Insights from Data

Paul Hewson

May 31, 2012

National HE STEM Programme

Pre-course activity 1

- You were asked to watch a short video clip, and answer some questions (online).
- There were few "correct" answers, the main point of the exercise was to get you thinking
- Please now imagine you are a public servant, with finite money to spend on
  - An intervention to reduce "Single vehicle loss of control" injuries, or
  - an intervention to reduce "Sorry Mate I didn't see you collisions".
- In groups, make a decision. Also, be prepared to explain *why* you chose the intervention you did.

Learning outcome 1

GAISE 2010: 1

Data beat Anecdotes

Do we really use data?

The plural of anecdote IS NOT data

- Saying "data beat anecdotes" is a cornerstone of statistical literacy

Do we really use data?

The plural of anecdote IS NOT data

- Saying "data beat anecdotes" is a cornerstone of statistical literacy
- But we contradict it regularly.

Do we really use data?

The plural of anecdote IS NOT data

- Saying "data beat anecdotes" is a cornerstone of statistical literacy
- But we contradict it regularly.
- How often do you hear about my dear old Aunt Sally who smokes 80 cigarettes a day, drinks 8 bottles of gin and has still lived to be one hundred and fourteen years old
- Who points out that this is "the exception that proves the rule"

Joel Best: Lies, Damned Lies and Statistics

- Who created this statistic?
- Why was this statistic created?
- How was this statistic created?

So what?

- Adopt outright cynicism: don't believe anything based on data ever
- Adopt naïve acceptance (especially if the "facts" suit us).

Both positions have the advantage of being thought free. However, for a professional, a critically cautious position between these two extremes is needed.

Example: consider two courses

- In two years, the dropout rate for Course A doubles
- In the same two years, the dropout rate for Course B increases by 50%
- You only have resources to intervene with one course. Which course do you decide to "sort out"?

Possible numbers

Conveniently (and implausibly) there are 100 students on each course:

Course	2008	2009
A	4	8
B	40	60

Possible numbers

Conveniently (and implausibly) there are 100 students on each course:

Course	2008	2009
A	4	8
B	40	60

Changes in dropout rate:

- A: The dropout rate has increases from 4% to 8%
- B: The dropout rate has increases from 40% to 60%

Possible numbers

Conveniently (and implausibly) there are 100 students on each course:

Course	2008	2009
A	4	8
B	40	60

Changes in dropout rate:

- A: The dropout rate has increases from 4% to 8%
- B: The dropout rate has increases from 40% to 60%

- A: The "relative risk" is $\frac{0.08}{0.04} = 2$.
- B: The "relative risk" is $\frac{0.6}{0.4} = 1.5$
- A: The "absolute difference in risk" is $0.08 - 0.04 = 0.04$
- B The "absolute difference in risk" is $0.6 - 0.4 = 0.2$

Relative and absolute risk

Activity

Never mind the jargon!!! The way you compare proportions (or percentages) can alter your impressions.

Note to self - mention percentage points!

Shere Hite (1987), Women in Love: A cultural revolution in progress

- There were 4,500 respondents to this survey - does that sounds like a big number?
- What do you feel is a good number for a survey (note, YouGov predict the election within a few percentage points on smaller numbers than this)

Women in Love?

- 84% of women are not emotionally satisfied with their relationships
- 70% of women who have been married for more than 5 years have had affairs
- 95% have been emotionally or psychologically harassed by the male
- 84% have suffered condescension from the male

What is this survey telling us about the wellbeing of women in the US in the 1980s?

National HE STEM Programme

Women who fill in surveys in Love

- 100,000 surveys were sent out, only 4,500 came back
- There were 127 essay type questions
- Are the 4,500 women who filled in such a survey typical of the 100,000 who were invited to take part? Are the 100,000 who were invited to take part typical of women in the US in the 1980s

Learning outcome 2

GAISE 2010: 2

Random sampling allows results of surveys to be extended to the population from which the sample was taken.

Surveys or data?

- There are large national surveys (carried out by agencies or national statistical bodies) which use fairly sophisticated variations on "random sampling", but done in a way they can produce results that are "representative" (think about election forecasting).
- What about the surveys we commission locally / use locally?
- What problems might there be with non-response?
- Do we think these are representative?
- What could we do to make them more representative?

National HE STEM Programme

What's the population

- A big part of understanding (and designing) surveys is thinking about the relevant population

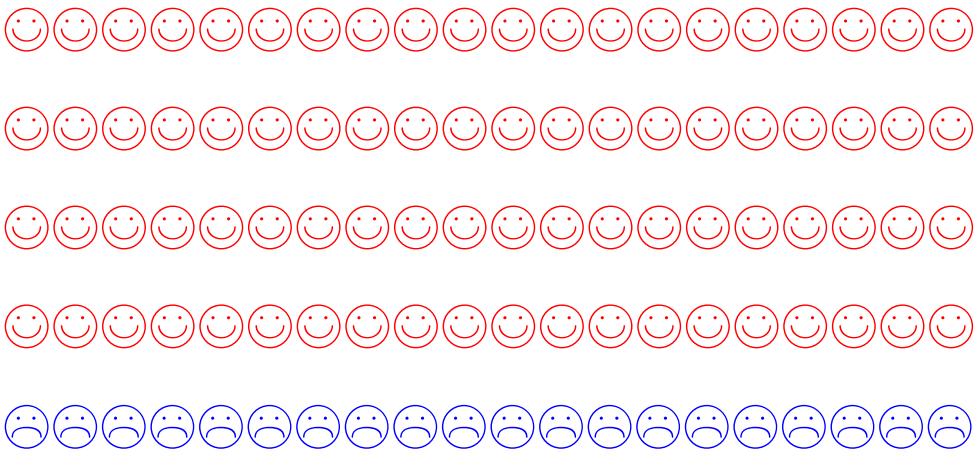
Mobile Phone Killer Crash Risk

Source: www.brake.org.uk/handheld-mobile-phone-use-one-rise-brake-reaction

- Survey of 21 year old students
- Is that typical

Happiness Census

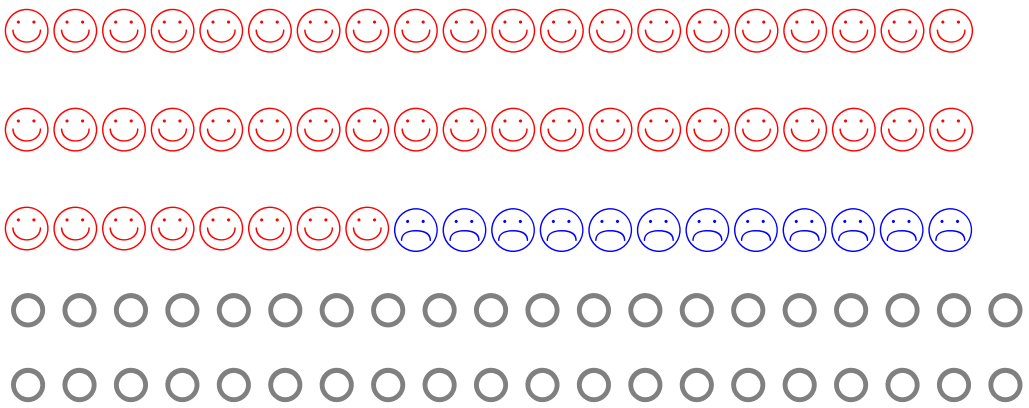
100 people questioned, all respond.



- 80 / 100 are happy, the proportion of happy people is 80%
- If this were a random sample of (say) 1,000 students we could do something "statistical" (compute a confidence interval for the proportion of happy students in the population). But we don't need to. Here, 100 is the population

Happiness Census

More realistically, 100 people questioned, 60 respond.

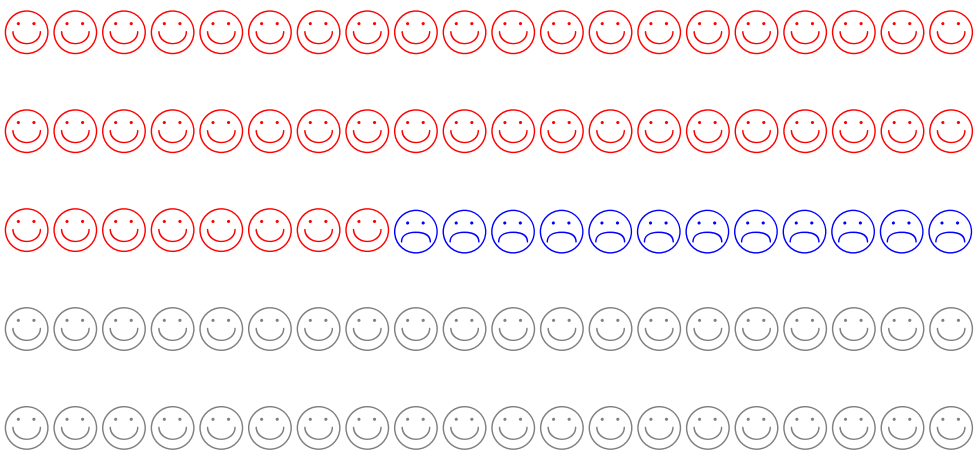


- 48 / 60 are happy; the proportion of happy people *who responded to the "survey"* is 80%
- We *can*[1] plug these numbers into a computer and do something statistical to get a 95% confidence interval for the *population* proportion as (68%, 88%). But just see how silly that is on the next two slides.

[1] although we shouldn't

If all the non-responders were too busy being happy to sit behind a computer filling in surveys

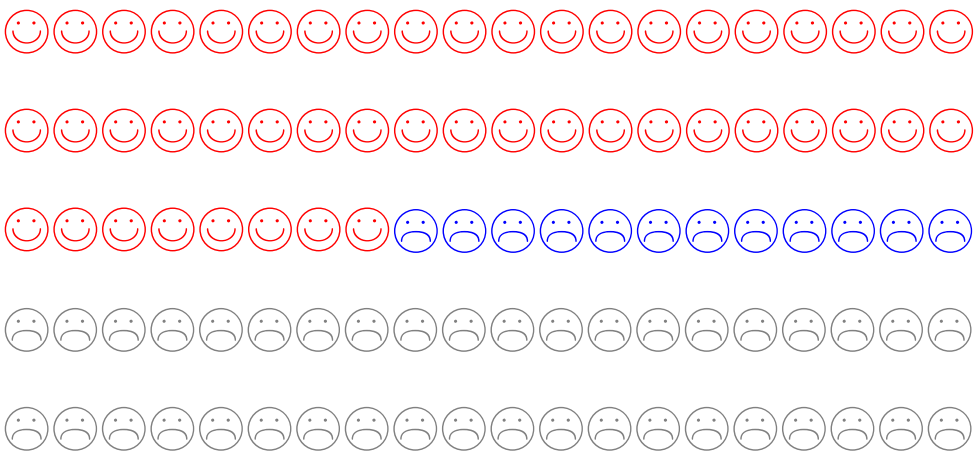
100 people questioned, 60 respond.



- In the population, we have 88 / 100 who are are happy; the proportion of happy people *in the population* is 88%

Or if all the non-responders were too fed up to fill in surveys (no computers, we didn't teach them how to use computers etc.)

100 people questioned, 60 respond.



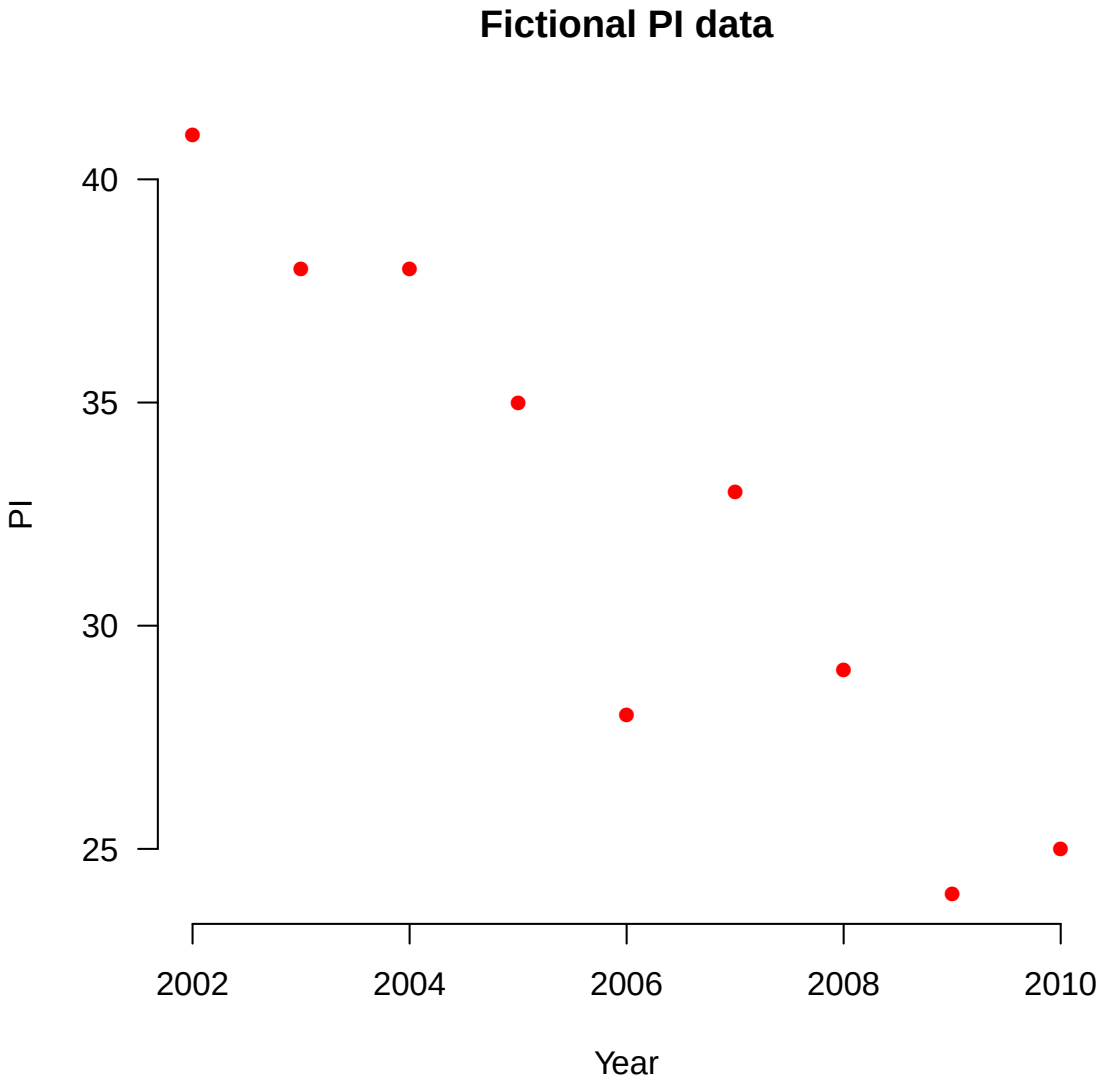
- In the population, we have 48 / 100 who are are happy; the proportion of happy people *in the population* is 48%

Non-response to surveys

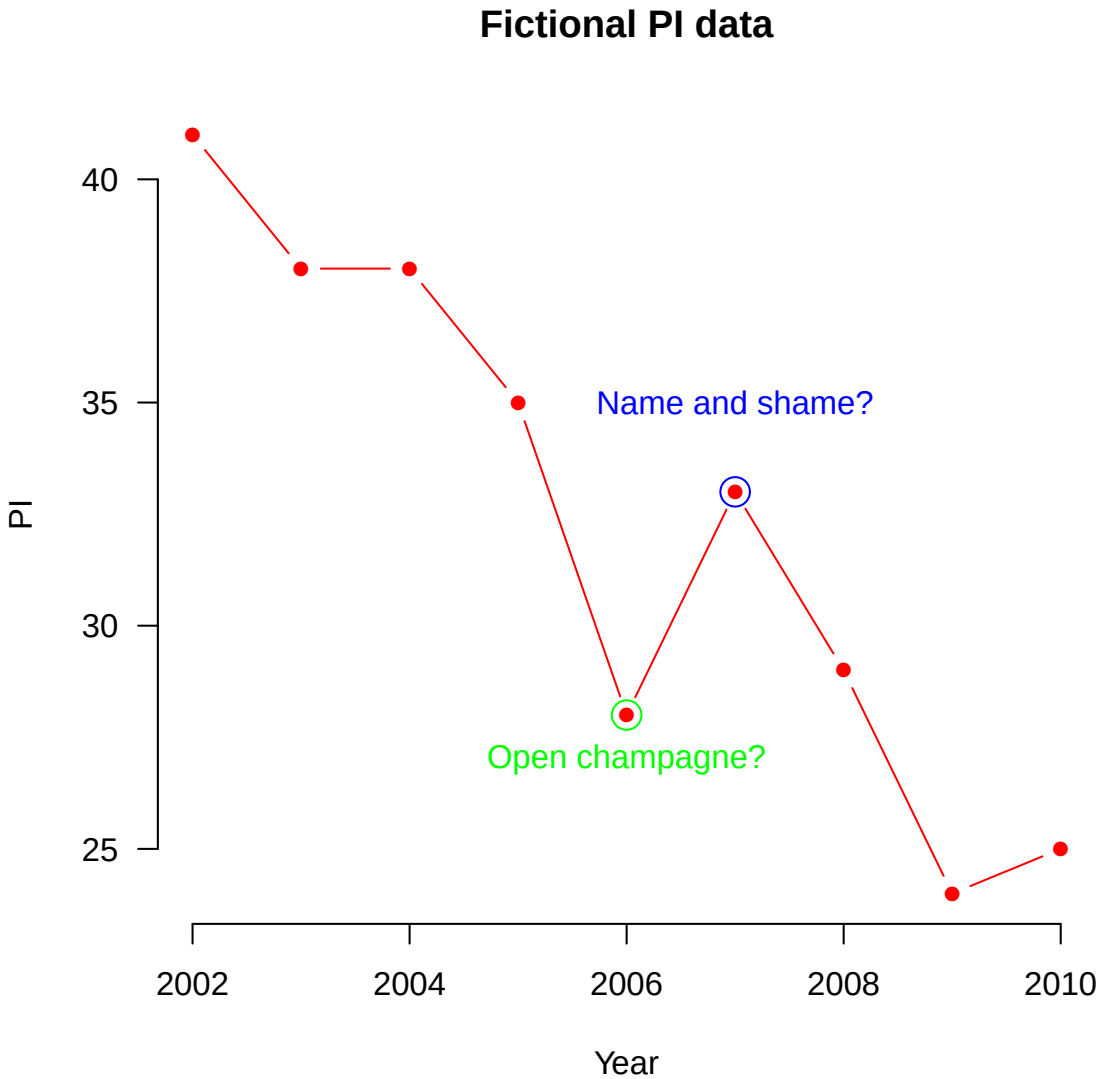
- Non-response is a real problem
- There are some (really quite advanced, i.e., rather specialised, i.e., think expensive expert) methods which try to help out. Clever as they are (i.e., I like playing with them) they are not a magic wand.
- Putting effort into minimising non-response (or at least understanding who is and isn't responding) is really quite important

National HE STEM Programme

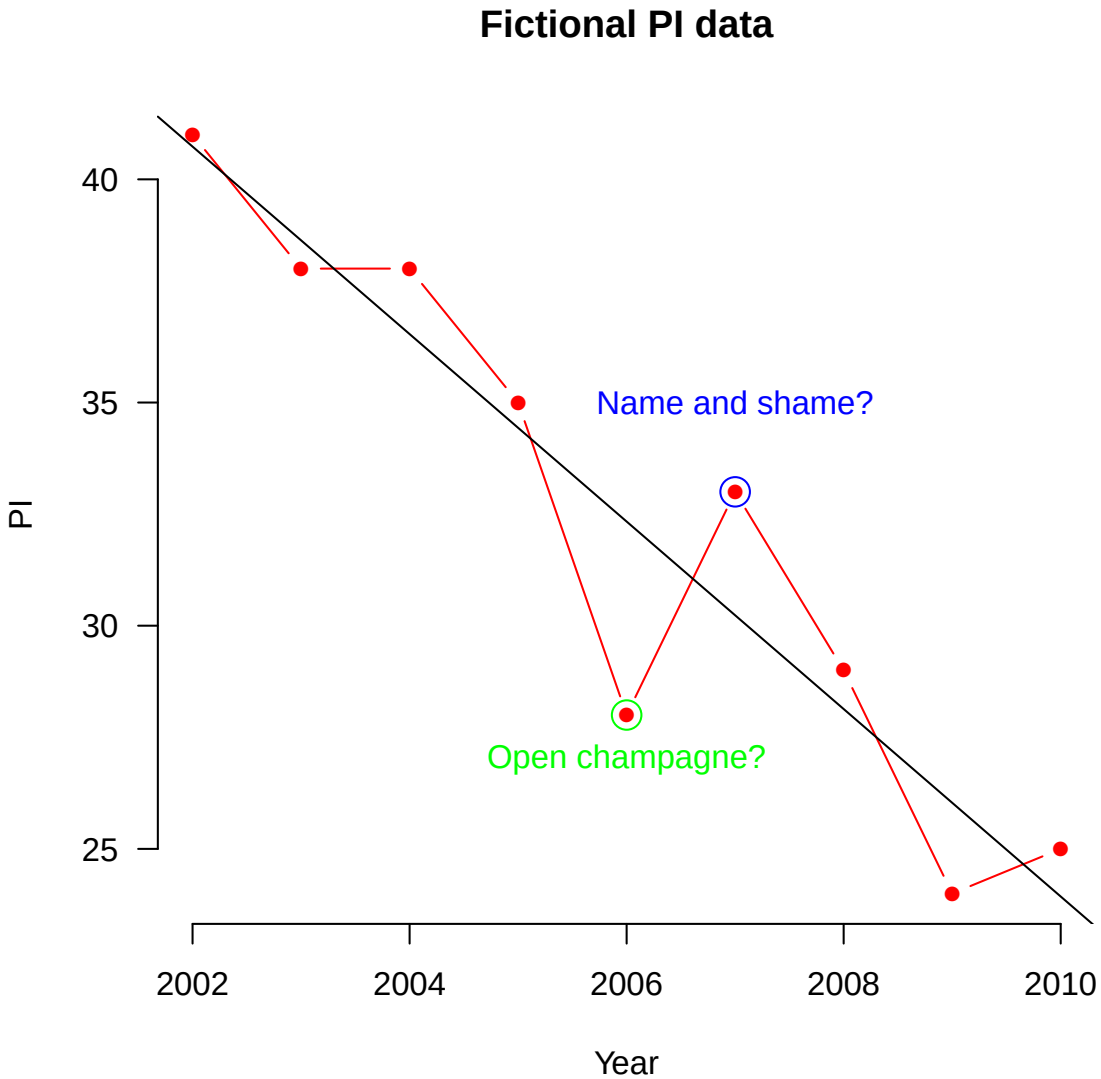
Fictional performance indicator



How we normally use these data



Isn't this just random blips either side of a trend?



Some real data

Year	Per capita	Hourly
1959	12,985	6.69
1964	14,707	7.33
1969	17,477	7.98
1974	18,989	8.24
1979	21,635	8.17
1984	23,171	7.80
1989	26,552	7.64
1994	28,156	7.40
1999	32,429	7.86

Table: (From US Department of Commerce "Economic report of the President 2000" cited by Joel Best, all prices USD)

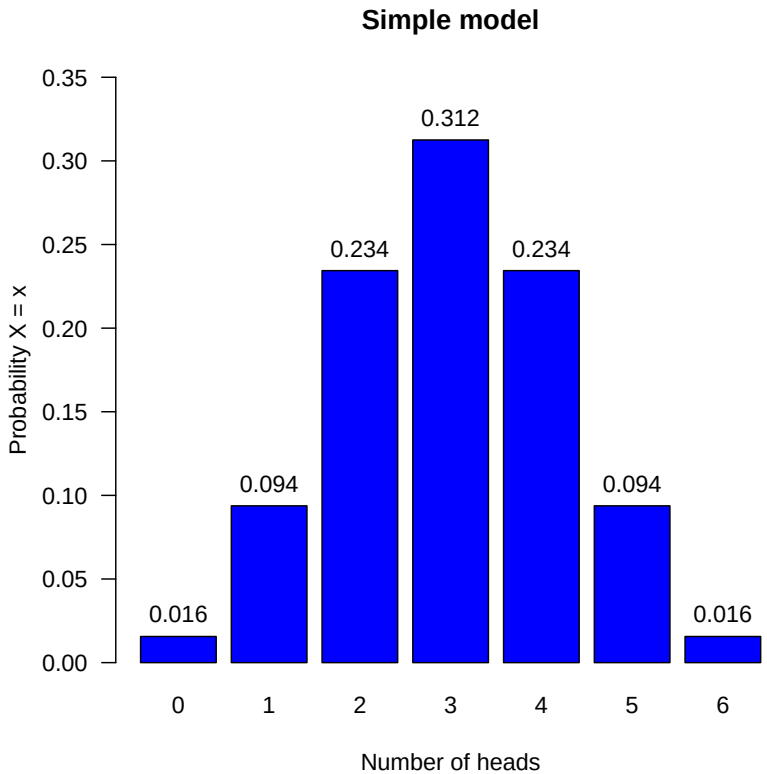
- How do we present these data?
- What is it telling us?

Modelling randomness

- Now we need everyone to toss a coin six times
- Count the number of heads
- (we may need to do this a second time if the numbers are small)
- Let's graph the results

One theoretical model

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

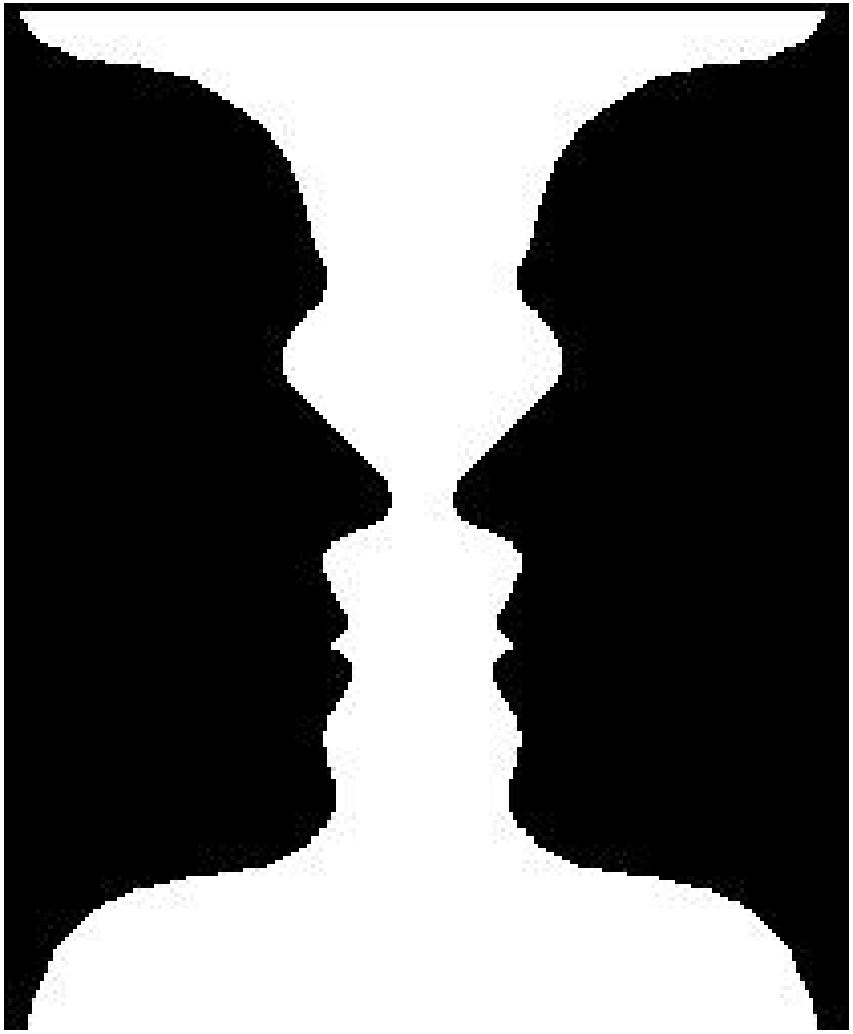


Learning outcome 3

GAISE 2012: 2

Variability is natural and is also predictable and quantifiable

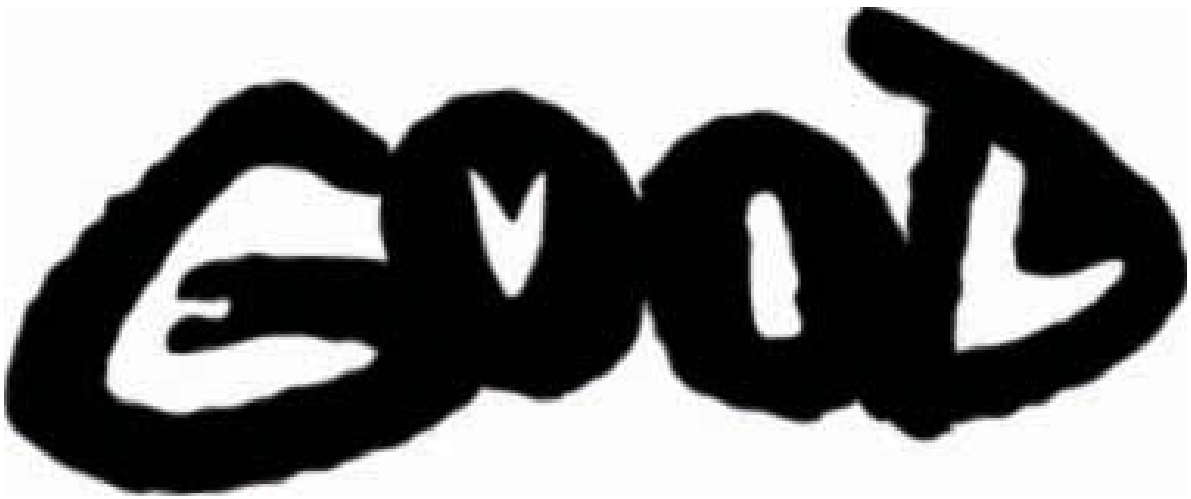
What's the first thing you see



What's the first thing you see

What's the character in the middle?

What's the first word you see?



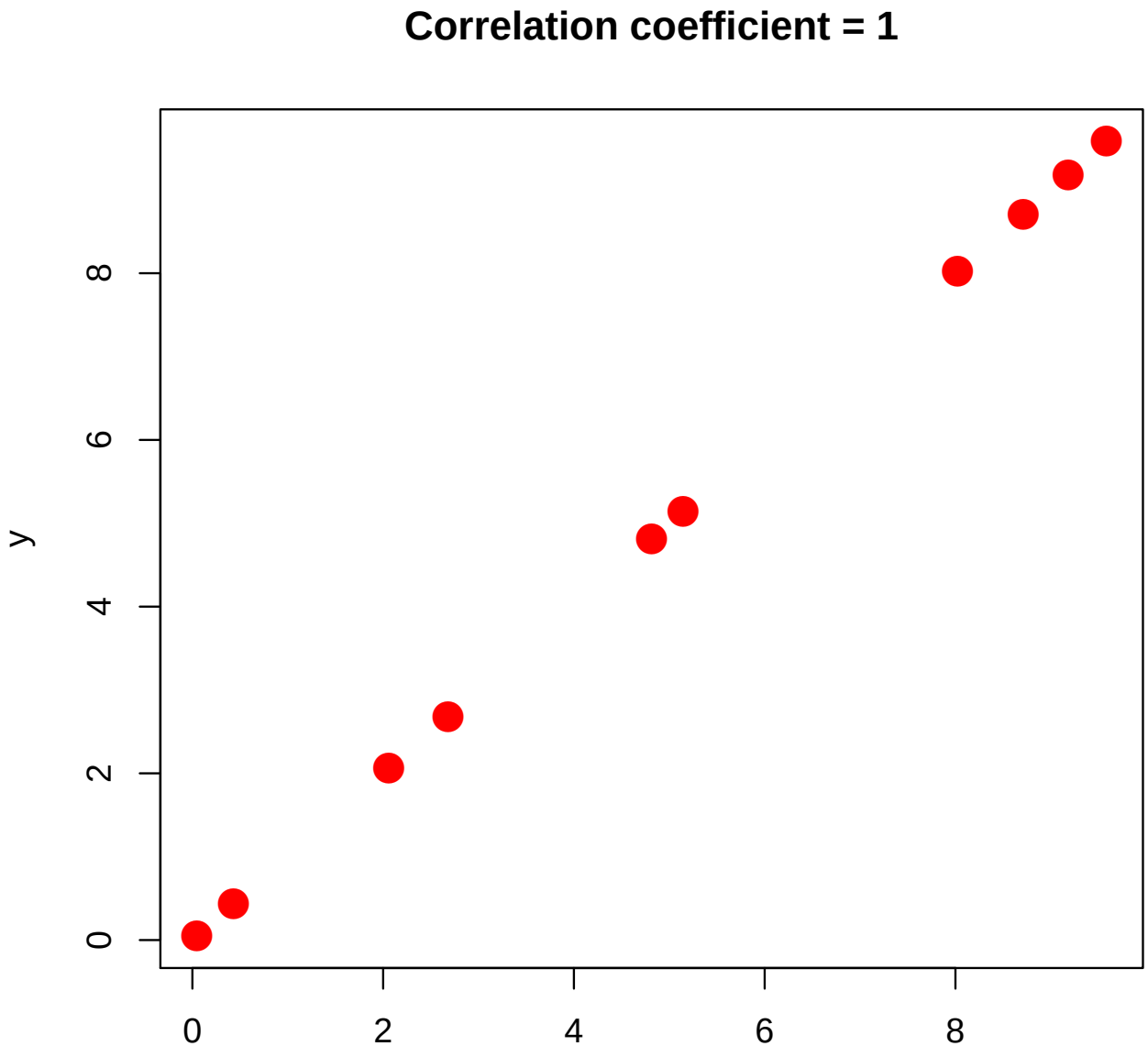
Opening question

- What does the word “correlation” mean to you?
- What do you think it means?

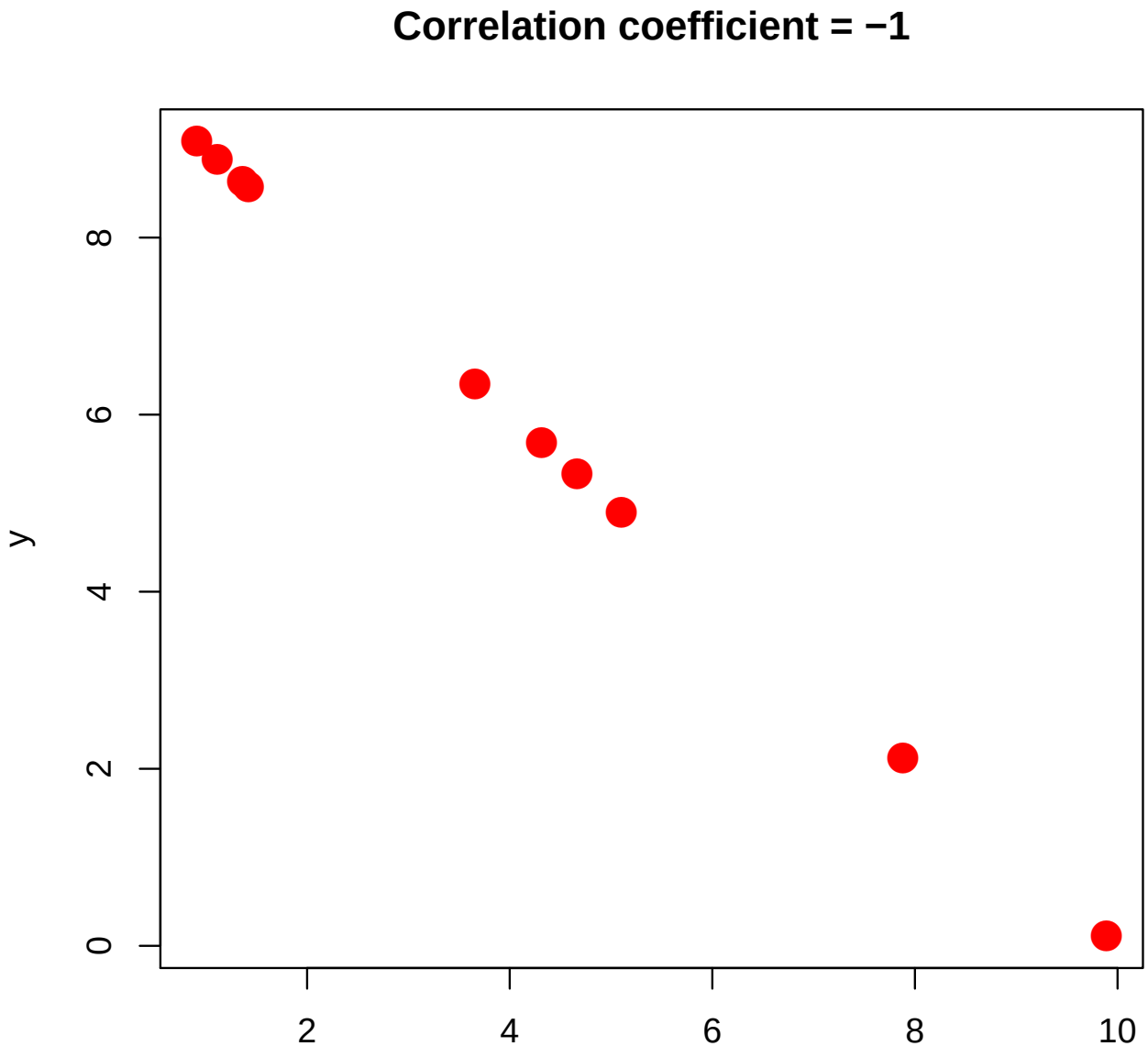
Correlation coefficient

- An attempt to produce a single number to describe the linear association between two variables
- Upper value is $+1$ (perfect positive linear association)
- Lower value is -1 (perfect negative linear association)
- Middle value is 0 (no linear association)

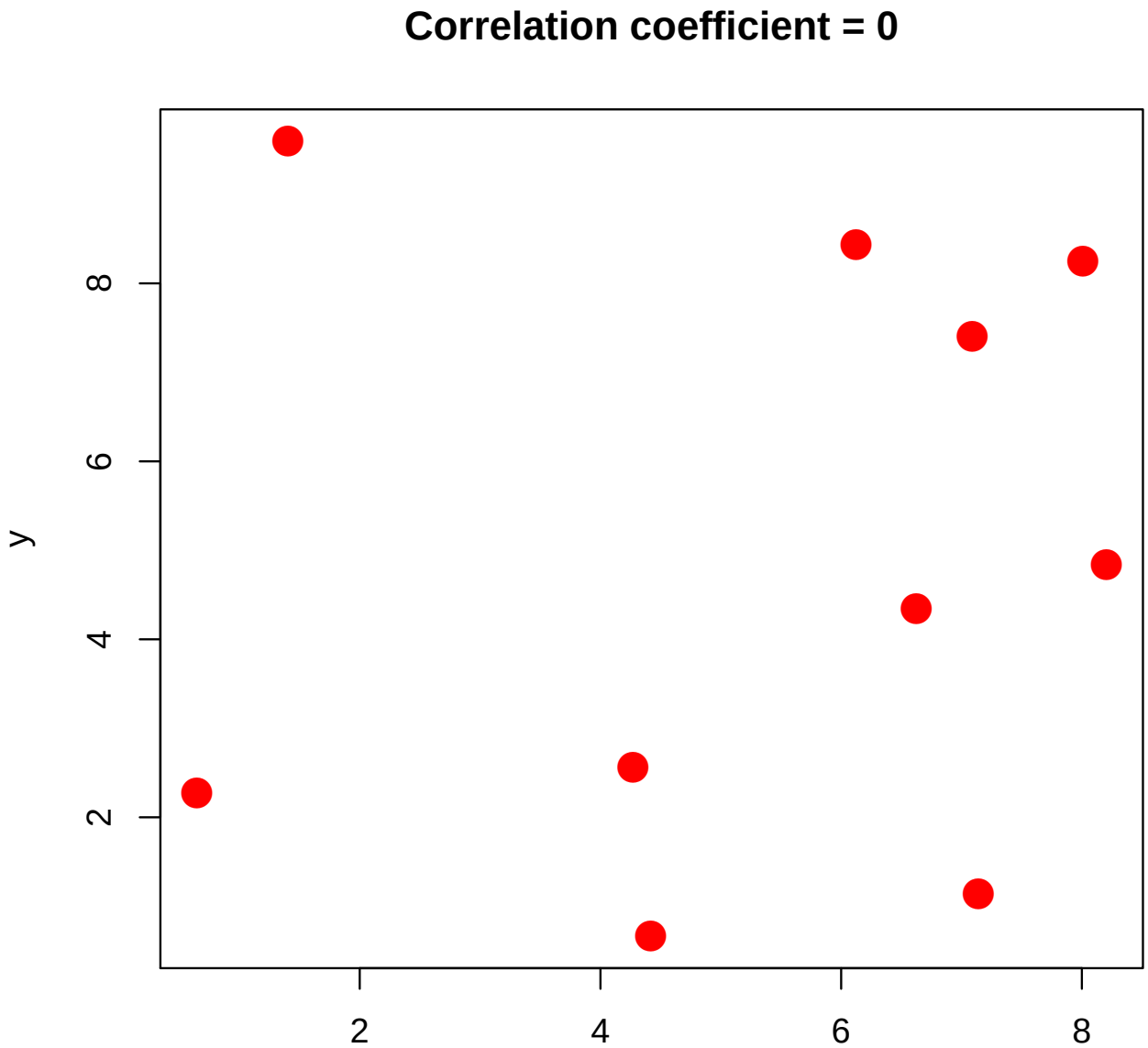
Correlation coefficient of 1



Correlation coefficient of -1

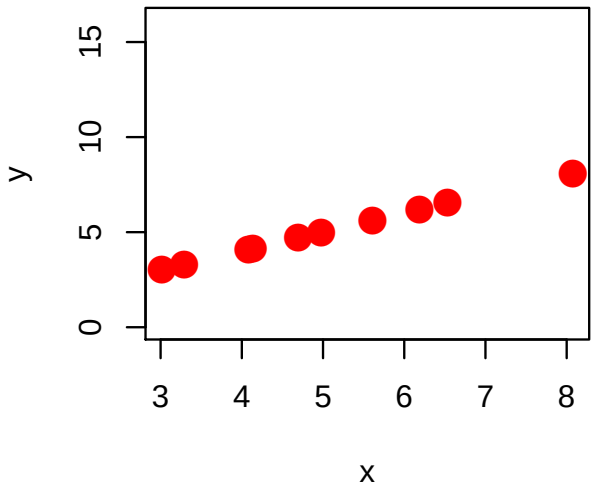


Correlation coefficient of 0

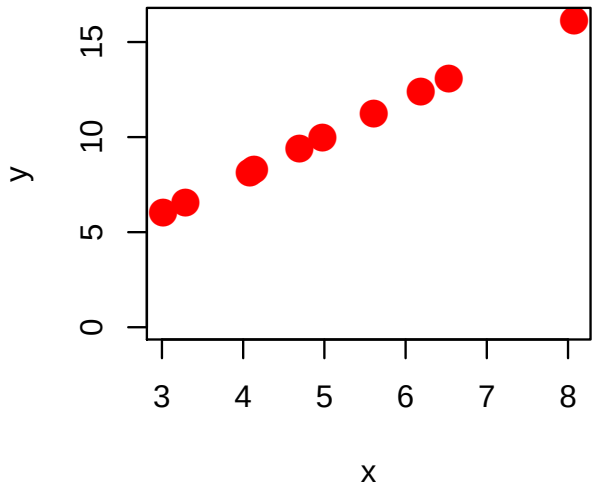


Higher correlation is (a) Left, (b) Right (c) Neither

Correlation coefficient = ?

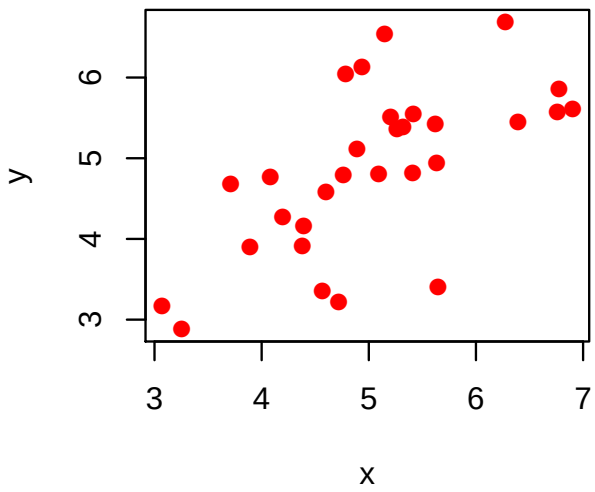


Correlation coefficient = ?

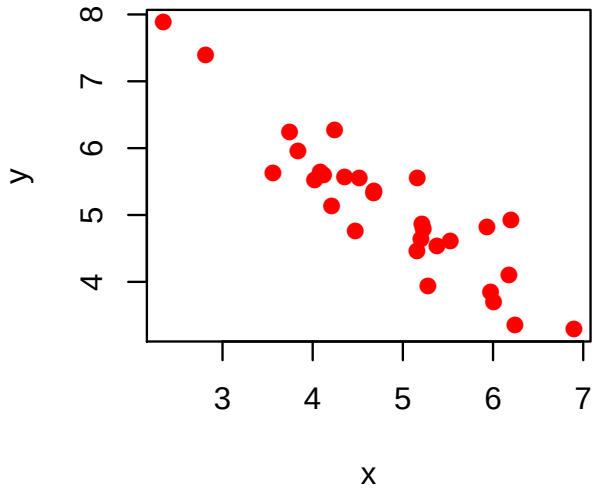


+0.5, -0.8, +0.9, -0.3

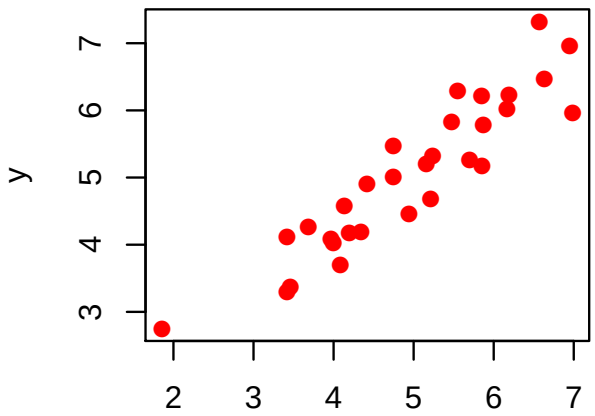
Correlation coefficient = ?



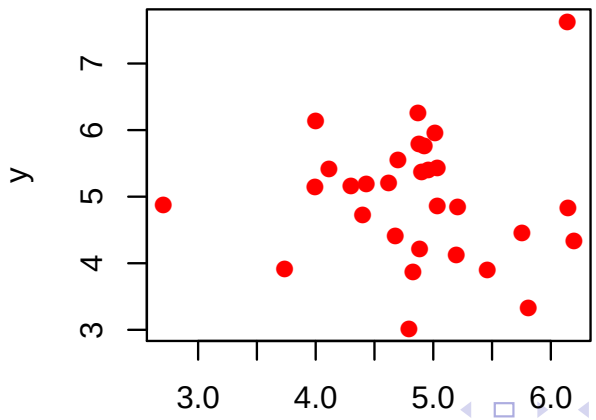
Correlation coefficient = ?



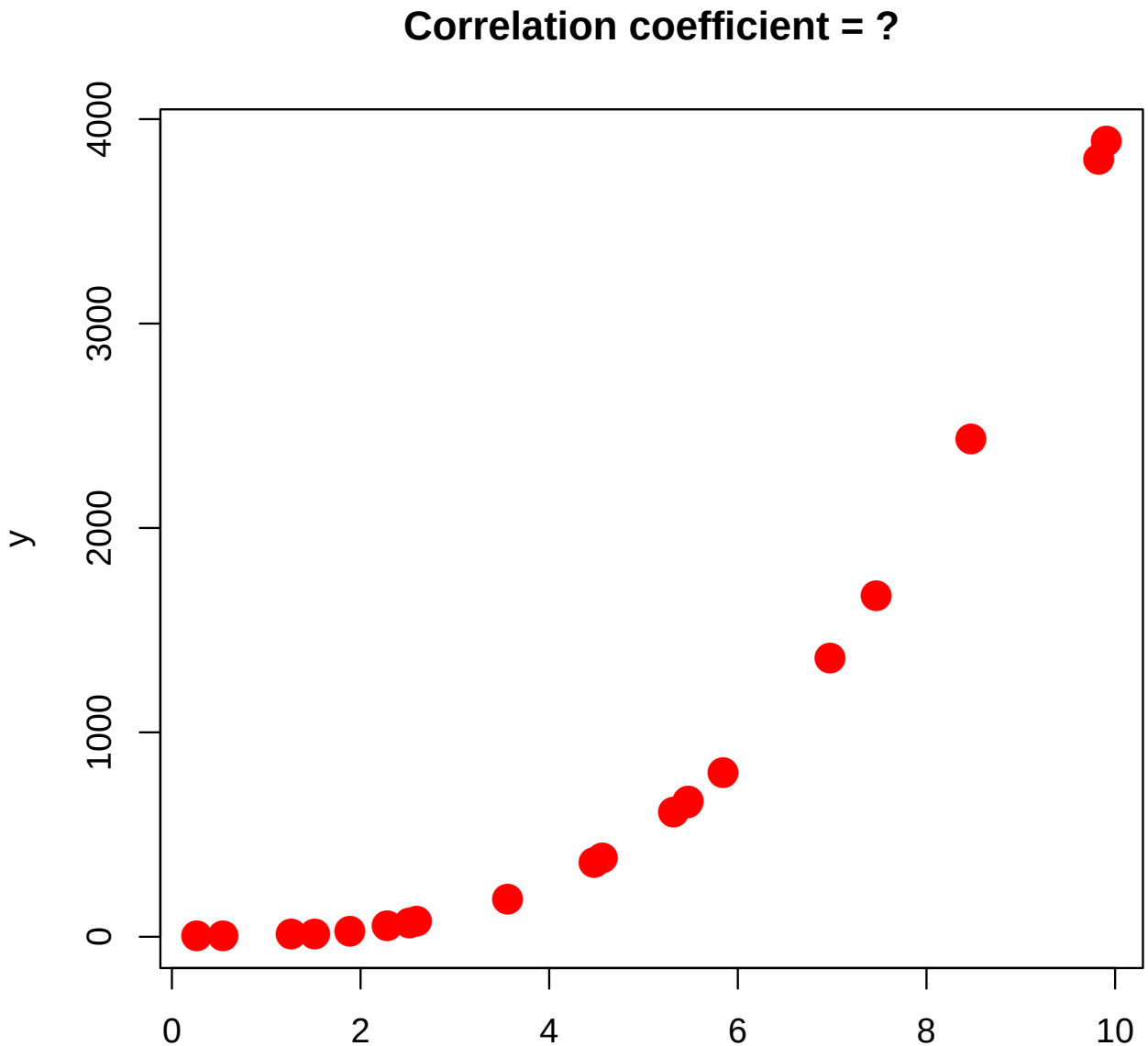
Correlation coefficient = ?



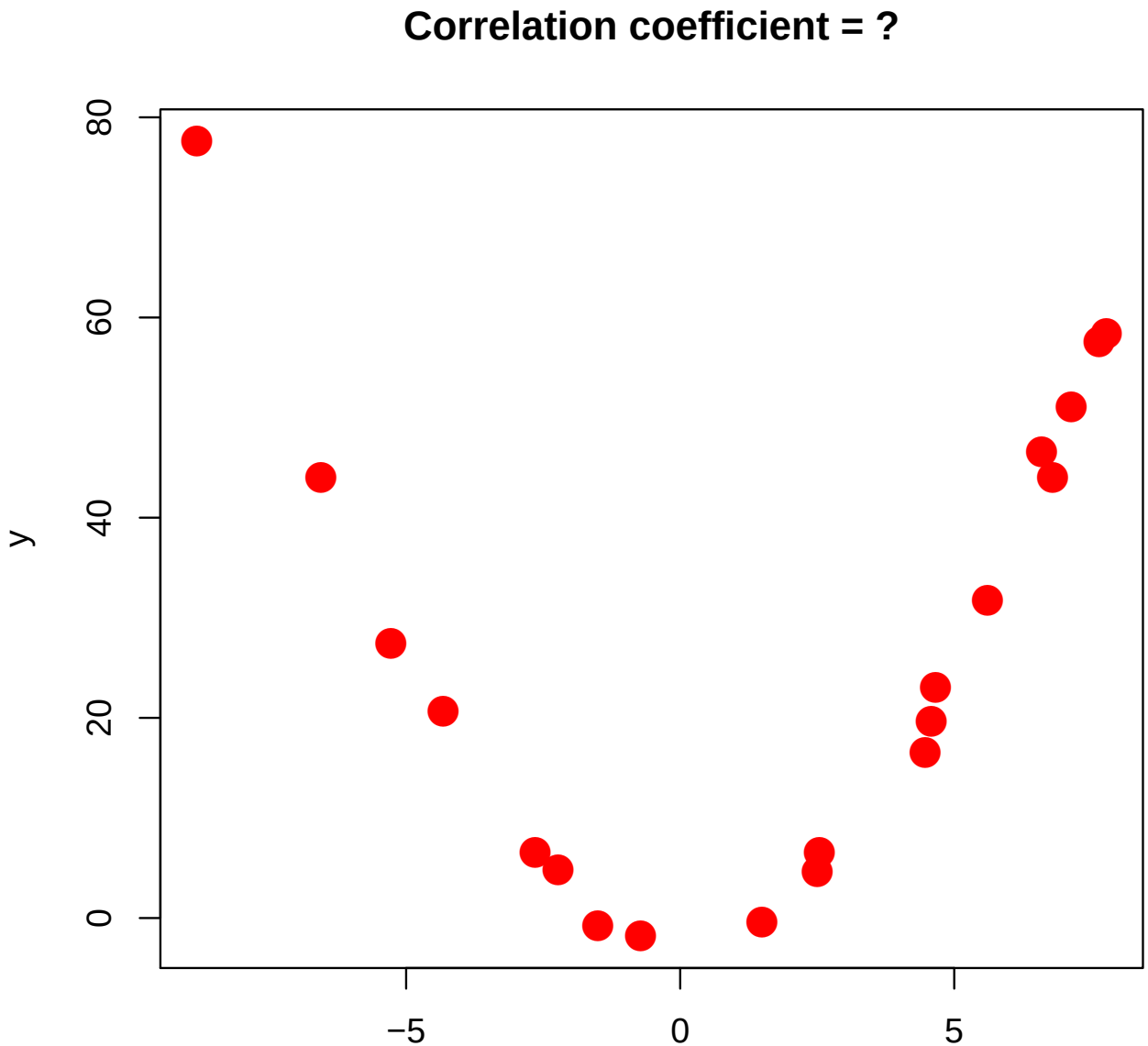
Correlation coefficient = ?



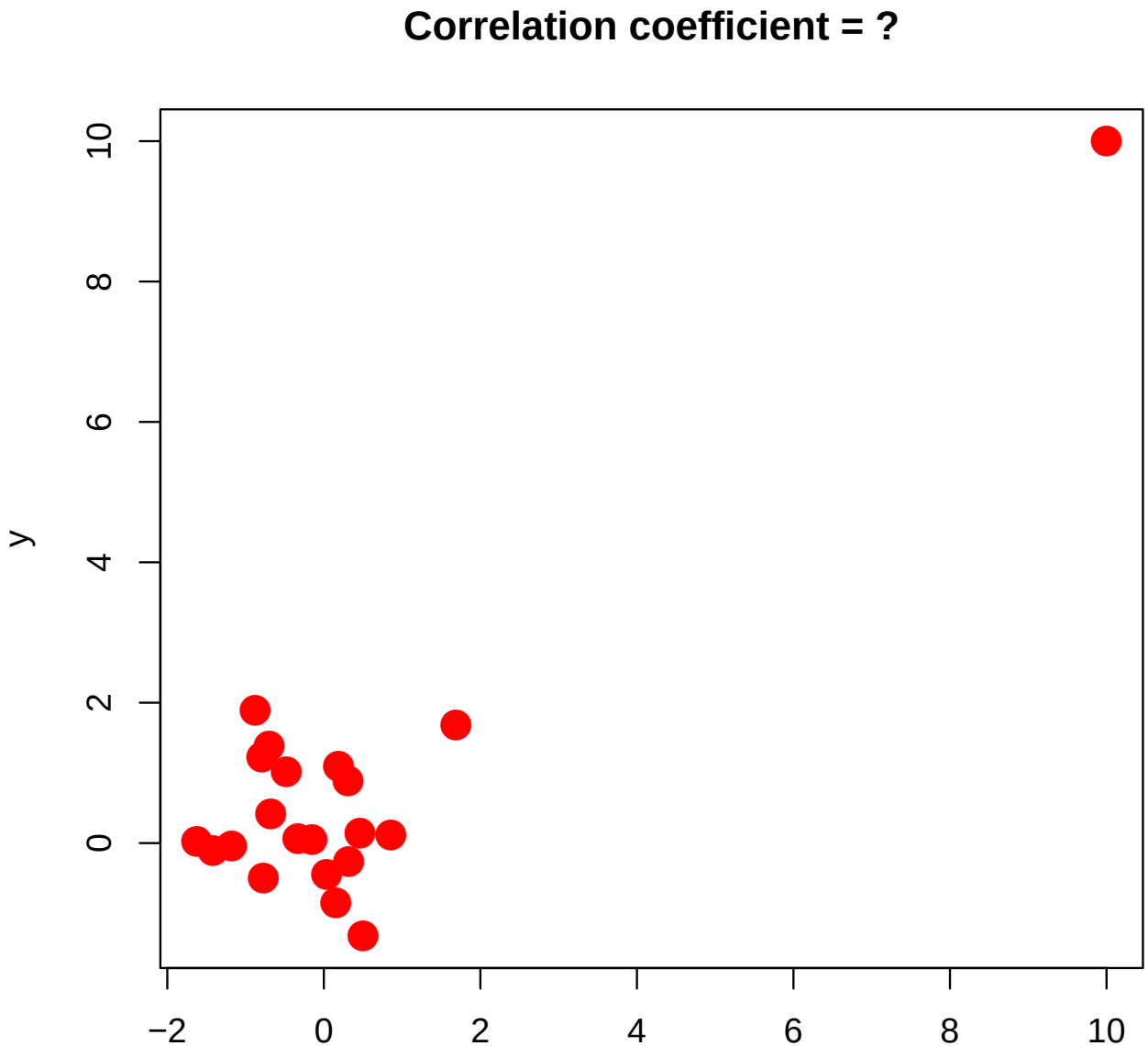
Non linear association



Non linear association



Also scope for outliers



Learning outcome 3a

GAISE 2010: 2

Correlation is a (rather simplistic) measure of *linear* association between two variables

We use a sample correlation coefficient as an estimate of the "population" correlation coefficient.

Important mantra for this week

Correlation does not imply causation

> *The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.*
> *Stephen Jay Gould (1941 - 2002)*

http://www.thepsychfiles.com/2009/11/episode-109-correlation-and-causation/

Contextual variables

What affects the success of your practice. Consider both:

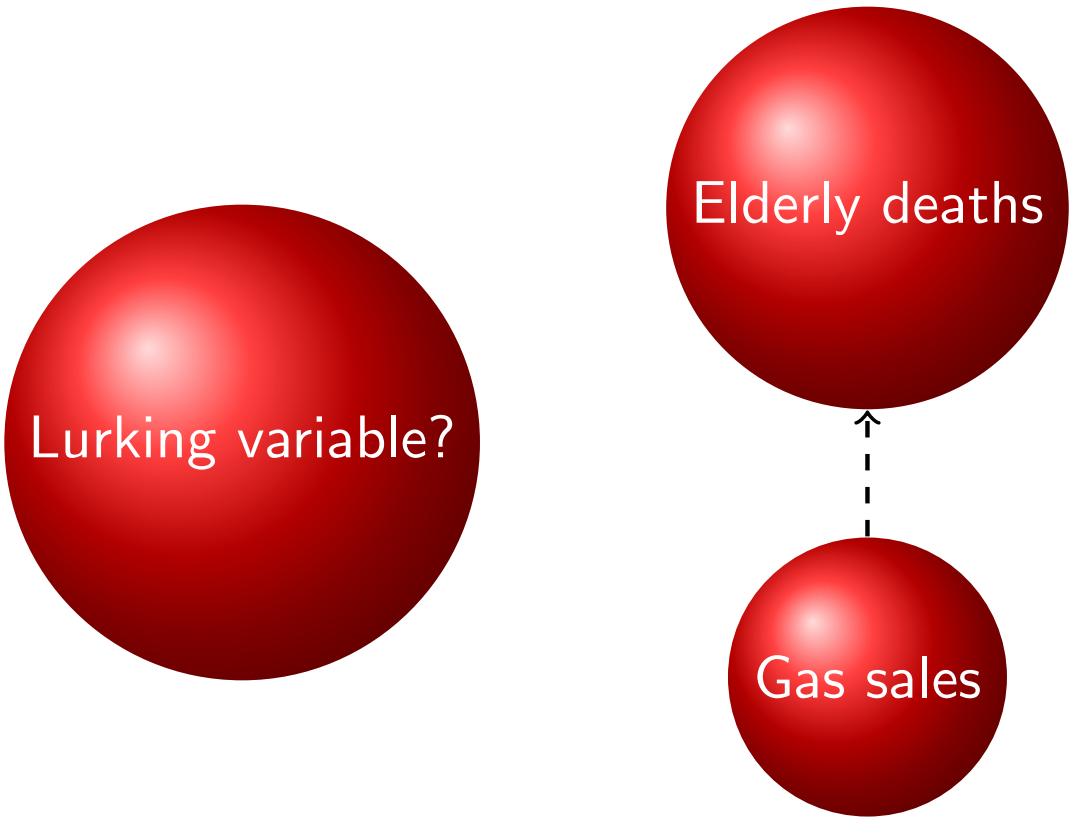
- Intrinsic factors
- Extrinsic factors

(This slide isn't intended to add to jargon overload - it's to highlight the fact you should think of things you control as well as things that are beyond your control)

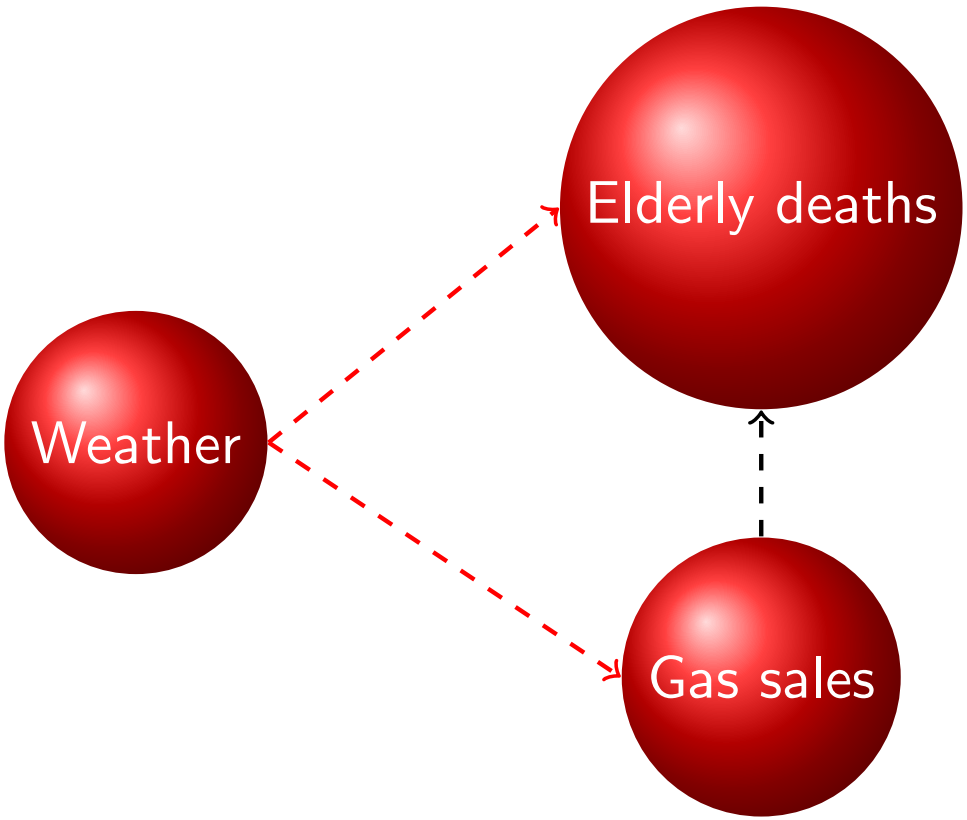
Illustrating Lurking variables

- There is a correlation between weekly gas sales and weekly deaths of old people (in the UK). Why?

Lurking variables



Weather as a lurking variable



In other words

- There is an apparent association between gas sales and elderly deaths
- But the reason for this association is probably due to weather. Weather is associated with both the elderly death *and* the gas sale variable

This is a key idea, and worth careful discussion. We shall review several of the news items that have been submitted and see where we need to think about lurking variables.

National HE STEM Programme

Exhibit 1: "Male road accidents soar in summer due to women's short skirts"

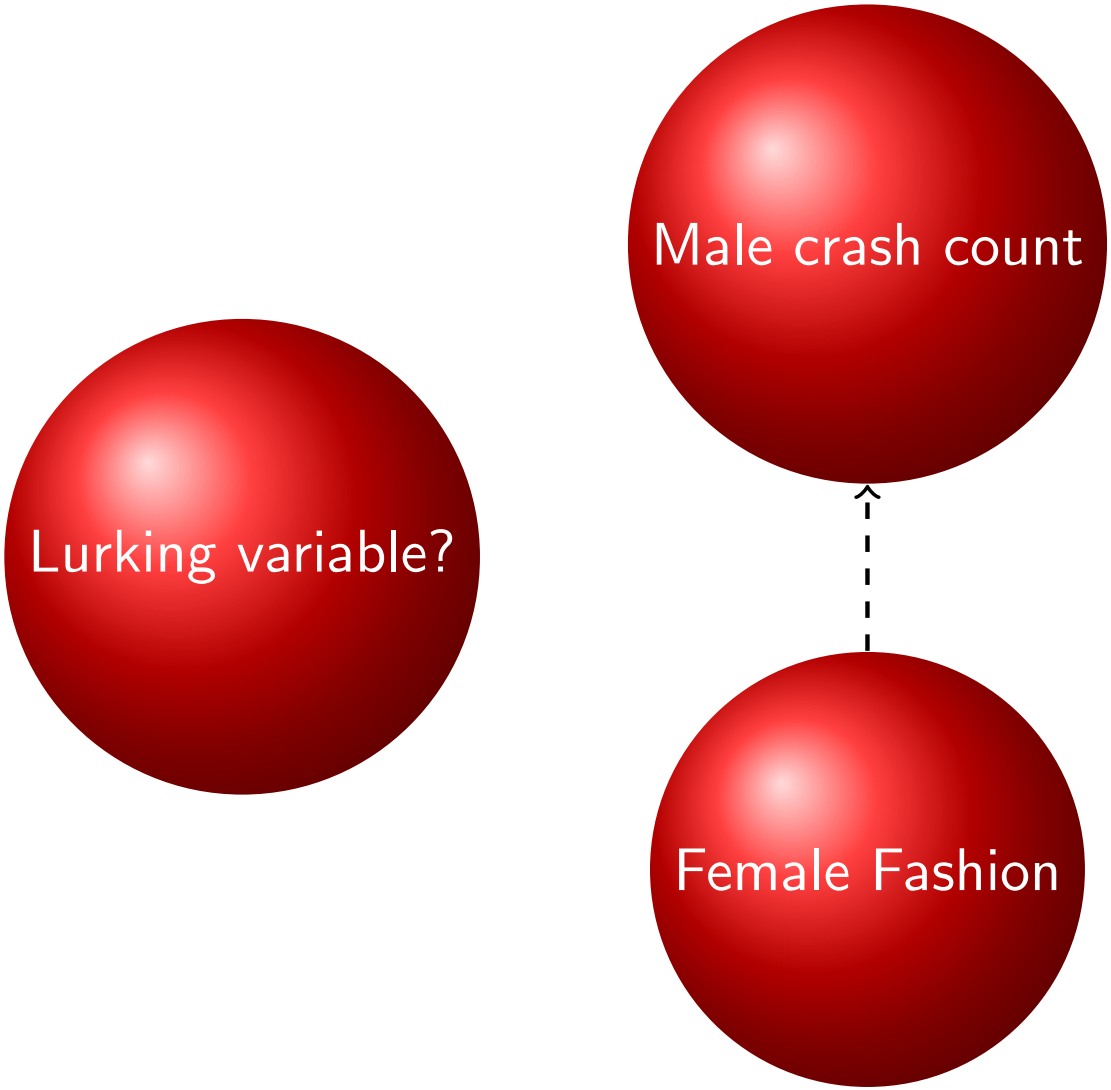
> *"The scorching heatwave in early July caused road accidents to soar because male drivers were distracted by womens' skimpy outfits, according to insurance claim figures."*

Source: Telegraph Online Posted 7:00am BST 30th Jul 2010
http://www.telegraph.co.uk/motoring/news/7917861/Male-road-accidents-soar-in-summer-due-to-womens-short-skirts.html

> *"And according to car insurance company Sheilas' Wheels, the summer smash phenomenon is getting worse each year - in 2009 men made 16.4 per cent more claims during the Summer than in any other month."*

Lurking variables

Suggest possible lurking variables:



A word on causation

1. Sometimes we collect data to prioritise, to explore a phenomenon
2. Sometimes we collect data in order to understand whether certain attributes (behaviours/people) are more likely to lead to worse outcomes
3. Sometimes we collect data as part of a study in order to decide whether an intervention "works"

Motive 1 is all about exploration. Motives 2 and 3 are about causation. However, for ethical reasons, we can usually only get really strong evidence of causation for Motive 3.

A simple memory test

An experiment to determine whether sugar in chocolate enhances performance in a memory test.

- Put all your pens down.
- You will be read a sequence of 8 numbers.
- You may then pick up your pens, and write down the sequence from memory.
- We will then mark how many you got right. Maximum mark is 8, but to get a mark the number must be in exactly the right position (so if you get four right, miss one altogether, and get the next three right you only get four marks as the last three will be out of position)
- When we've done this once, I will tell you to eat the chocolate.
- After 12 minutes, we will repeat the exercise

National HE STEM Programme

Flu remedy

Consider the following fictitious data:

Treatment	Pain relief	No pain relief	Total number	Percent with pain relief
Remedy	386	414	800	48%
Control	317	483	800	40%
Total	703	897	1,600	

- It is summarised as A 2×2 contingency table
- Your question: is the "Remedy" better than the control

Females

Now consider a subgroup analysis of half the study group.

Treatment	Pain relief	No pain relief	Total number	Percent with pain relief
Remedy	351	249	600	59%
Control	142	58	200	71%

- Your question: does the remedy "work" for this subgroup

Males

And just to be fair, let's consider the other subgroup

Treatment	Pain relief	No pain relief	Total number	Percent with pain relief
Remedy	35	165	200	18%
Control	175	425	600	30%

- Your question: does the remedy work for this subgroup?

What's going on

- I think the top level table implies that the remedy "works". 48% of people taking the remedy had pain relief, which is better than the 40% who took the control

What's going on

- I think the top level table implies that the remedy "works". 48% of people taking the remedy had pain relief, which is better than the 40% who took the control
- Clearly, plenty of people who took the control had pain relief.

What's going on

- I think the top level table implies that the remedy "works". 48% of people taking the remedy had pain relief, which is better than the 40% who took the control
- Clearly, plenty of people who took the control had pain relief.
- But in some sense, you seem to increase your chance of pain relief if you take the remedy

National HE STEM Programme

What's going on

- I think the top level table implies that the remedy "works". 48% of people taking the remedy had pain relief, which is better than the 40% who took the control
- Clearly, plenty of people who took the control had pain relief.
- But in some sense, you seem to increase your chance of pain relief if you take the remedy
- Provided there are no worrysome costs (cash or side effect wise) incurred by taking the remedy it seems better to take the remedy

National HE STEM Programme

The paradox

- More females (600) than males (200) took the remedy
- Whatever treatment they take, females are more likely (59%, 71%) to report pain relief than males (18%, 30%)
- Combine the two points, and we have a situation where the top level results completely contradict the subgroup results
- Although these are fictitious data, we can find plenty of real world examples

National HE STEM Programme

Random allocation

- Had we randomly allocated subjects to remedy or control, we would have (roughly) as many males and females in each group.
- We could "randomise away" the systematic differences between males and females.
- Now, if we have all kinds of other systematic differences (long sighted people / short sighted people, quick reaction time/slow reaction time), we couldn't identify all the possible subgroups.
- However, hopefully the random allocation process spreads these systematic differences out evenly
- So hopefully, the only systematic difference between the two groups is that they received a different experimental treatment
- And we therefore hope that any differences we observe between the two groups can be attributed to that treatment.

National HE STEM Programme

Learning Outcome

GAISE 2010: 4

Random assignment in comparative experiments allows cause and effect conclusions to be drawn.

- The fundamental contribution of statistics to experiments is in terms of *randomisation*.
- This is a most powerful way of dealing with potential confounding variables. Not only do we reduce the risk of confounding (by averaging it out across all experimental groups) but we can actually measure the remaining uncertainty.

But this was just made up data

- I usually think the above examples illustrates the need for randomised controlled experiments
- But what if we can't conduct an experiment. A very famous example are admissions to Graduate Schools at University of California, Berkeley (1973)

Gender	Applicants	Admitted
Male	8442	3546
Female	4321	1512

Is there any evidence of Gender Bias?

By school

	M	Admitted	F	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

What do you see?

National HE STEM Programme

Less dramatic problems

- An institution's NSS results depend on responses from a variety of courses, genders etc. Gender certainly affects satisfaction
- Drop out rates (again): imagine slightly more plausible class sizes (they vary)

	2008	2009	Total
A	0/30	5/70	5/100
B	1/70	3/30	4/100

A recreational problem

- Some plagiarism software is on offer to you. It will detect plagiarised work 80% of the time, and will report non-plagiarised work as such 90% of the time
- Your "plagiarism detector" flags up a coursework as plagiarised.
- What's the chance you have a piece of plagiarised work in your hands

Solution

Let's say 1 piece of coursework in 1000 is plagiarised

	Plagiarised	OK	Total
Detected	0.8	99.9	100.7
Not-detected	0.2	899.1	899.3
Total	1	9999	1000

So, $\frac{0.8}{100.7} = 0.0079$

Bayes Theorem

Activity

The take away point is meant to be that "the probability of A given B" is not the same as "the probability of B given A". The "wording" of a probability is as important (more important) than the actual number.