Insights from Data

Paul Hewson

April 23, 2012

Contents

1 Pre-amble

This course has been developed with funding from HE-STEM and in partnership with Devon County Council. The support of both is gratefully acknowledged.

The aim of the course is to provide a "one day" introduction to statistical literacy for use in the workplace. In doing that, we are trying to make a variety of delivery modes available, a blended course (a combination of online pre-course work, a group contact session and a little online post-course work) as well as a fully online course. At the time of writing we continue to evaluate the relative merits of each delivery mode.

What this does mean is that for any course, there are online materials, in the form of a Moodle course, which accompanies these notes. Currently, a live version of these materials are being hosted by the Royal Statistical Society Centre for Statistical Education, and these should allow guest access. A zip archive of the course is also available for anyone wishing to adopt the course.

All course materials are being offered under the GNU Public Licence. These learning materials are free to use by anyone, with the sole restriction that you cannot subsequently restrict anyone else's use of these materials.

The course is intended to consist of:

- Some pre-course activities
  - watch a video, answer some questions about the video
  - submit "newspaper" articles into a database
- A course - either a "contact" session or an entirely online course
- Some post-course activities

1.1 Activity 1

Following any housekeeping and ice-breaker activities, the course starts with a discussion of the video (and quiz) that was set for pre-reading.

> ? *Activity*
>
> Discuss how we priortise an action in this scenario

The fundamental goal of the video exercise was to get people thinking about the evidence we need in order to make a decision about an intervention. Road travel is a ubiquitous activity. Many people expect the state (central/local government/police/other) to do "something" about road injury. But how do we decide the priorities. Part of the discussion could involve the following ideas:

- Direct resources to what I think is the most important
- Direct resources to what elected representatives think is most important
- Direct resources to what the public think is most important
- Inform a debate on resources using something approaching objective evidence?
- Balance the needs with the likely effectiveness of any intervention (maybe one problem is more common, but I don't really know how to fix it)

Hopefully, we arrive at a position where (a) we see some need for data to support a decision and (b) we acknowledge that these data are unlikely on their own to make the decision for us. In this case, one major complication may be that some problems are more amenable to our interventions than others. Even if one crash type were more numerically common than the other, maybe we don't have an effective remedy, and would be better spending our money on something that makes a difference. However, we should be basing our decisions on data about how common things are, and how effective treatments are. We shouldn't be basing it on guesswork or anecdote.

Learning outcome 1

GAISE 2010: 1

Data beat Anecdotes

1.2 Activity 2

We first need to address the question of data beating anecdotes.

Do we really use data?

The plural of anecdote IS NOT data

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

With a mature audience, we need to do this in a way that acknowledges the fact that no administrative data (if any data) give a perfect representation of reality. We therefore have to move to a discussion of decision making in the face of imperfect data.

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.

The aim of of this slide is to prompt a more informed discussion about the actual data that concerns course participants. What does actually get recorded? How does it get recorded. In injury prevention, one "Gold Standard" source of data are the police collected, so-called "STATs 19" data.

- Who: collected by the police, in response to information about a collision (so for example, minor collisions involving uninsured drivers may never get reported to the police)
- Why: apart from the urgent human needs, police involvement is directed towards determining whether an offence has taken place, and whether a prosecution is possible. This might limit the value of the data for preventative action (if I get hit as a pedestrian / cyclists, I might not care that it was the car driver's fault. I might care about ways I can prevent it happening again)
- How: a long, complicated form, sometimes collated in response to a public report a while after the collision occurred. All evidence is retrospective and "first impressions"

Activity

Using other examples provided by the course participants, discuss the limitations of the data using Best's Who/How/Why.

2 The way we collect data

(Maybe it's time to find a better example). This case study was used by Sharon Lohr in her excellent (one of the best around) textbook on survey sampling. There's also an interesting parallel with road injury. Few people admit to being a "below average" driver. Likewise few people admit to being a "below average" lover, so it does seem to have some audience resonance. We start by stating a few bare "facts" from the summary of results.

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)

The aim of this slide is to prompt a discussion about survey methods. Hopefully, most people would accept that 4,500 is quite large by survey standards. It is though well worth prompting a discussion about what we mean by "quite large", as well as asking about the size of surveys people rely on in their own practice.

Activity

Discuss the validity of results based on a sample size of 4,500. Discuss the sample sizes used by participants in their own practice.

Once we've exhausted discussion about sample size we can move on to consider the results.

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?

This should fuel a good open ended discussion. The findings out to be a little challenging. It is worth letting this discussion run, with careful prompting. I suppose it's bad pedagogy to let anyone comprehensively challenge a theory of human relationships before you display the next slide, but the more people try to engage with the results the better

Activity

Discuss the findings of Shere Hite's survey.

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

It is this last slide on response rate that really provides the key to the results presented. It is worth giving more information on the way the data were collected. Actually 100,000 surveys were sent out, only 4,500 came back. It may well be that those 100,000 surveys were sent to an essentially representative sample frame (alhough that is a little unlikely). But the key point is that we have a 4.5% response rate. In addition, there were 127 essay type questions. So the concluding question we leave with is "will the 4.5% of respondents (who are willing to write all those essays about intimate aspects of their life) be typical of US women?"

Activity

Discuss the validity of the conclusions in the light of the low response rate

Hopefully at this point the audience will be ready to discuss the next point.

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?

First, we need to make a technical point (about non-response) and the value of random sampling. It can be a strange idea to think that randomly selected people can "represent" a population.

But secondly, we need to think carefully about the "population to which these results apply"

What's the population

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

Activity

Discuss the target "population" which is being examined by this survey. Consider examples provided by the course participants.

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

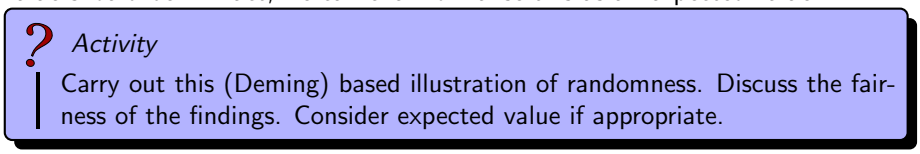
The slide above is typical of one that has been submitted by a course participant in the pre-course activities. The aim now is to discuss carefully what this headline is telling us

3 Randomness is Natural

I've just bought myself a full size Alan Sugar facemask to make this next exercise a little more interesting. However, the basic idea is one of W.Edwards Deming.

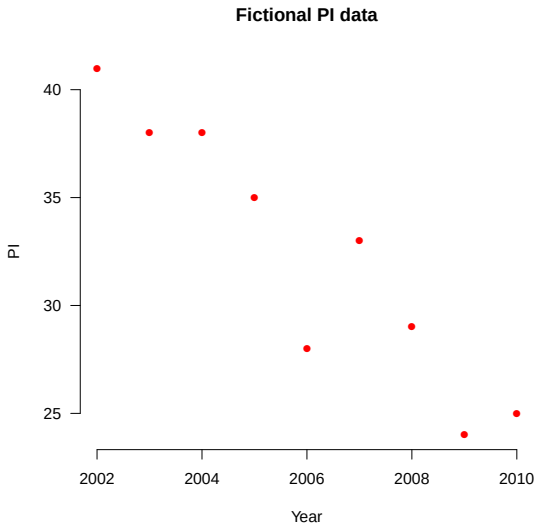
- Get (about) 8 volunteers
- Give them a task. I have used sampling beads from an urn ("make batches of yellow beads", I have used die ("get a five or a six") and I have used coins ("throw the coin eight times and get at least six heads").
- Collect results from the volunteers. Put on the Alan Sugar mask and "fire" the worst performing individual. Promote the best performing individual (who is now called Karen or Nick). Maybe put a big "D" hat on anyone who almost got fired. Run another round.
- Fire/promote/reward as necessary
- Repeat until there is a winner, who gets a bag of chocolates or other token prize

At this point we can have an interesting discussion about "fairness". One advantage of the dice/coins is that we can fairly easily get a sense of what a typical value should be. In fact, we can even formalise this as an expected value.



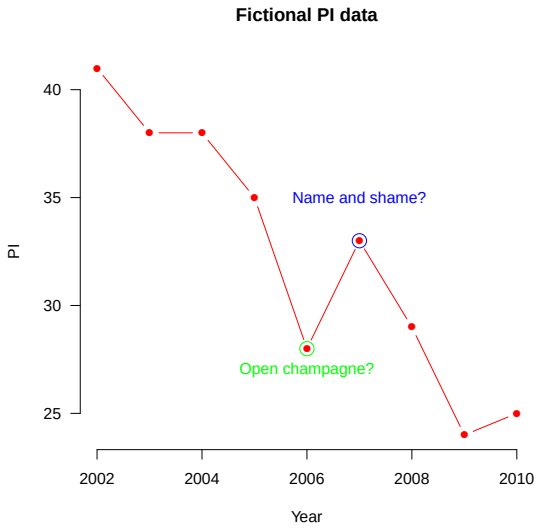
We close this session with some fictional (or some real, if it's available) Performance INdicator data.

Fictional performance indicator



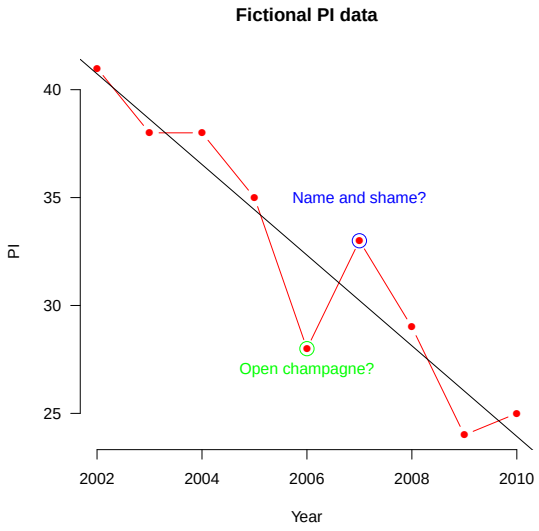
This kind of slide is familiar (maybe we don't even need the trend to start with)

How we normally use these data



And this kind of treatment is familiar.

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



So the final slide should prompt a bit of a discussion. How do we determine trends (and should they be rewarded for the trend?); how do we determine blips. I've often ended up explaining things like CUSUM charts at this point - maybe something should be added. We want to avoid cynicism (it's all randomness). But most of the course participants I've had to date are the subject of performance monitoring and perhaps would expect the producers to give them the information as a CUSUM chart.

> ? *Activity*
>
> Discuss the importance of distinguishing signal from noise (randomness from substantive)

We can do a simple experiment to try to persuade people we have models for randomness. Get everyone to toss six coins, and record the number of heads they get (if it's a very small group they may have to do this twice).

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

We can collate the results as a bar/tally graph, and compare them with a simple theoretical model (the binomial)

Heads	People
0	
1	×
2	× × ×
3	× × × × × × ×
4	××
5	
6	×

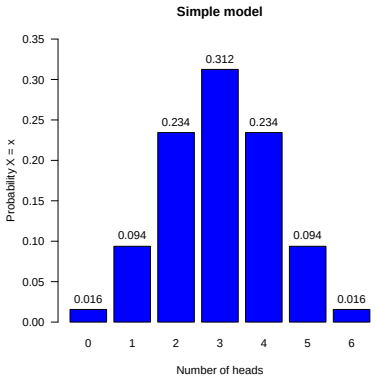
We can compare this flipchart/whiteboard with the theoretical model below

Activity

Get everyone to carry out a coin tossing experiment. Collate the results. Compare them to the "theoretical" model.

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

4 Contextual variables

We are going to "drift" towards causation (it's such a difficult subject). Again, it helps partipants reflect on their own practice to think about contextual variables that are relevant to them.

Your task is to consider all the things you could potentially measure which might have a bearing on your practice. In a road safety context for example we might consider weather, vehicle safety, highway design, training programmes.

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)

Activity

Collate lists of relevant contextual variables for the participants practice

To make this illustration a little concrete, here are a number of possibilities for road injury prevention.

For road injury prevention

- Newspaper articles
- Accident signs (requests for witnesses, "hotspots")
- Insurance
- 999 calls

- STATs19, Ambulance, F&R, HES, A&E
- Other

Contextual variables

- New / existing driving licences
- New / existing cars; car tax, MOTs
- Trafic surveys
- Perception surveys
- Demographics
- Car tax
- Census
- Car manufacturer surveys
- Many more

We had quite an interesting time teasing this out (for example newspapers might report insurance figures, so what's data and what's not).

4.1 Lurking variables

As part of our drift towards causation, we wanted to talk about lurking variables.

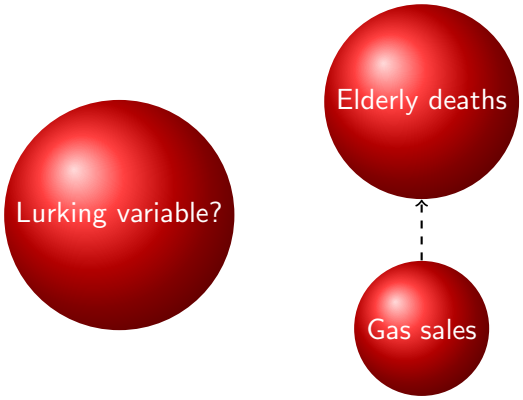
We now run through a series of visuals which illustrate the idea of a lurking variable (epidemiologists call this a confounding variable - I prefer "lurking" as it's less jargon sounding).

Illustrating Lurking variables

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

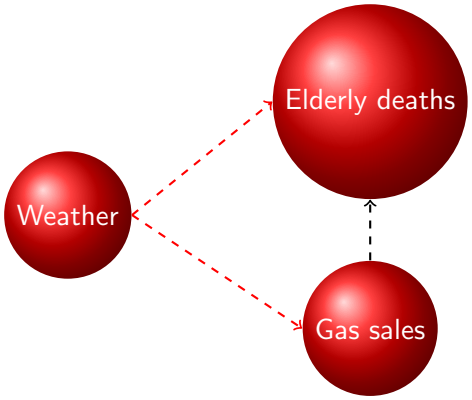
We set up one slide to explain the general idea.

Lurking variables



Then we illustrate it with a rather famous (if overused) example.

Weather as a lurking variable



And then explain the idea of "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.

Finally, we summarise the basic idea.

The next step is to consolidate this, using examples that have been submitted by the course participants.

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

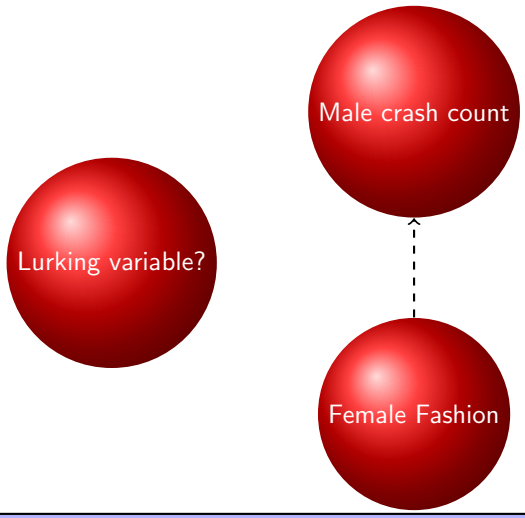
> "The scorching heatwave in early July caused road accidents to soar because male drivers were distracted by womens' skimpy outfits, according to insurance claim figures."

Source: Telegraph Online Posted 7:00am BST 30th Jul 2010 http://www.telegraph.co.uk/motoring/news/791786 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:



Activity

Using material provided by the participants, consider possible lurking variables for any apparent associations seen.

For the Sheila's wheel example, we came up with many possible lurking variables (essential all to do with summer, e.g. in summer you get dazzled, you travel to the beach and so on). Another example could be Speed and Accident rate. A lurking variable could be "Young drivers" (who have various bad practices including dangerous overtaking and tailgating)

5 Experiments

I'm almost frightened of using the phrase "causation", but we have to cover the topic.

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.

In my (biased and limited) experience, I'm usually dealing with people who have observational data (and quite often it's secondary observational data). They also want to understand causation. And, as Oscar Wilde said, (approximiately) sometimes you have to pay attention to circumstantial evidence, such as when you find a fish in your milk. So saying that observational studies tell us nothing about causation is going to be very challenging idea indeed. It also might suit some people to be able to dismiss any and all observational studies they don't like, and revert to their common sense as a guide to causation.

So I think we need to travel this topic carefully.

There are two sets of materials. One is that I run an experiment, based on an idea passed on to me by Mark Kent of (now) Solihull College. This needs to be timed carefully.

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

Possibly this is bad practice. I run this chocolate experiment with numbers such as these:

Before 3,1,5,4,3,8,5,6

After 8,2,5,2,6,4,5,9

Generally (but not always) we get better performance second time around. The aim of the exercise is to discuss whether we think we've proved the idea that the sugar in chocolate improves performance in a memory test.

Usually, the discussion converges around the idea that there is a training effect (second time round the students know what on earth was going on). It is possible to use diabetic chocolate to entirely rule out the idea that sugar had anything to do with it. I also forgot the chocolate once and had to use water with a story about brain hydration. You usually get some very good suggestions (at a Masterclass, a year 9 pupil told me they did better because I had told them they would do better). But essentially, this little demonstration usually gets us to the idea that we needed a control group (and maybe they should have a placebo of sugar free/caffeine free chocolate).

Activity

Carry out the memory test/chocolate eating experiment.

Partly as a time filler, partly as an explanation of *why* randomisation works, I

run through the following slides while we are waiting for the 12minutes to expire.

Without getting into the jargon (i.e. never mentioning Simpson's / Yule's paradox) we run through a few slides which demonstrate (again) another lurking variable problem.

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

The first slide should also allow room for discussion on what we mean by a good treatment effect.

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

The second slide is a good place to ask people what is going to happen when we show the next slide. A few smarties guess that we are playing tricks on them, but provided we focus on reasoning people should be clear that if headline figures go one way, and females go another, then males should very strongly match the headline direction.

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?

It can be good to run through some of the arithmetic just to verify what's going on.

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

Nowadays, following advice given in a causeweb seminar, I never mention the name of this paradox (too much jargon). I also try to get the group to work out how this has happened before I give the answer away (nearly always someone spots that males and females have taken different numbers of remedy / control). Rarely someone makes the full link (that females recover better from flu than males).

I then try to conclude the chocolate experiment before moving on. The next two slides summarise everything here.

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

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.

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.

The conclusion here is that in some circumstances, there ought to be data from controlled experiments on the effectiveness of interventions. Perhaps such studies should guide our practice, rather than our own tinkerings.

However, determining causality in order to understand why (bad) things happen is a little more complex, and may require futher discussion.

Post course work

-
- Motorcycle crashes and ownership over time.
- Darrel Huff - some misleading visuals