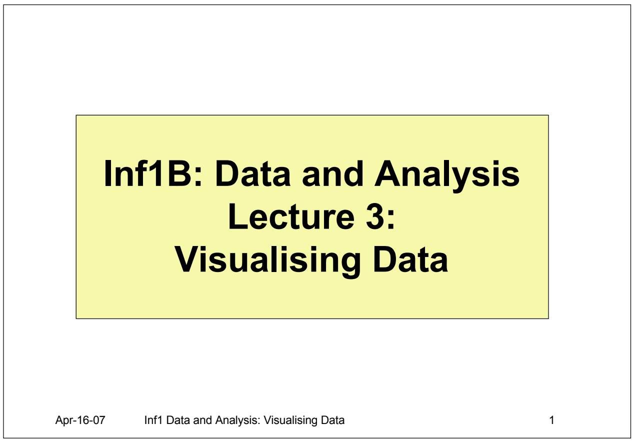
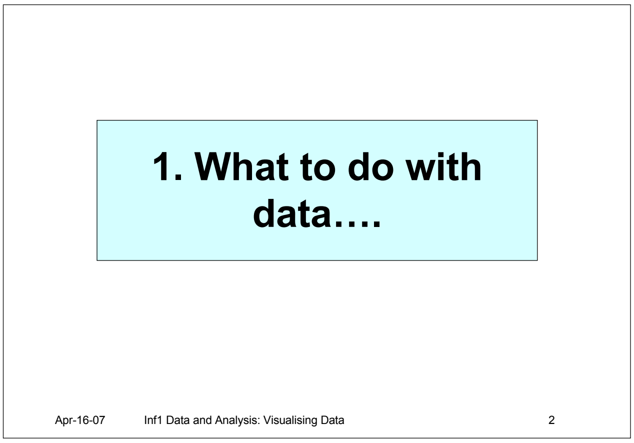
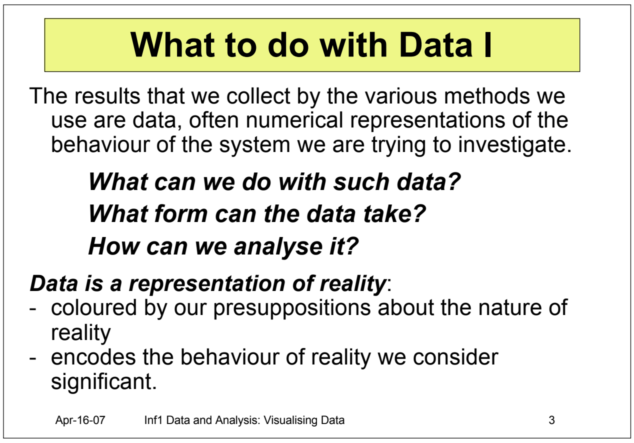


Inf1B: Data and Analysis Lecture 3: Visualising Data



1. What to do with data….



What to do with Data I

The results that we collect by the various methods we use are data, often numerical representations of the behaviour of the system we are trying to investigate.

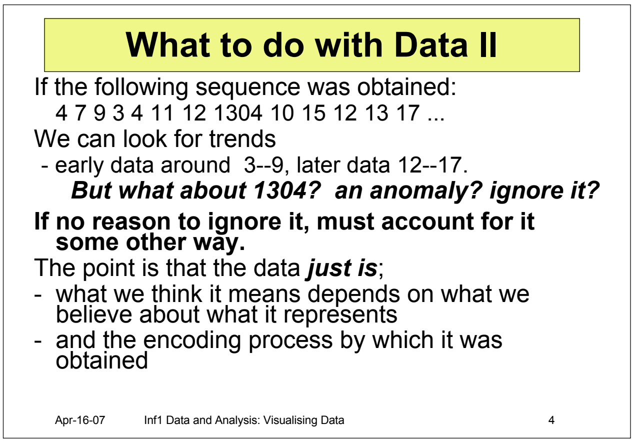
What can we do with such data?

What form can the data take?

How can we analyse it?

Data is a representation of reality:

- coloured by our presuppositions about the nature of reality
- encodes the behaviour of reality we consider significant.



What to do with Data II

If the following sequence was obtained:

4 7 9 3 4 11 12 1304 10 15 12 13 17 ...

We can look for trends

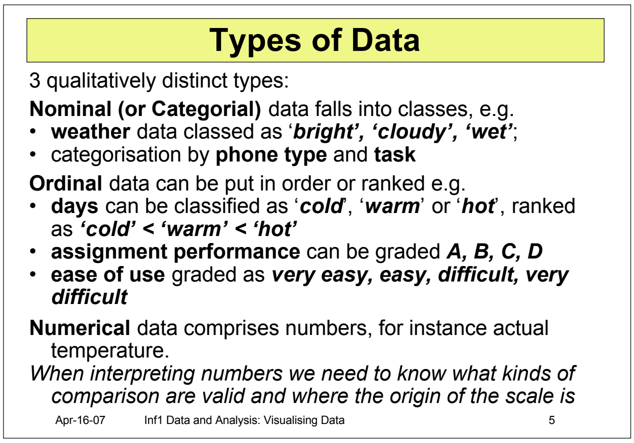
- early data around 3--9, later data 12--17.

But what about 1304? an anomaly? ignore it?

If no reason to ignore it, must account for it some other way.

The point is that the data ***just is***;

- what we think it means depends on what we believe about what it represents
- and the encoding process by which it was obtained



Types of Data

3 qualitatively distinct types:

Nominal (or Categorial) data falls into classes, e.g.

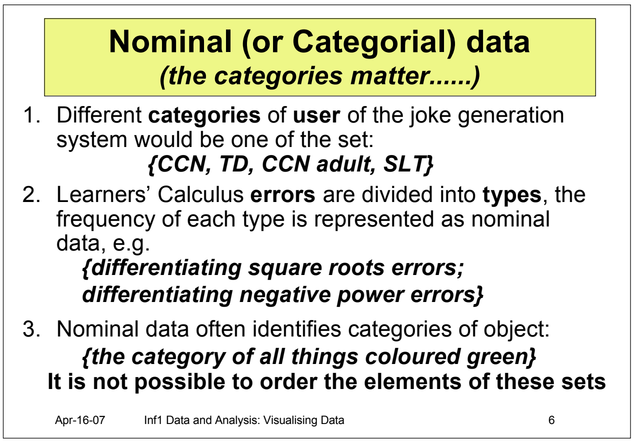
- **weather** data classed as '***bright', 'cloudy', 'wet'***;
- categorisation by **phone type** and **task**

Ordinal data can be put in order or ranked e.g.

- **days** can be classified as '***cold***', '***warm***' or '***hot***', ranked as ***'cold' < 'warm' < 'hot'***
- **assignment performance** can be graded ***A, B, C, D***
- **ease of use** graded as ***very easy, easy, difficult, very difficult***

Numerical data comprises numbers, for instance actual temperature.

When interpreting numbers we need to know what kinds of comparison are valid and where the origin of the scale is

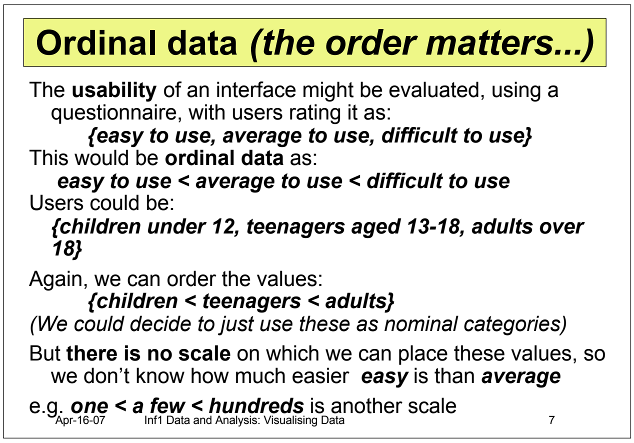


Nominal (or Categorial) data

(the categories matter......)

1. Different **categories** of **user** of the joke generation system would be one of the set:
   {CCN, TD, CCN adult, SLT}
2. Learners' Calculus **errors** are divided into **types**, the frequency of each type is represented as nominal data, e.g.
   {differentiating square roots errors; differentiating negative power errors}
3. Nominal data often identifies categories of object:
   {the category of all things coloured green}

It is not possible to order the elements of these sets



Ordinal data *(the order matters...)*

The **usability** of an interface might be evaluated, using a questionnaire, with users rating it as:

{easy to use, average to use, difficult to use}

This would be **ordinal data** as:

easy to use < average to use < difficult to use

Users could be:

{children under 12, teenagers aged 13-18, adults over 18}

Again, we can order the values:

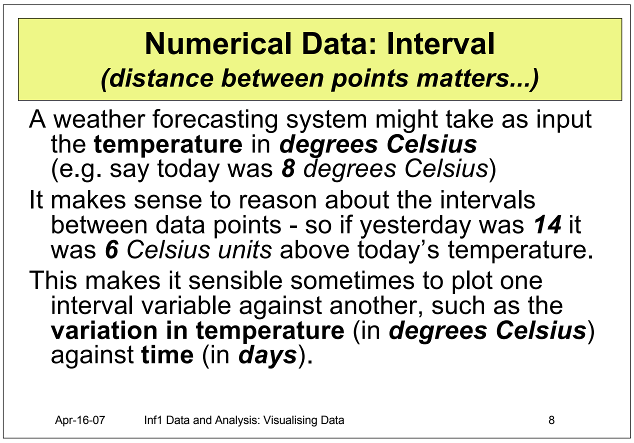
{children < teenagers < adults}

(We could decide to just use these as nominal categories)

But **there is no scale** on which we can place these values, so we don't know how much easier ***easy*** is than ***average***

e.g. ***one < a few < hundreds*** is another scale

Apr-16-07 Inf1 Data and Analysis: Visualising Data 7



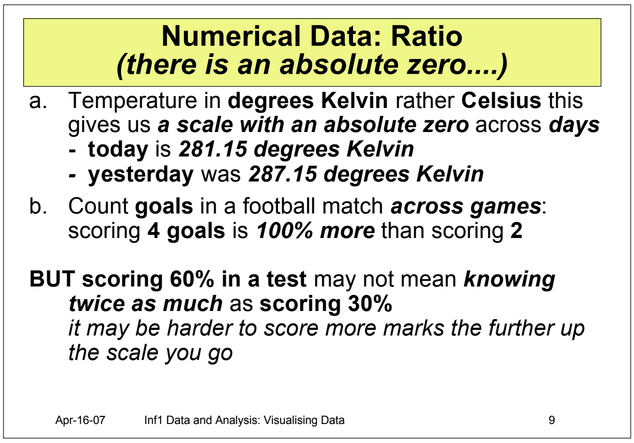
Numerical Data: Interval
(distance between points matters...)

A weather forecasting system might take as input the **temperature** in ***degrees Celsius*** (e.g. say today was ***8*** *degrees Celsius*)

It makes sense to reason about the intervals between data points - so if yesterday was ***14*** it was ***6*** *Celsius units* above today's temperature**.**

This makes it sensible sometimes to plot one interval variable against another, such as the **variation in temperature** (in ***degrees Celsius***) against **time** (in ***days***).

Apr-16-07 Inf1 Data and Analysis: Visualising Data 8



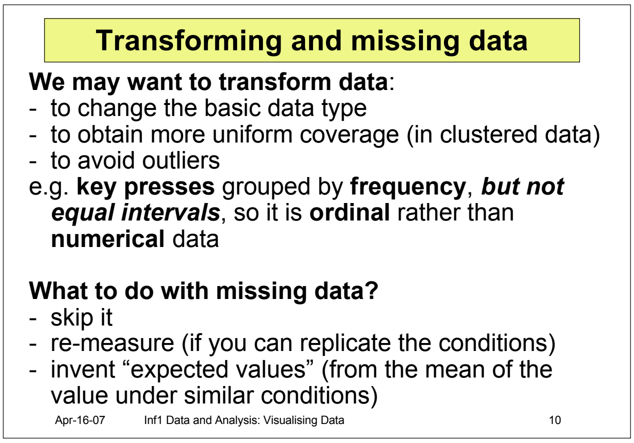
Numerical Data: Ratio
(there is an absolute zero....)

a. Temperature in **degrees Kelvin** rather **Celsius** this gives us ***a scale with an absolute zero*** across ***days***
   - **today** is ***281.15 degrees Kelvin***
   - **yesterday** was ***287.15 degrees Kelvin***

b. Count **goals** in a football match ***across games***: scoring **4 goals** is ***100% more*** than scoring **2**

BUT scoring 60% in a test may not mean ***knowing twice as much*** as **scoring 30%**
it may be harder to score more marks the further up the scale you go

Apr-16-07 Inf1 Data and Analysis: Visualising Data 9



Transforming and missing data

We may want to transform data:

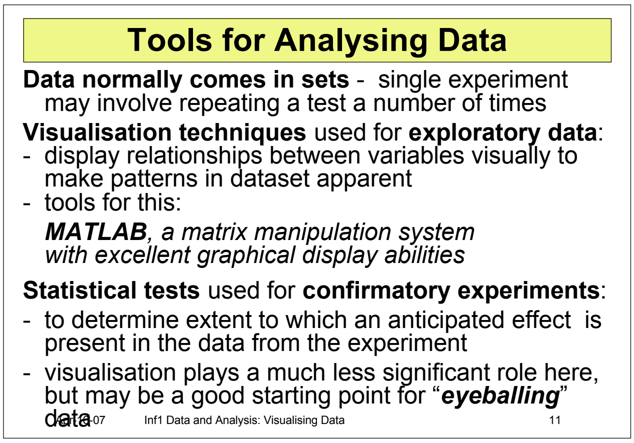
- to change the basic data type
- to obtain more uniform coverage (in clustered data)
- to avoid outliers

e.g. **key presses** grouped by **frequency**, ***but not equal intervals***, so it is **ordinal** rather than **numerical** data

What to do with missing data?

- skip it
- re-measure (if you can replicate the conditions)
- invent "expected values" (from the mean of the value under similar conditions)

Apr-16-07 Inf1 Data and Analysis: Visualising Data 10



Tools for Analysing Data

Data normally comes in sets - single experiment may involve repeating a test a number of times

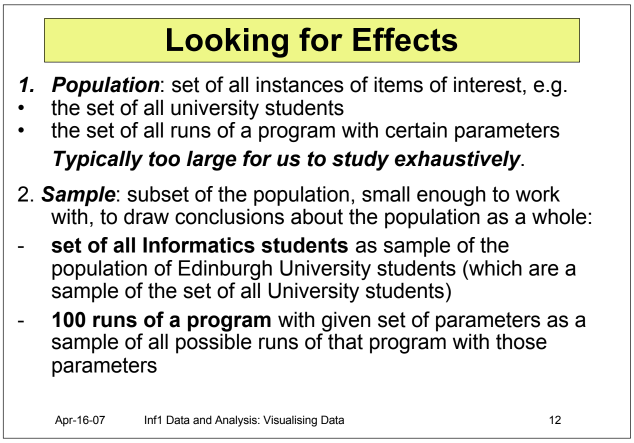
Visualisation techniques used for **exploratory data**:

- display relationships between variables visually to make patterns in dataset apparent
- tools for this:
  MATLAB*, a matrix manipulation system with excellent graphical display abilities*

Statistical tests used for **confirmatory experiments**:

- to determine extent to which an anticipated effect is present in the data from the experiment
- visualisation plays a much less significant role here, but may be a good starting point for "***eyeballing***" data

Apr-16-07 Inf1 Data and Analysis: Visualising Data 11



Looking for Effects

1. Population: set of all instances of items of interest, e.g.

- the set of all university students
- the set of all runs of a program with certain parameters

Typically too large for us to study exhaustively.

2. ***Sample***: subset of the population, small enough to work with, to draw conclusions about the population as a whole:

- **set of all Informatics students** as sample of the population of Edinburgh University students (which are a sample of the set of all University students)
- **100 runs of a program** with given set of parameters as a sample of all possible runs of that program with those parameters

Apr-16-07 Inf1 Data and Analysis: Visualising Data 12

Sampling

Sample must be representative of the population it is taken from

e.g. ***Informatics students*** not representative of ***students at Edinburgh University***

But a ***random*** sample of ***20 Inf1B students*** could be ***representative of the Inf1 class***

How is the sample selected?

- must ensure a representative sample, for the purposes of the experiment at hand

How large must the sample be?

- the larger the sample the more work is involved but the more secure the conclusions will be

Statistic(s)

A ***statistic*** is a ***numerical encoding of some property of a population or sample***, hopefully characteristic of that population or sample, which we can use instead of the sample for reasoning about the properties of the sample.

e.g. **the average age of students taking Inf1** summarizes certain properties of the population of such students.

Statistics is the subject that studies the properties of such encodings

Summary statistics

Summary Statistics express a property of the data set in a single number or set of a few numbers.

Most common are ***mean***, ***mode*** (most frequent value), ***median*** (middle value), ***variance*** and ***standard deviation***

Mean (μ) gives the centre of mass of the set:

$$\mu = \frac{\sum \{x1,\ldots xn\}}{n}$$

So mean of {2, 3, 6, 1, 5, 1} = 18/6 = 3

Summary statistics

{6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6}

Mean = $\mu = \dfrac{\sum \{x1,\ldots xn\}}{n}$

= (6+ 2+ 3+ 6+1+5+1+7+2+5+6)/11

= 44/11 = **4**

Median = middle value

1,1,2,2,3,5,5,6,6,6,7 = **5**

Mode = most frequent value = **6**

Variance

Variance is the mean deviation from the centre:

$$\text{Variance} = \frac{\sum \{(x1 - \mu)\text{\textasciicircum}2,\ldots(xn - \mu)\text{\textasciicircum}2\}}{N}$$

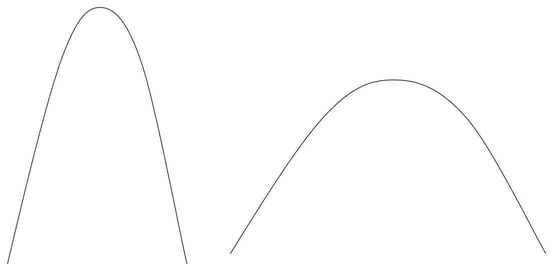
$$= (\sum (X - \mu)\text{\textasciicircum}2)/N$$

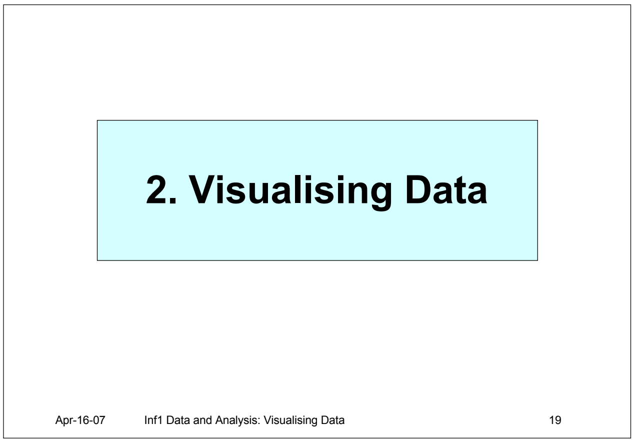
The Standard deviation σ is the square root of the variance:

$$\sigma = \sqrt{(\,(\sum (X - \mu)\text{\textasciicircum}2)/N\,)}$$

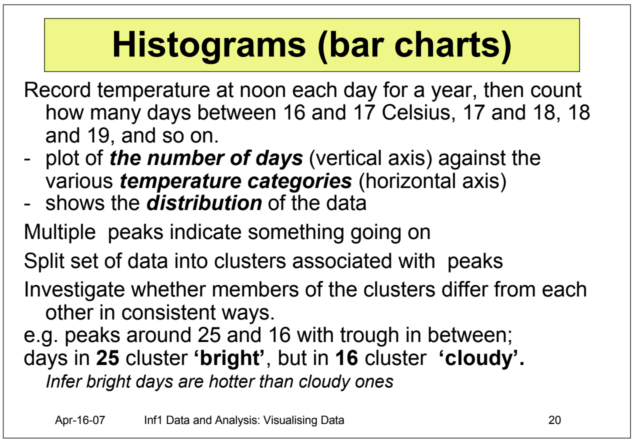
Normal distributions

Left curve shows less variance than right curve





2. Visualising Data



Histograms (bar charts)

Record temperature at noon each day for a year, then count how many days between 16 and 17 Celsius, 17 and 18, 18 and 19, and so on.

- plot of ***the number of days*** (vertical axis) against the various ***temperature categories*** (horizontal axis)
- shows the ***distribution*** of the data

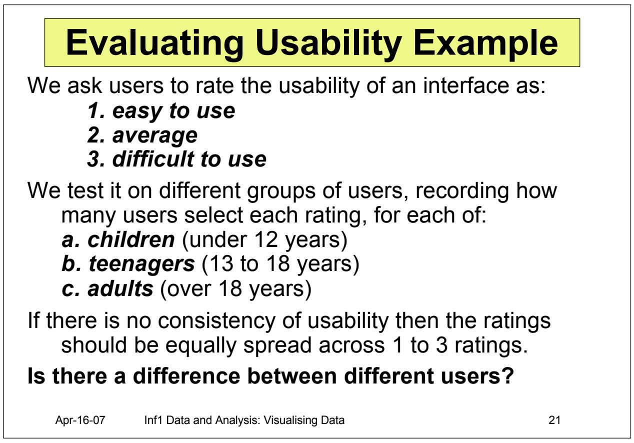
Multiple peaks indicate something going on

Split set of data into clusters associated with peaks

Investigate whether members of the clusters differ from each other in consistent ways.

e.g. peaks around 25 and 16 with trough in between;
days in **25** cluster **'bright'**, but in **16** cluster **'cloudy'.**

Infer bright days are hotter than cloudy ones



Evaluating Usability Example

We ask users to rate the usability of an interface as:

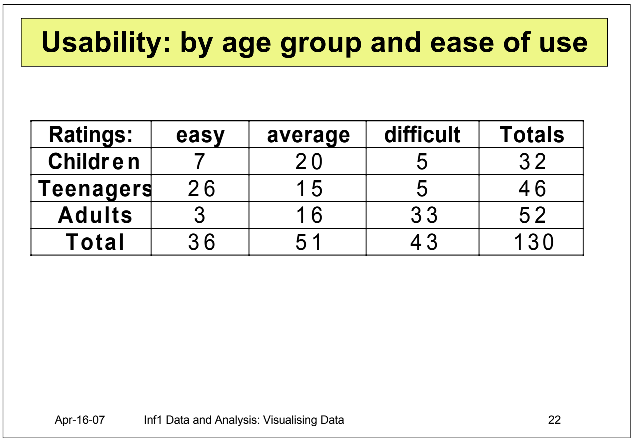
1. easy to use
2. average
3. difficult to use

We test it on different groups of users, recording how many users select each rating, for each of:

a. children (under 12 years)
b. teenagers (13 to 18 years)
c. adults (over 18 years)

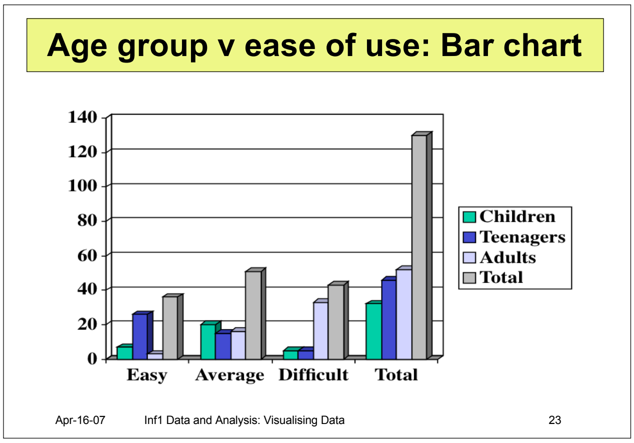
If there is no consistency of usability then the ratings should be equally spread across 1 to 3 ratings.

Is there a difference between different users?

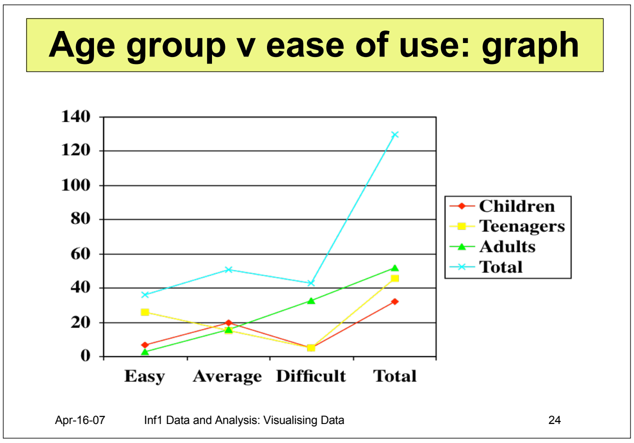


Usability: by age group and ease of use

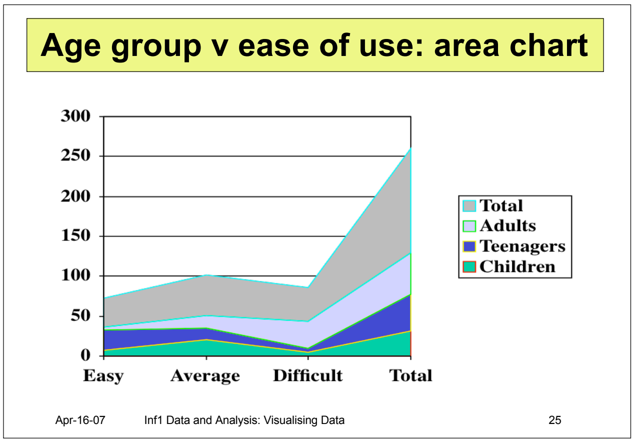
Ratings:	easy	average	difficult	Totals
Children	7	20	5	32
Teenagers	26	15	5	46
Adults	3	16	33	52
Total	36	51	43	130



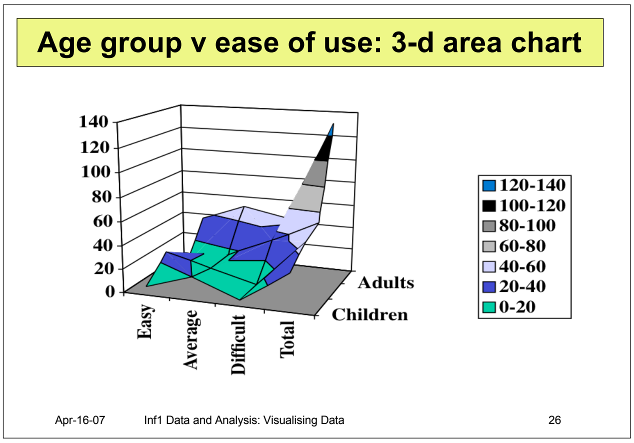
Age group v ease of use: Bar chart



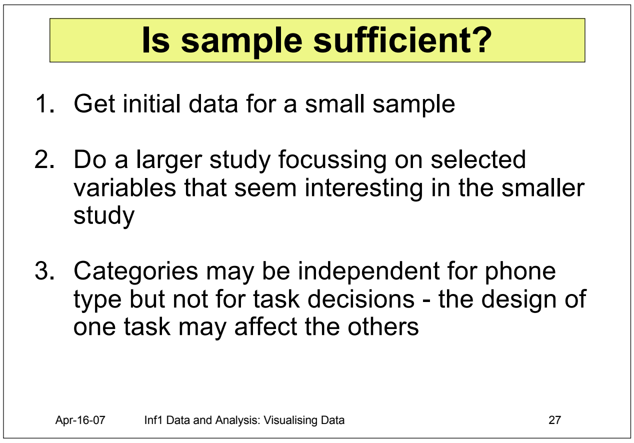
Age group v ease of use: graph



Age group v ease of use: area chart

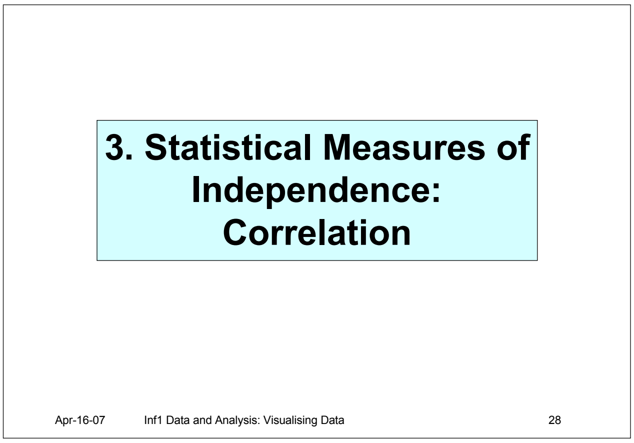


Age group v ease of use: 3-d area chart

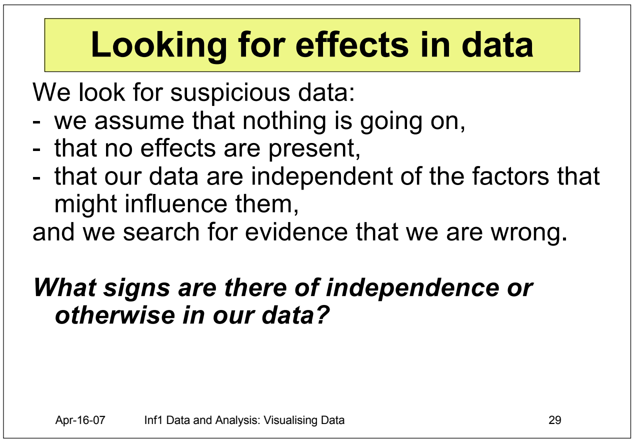


Is sample sufficient?

1. Get initial data for a small sample
2. Do a larger study focussing on selected variables that seem interesting in the smaller study
3. Categories may be independent for phone type but not for task decisions - the design of one task may affect the others



3. Statistical Measures of Independence: Correlation

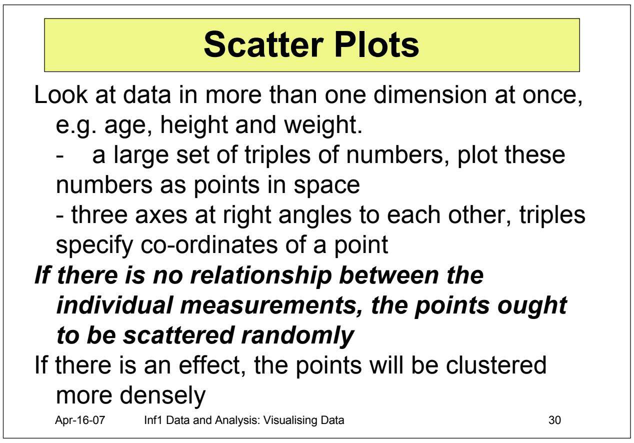


Looking for effects in data

We look for suspicious data:
- we assume that nothing is going on,
- that no effects are present,
- that our data are independent of the factors that might influence them,

and we search for evidence that we are wrong.

What signs are there of independence or otherwise in our data?



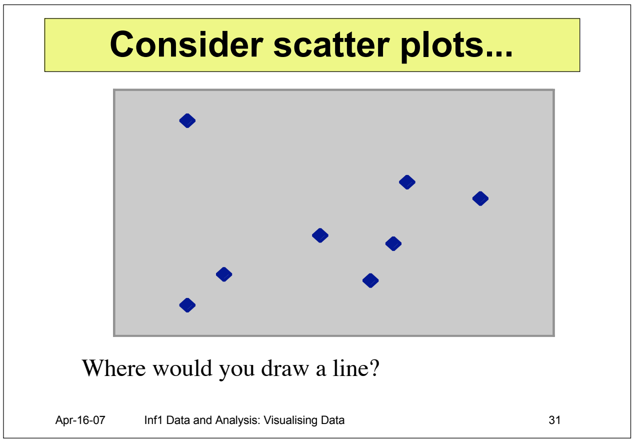
Scatter Plots

Look at data in more than one dimension at once, e.g. age, height and weight.

- a large set of triples of numbers, plot these numbers as points in space
- three axes at right angles to each other, triples specify co-ordinates of a point

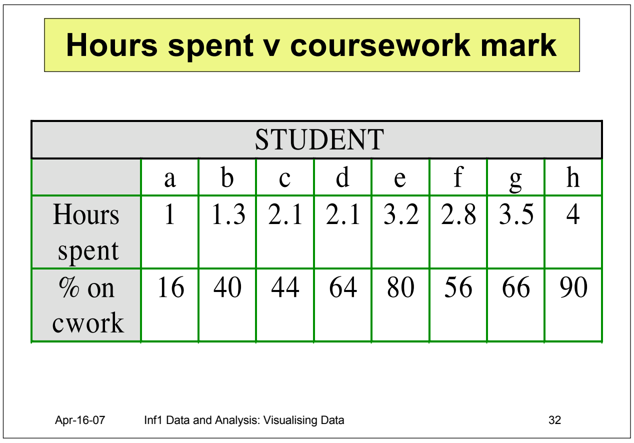
If there is no relationship between the individual measurements, the points ought to be scattered randomly

If there is an effect, the points will be clustered more densely



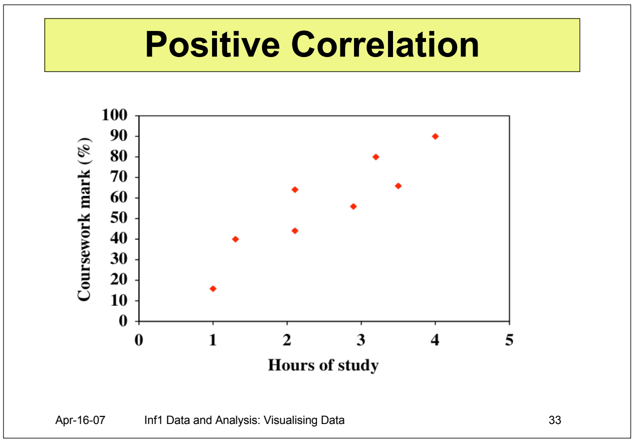
Consider scatter plots...

Where would you draw a line?

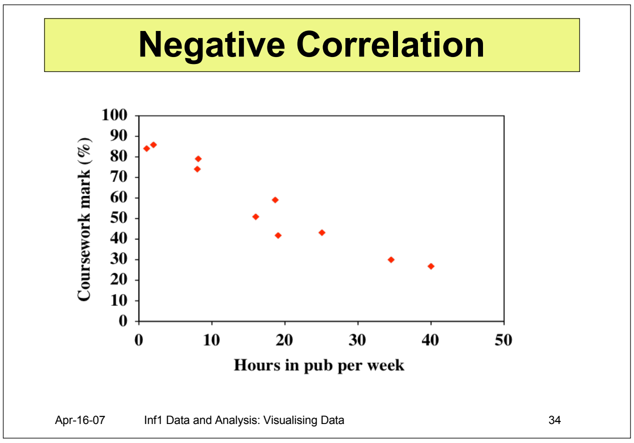


Hours spent v coursework mark

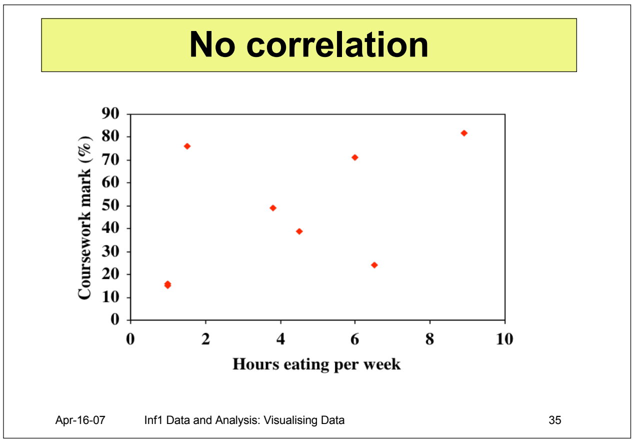
	STUDENT							
	a	b	c	d	e	f	g	h
Hours spent	1	1.3	2.1	2.1	3.2	2.8	3.5	4
% on cwork	16	40	44	64	80	56	66	90



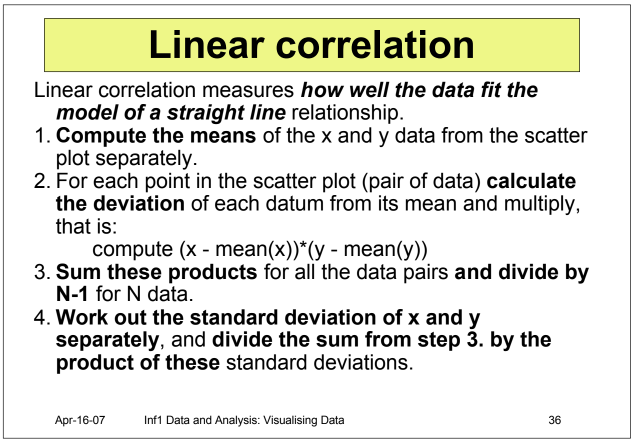
Positive Correlation



Negative Correlation



No correlation



Linear correlation

Linear correlation measures ***how well the data fit the model of a straight line*** relationship.

1. **Compute the means** of the x and y data from the scatter plot separately.
2. For each point in the scatter plot (pair of data) **calculate the deviation** of each datum from its mean and multiply, that is:

   compute (x - mean(x))*(y - mean(y))
3. **Sum these products** for all the data pairs **and divide by N-1** for N data.
4. **Work out the standard deviation of x and y separately**, and **divide the sum from step 3. by the product of these** standard deviations.

Pearson's Correlation Coefficient

Measures how well the data fit the straight line model it assumes:

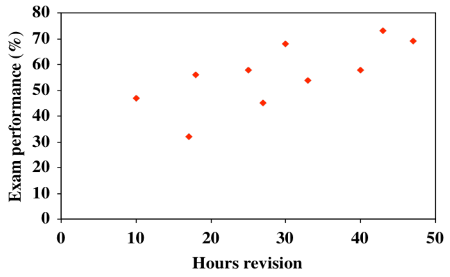
$$\text{correlation} = \frac{\sum \{(x - \mu x)(y - \mu y)\}}{(N-1)\ \sigma x\ \sigma y}$$

Lies between -1 (low X means high Y) and +1 (high X means high Y) with 0 meaning no correlation

Revision v exam performance
(example from Hinton, 1995)

	STUDENT									
	a	b	c	d	e	f	g	h	i	j
Hours studied	40	43	18	10	25	33	27	17	30	47
% on exam	58	73	56	47	58	54	45	32	68	69

Revision v Exam performance



Using Pearson's Correlation Coefficient

$$\text{correlation} = \frac{\sum \{(x - \mu x)(y - \mu y)\}}{(N-1)\ \sigma x\ \sigma y} = 0.72$$

To see if this might be due to chance, we need to know the **degrees of freedom** = n-2 = 8

One-tailed test - is correlation +ve or -ve?

Two-tailed test - is there a significant correlation?

Here, +ve correlation predicted, so one-tailed

From tables of probability for one tailed = 0.05, for 8 d.f. r = 0.5494

0.72 is greater than that, **so significant correlation with less than 5% probability it is due to chance**

Comments on Correlation…

A **high positive correlation** between two variables **doesn't mean that one causes the other**……

Say we get a correlation of 0.8 between exam performance and hours of study:

- Does this mean that the longer you study the better your exam results will be?
- or the better the exam results the more you will study?
- or some other variable influencing both (you are conscientious and bright)

Or ***time spent watching television and incidence of lung cancer are correlated***, but neither causes the other:

- both are caused by economic factors providing people with leisure time and money to buy cigarettes…

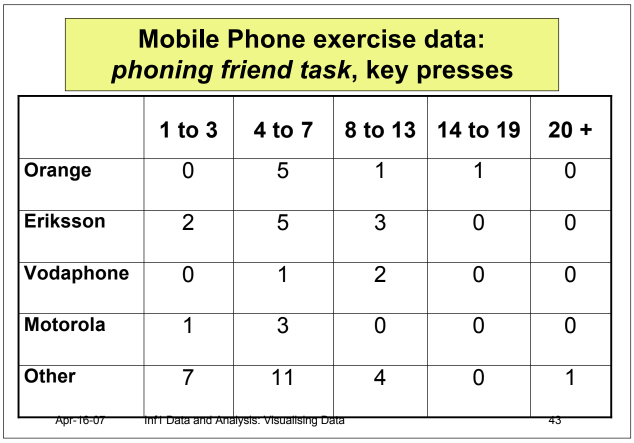
Statistical dependence is not the same thing as causal dependence

Reading

Dix, A., Finlay, J., Abowd, R. and Beale, R. (2004) *Human-Computer Interaction.* Prentice Hall
Chapter 9: Evaluation Techniques pp 318 - 364

(copies are available from ITO)

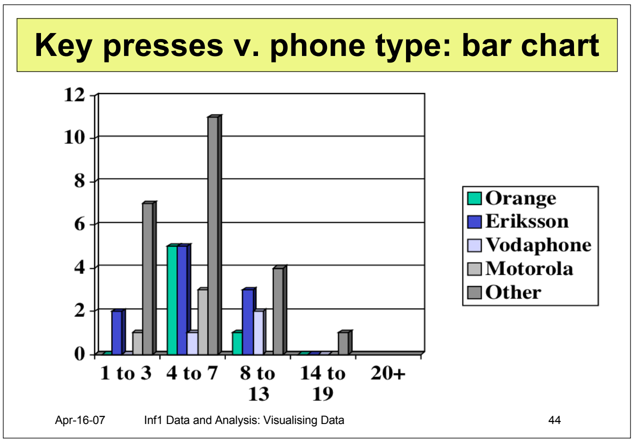
Hinton, P. (1995) *Statistics Explained,* Routledge, London, UK



Mobile Phone exercise data: *phoning friend task*, key presses

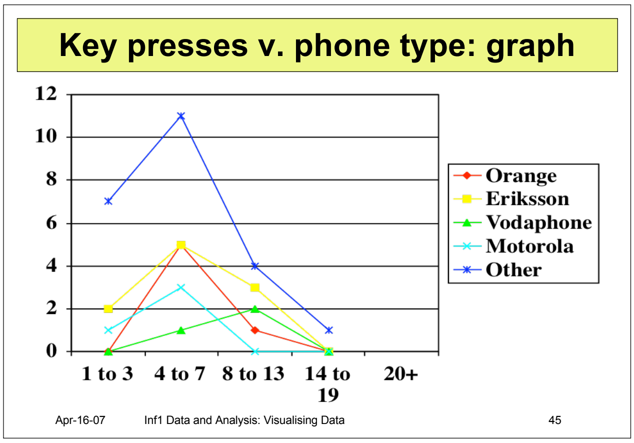
	1 to 3	4 to 7	8 to 13	14 to 19	20 +
Orange	0	5	1	1	0
Eriksson	2	5	3	0	0
Vodaphone	0	1	2	0	0
Motorola	1	3	0	0	0
Other	7	11	4	0	1

Apr-16-07 Inf1 Data and Analysis: Visualising Data 43



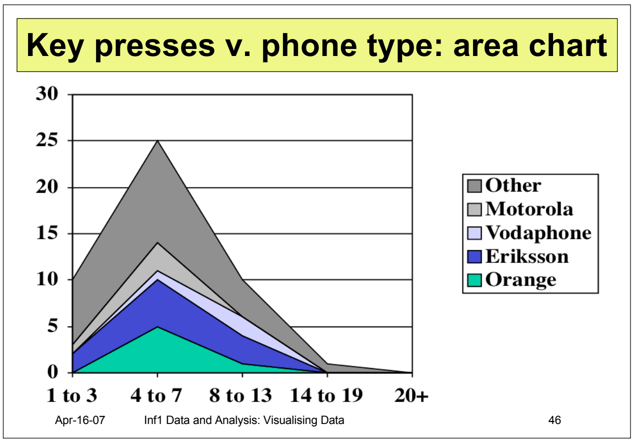
Key presses v. phone type: bar chart

Apr-16-07 Inf1 Data and Analysis: Visualising Data 44



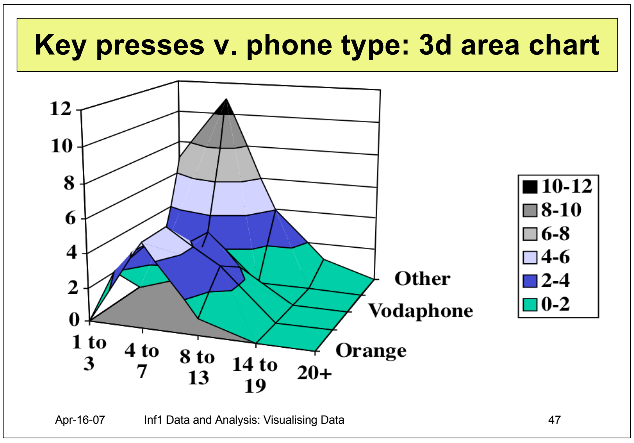
Key presses v. phone type: graph

Apr-16-07 Inf1 Data and Analysis: Visualising Data 45



Key presses v. phone type: area chart

Apr-16-07 Inf1 Data and Analysis: Visualising Data 46



Key presses v. phone type: 3d area chart

Apr-16-07 Inf1 Data and Analysis: Visualising Data 47