So you think you have all the data?

Causes and consequences of selection bias

David J. Hand
Imperial College London
and
Winton Capital Management

16th September 2016

BACKGROUND

The promise of big data:

> *"Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data,* ***the numbers speak for themselves.****"*

Chris Anderson *Wired* in an article called '*The end of theory: the data deluge makes scientific method obsolete*'

and endless similar by others

But things are not that simple

Big data carries risks

The risks associated with ‘small data’ and more

Today I will discuss just one of those risks

SELECTION BIAS

OUTLINE

Where I first met selection bias

The ubiquity of selection bias

Some drivers of selection bias

What to do about selection bias

Some more examples

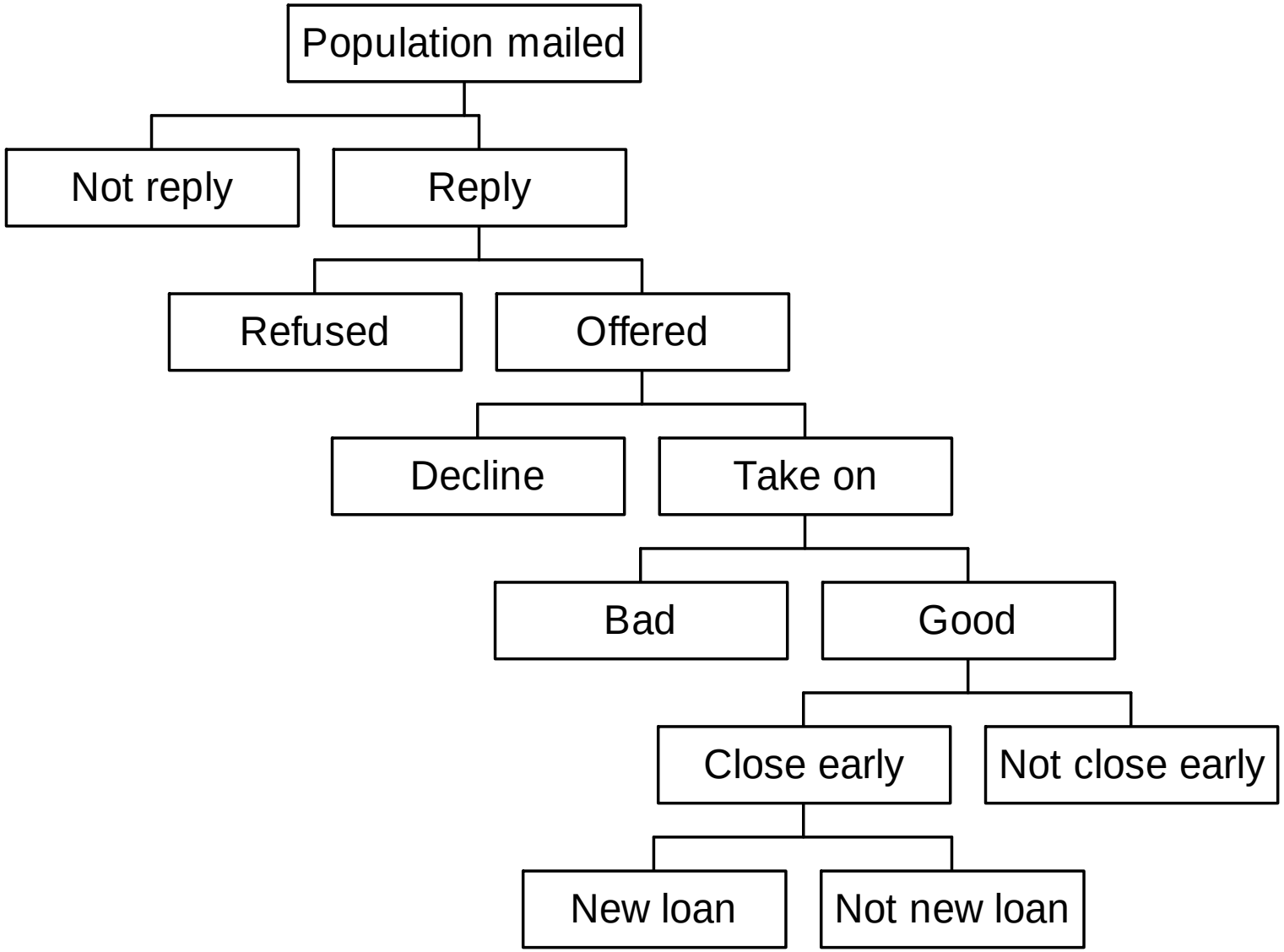
Conclusion

WHERE I FIRST MET SELECTION BIAS:

1984 consultancy project for a UK consumer bank

The task: could I build an improved scorecard?
Available data: sample of people with known descriptors (application form) and with known (good/bad) outcome

The problem: the available data were not a random sample (or even a probability sample) from the population of future applicants:


Population mailed
Not reply
Reply
Refused
Offered
Decline
Take on
Bad
Good
Close early
Not close early
New loan
Not new loan

e.g. Number of weeks since last CCJ

(from a different data set)

Weeks	% G in whole pop	% G in design data	Ratio .
1-26	22.7	44.6	1.964
27-52	30.4	46.9	1.542
53-104	31.4	49.2	1.567
105-208	37.8	55.2	1.460
209-312	42.6	63.1	1.481
> 312	55.6	69.2	1.245

SELECTION BIAS IS UBIQUITOUS:

Example 1: Potholes

- *Streetbump* smartphone app
- Detects potholes using accelerometer
  and emails location to local authority using GPS
- "Big data", but no sophisticated computation or analytics

SELECTION BIAS IS UBIQUITOUS:

Example 1: Potholes

- *Streetbump* smartphone app

- Detects potholes using accelerometer
  and emails location to local authority using GPS

- "Big data", but no sophisticated computation or analytics

- ***But*** lower income people less likely to have smartphones
  and cars, older people less likely to have smartphones, ...

→ streets in richer areas get fixed

Example 2: Hurricane Sandy

20 million tweets between 27 October and 1 November 2012

But a distorted impression of where problems are:

- most tweets came from Manhattan
- few from “more severely affected locations, such as Breezy Point, Coney Island and Rockaway”

- because of relative density of population/smartphones
- because power outages meant phones not recharged

→ distorted impression of where the damage occurred

Example 3: Publication bias

Relevant factors include:

- tendency not to submit negative results (file-drawer effect)
- positive results are more interesting to editors;
- anomalous results may be regarded as errors, and not submitted;

In an exploration of publication bias in the Cochrane database of systematic reviews:

"In the meta-analyses of efficacy, outcomes favoring treatment had on average a 27% ... higher probability to be included than other outcomes. In the meta-analyses of safety, results showing no evidence of adverse effects were on average 78% ... more likely to be included than results demonstrating that adverse effects existed."

Kicinski et al 5015

SOME DRIVERS OF SELECTION BIAS:

1) Natural mechanisms

Abraham Wald and the WWII bomber armour

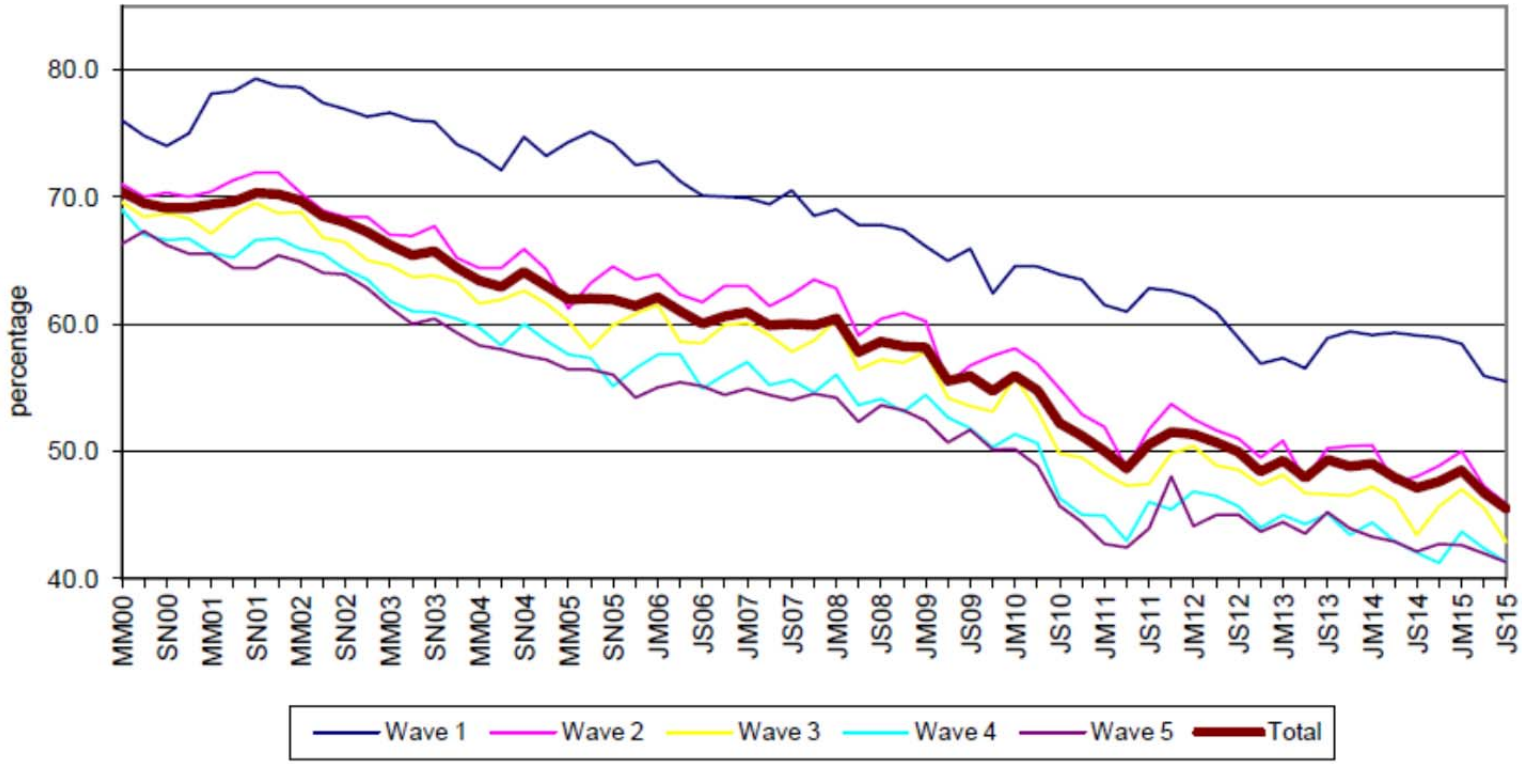
The bullet holes in returning bombers showed where they could be hit without bringing them down

A lesson for business schools? Look at the failures, not the successes

Francis Bacon

"when they showed him hanging in a temple a picture of those who had paid their vows as having escaped shipwreck, and would have him say whether he did not now acknowledge the power of the gods — 'Aye,' asked he again, 'but where are they painted that were drowned after their vows?' "

2) Non-response, refusals, and dropouts



LFS quarterly survey wave-specific response rates:
March-May 2000 to July-Sept 2015

http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-force-survey/index.html

3) Self-selection

(i) The magazine survey which asks the one question: *do you reply to magazine surveys?*

(ii) The *Literary Digest* disastrous prediction that Landon would beat Roosevelt in the 1936 presidential election

Standard explanation: the prediction was based on polling people with phones, who are more likely to be Republican

But this is a myth

In fact 10m people were polled, but only 2.3m replied

A self-selected sample, and in this election the anti-Roosevelt voters felt more strongly than the pro

(iii) ***The Actuary*** edition of July 2006 included an editorial which said '*A couple of months ago I invited you - all 16,245 of you - to participate in our online survey concerning the sex of actuarial offspring. ... Well, I'm pleased to say that a number of you (13, in fact) replied to our poll.*'

Particularly web-based surveys

- who replies?
- under-representation of some groups
- multiple responding

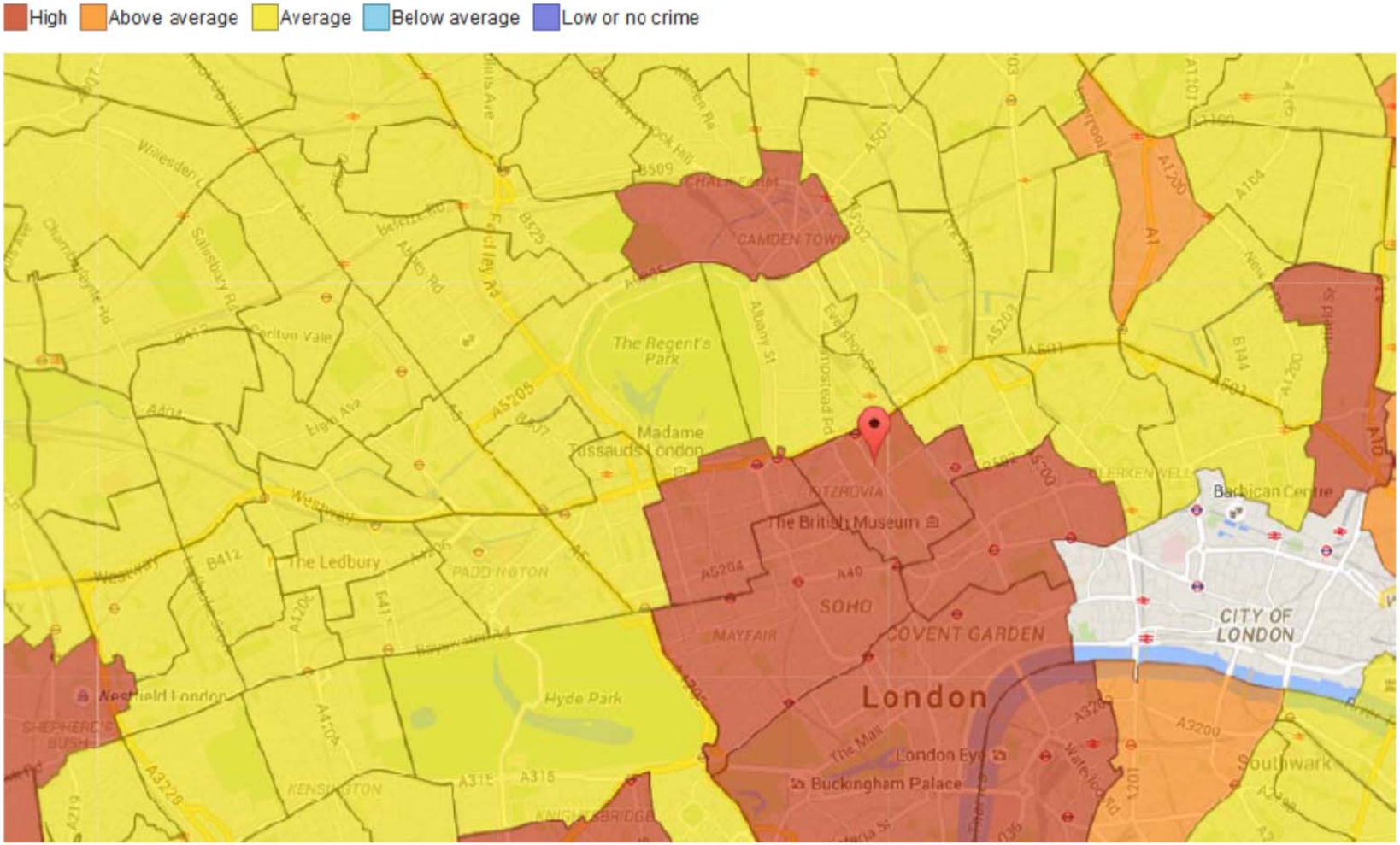
4) Feedback and asymmetric information

(i) The market for lemons

The ***buyer*** of a used car, with no further information on the vehicle in question, offers the *average* price of such vehicles

The ***seller*** can keep the better quality ones and sell only the poor quality ones

(ii) Crimemaps



But

People will not bother to report minor crime if they feel there's no point

or for other reasons

> *"More than 5.2 million people have not reported crimes for fear of deterring home buyers or renters since the online crime map was launched in February 2011"*
>
> *"A quarter (24 per cent) of people would not report a crime for fear it would harm their chances of selling or renting their property"*
>
> http://www.directline.com/media/archive-2011/news-11072011

WHAT TO DO ABOUT SELECTION BIAS:

1) Construct and stick to sampling frame

Or use “gold samples”

Draw ***some*** cases from throughout the sample space

Then standardise

2) Registers

e.g. in surveys of people

e.g.pre-registration in clinical trials

September 2004: *NEJM, Lancet, Annals of Internal Medicine, JAMA*: required drug research sponsored by pharmaceutical companies to be pre-registered in a a public database as a pre-condition for publication

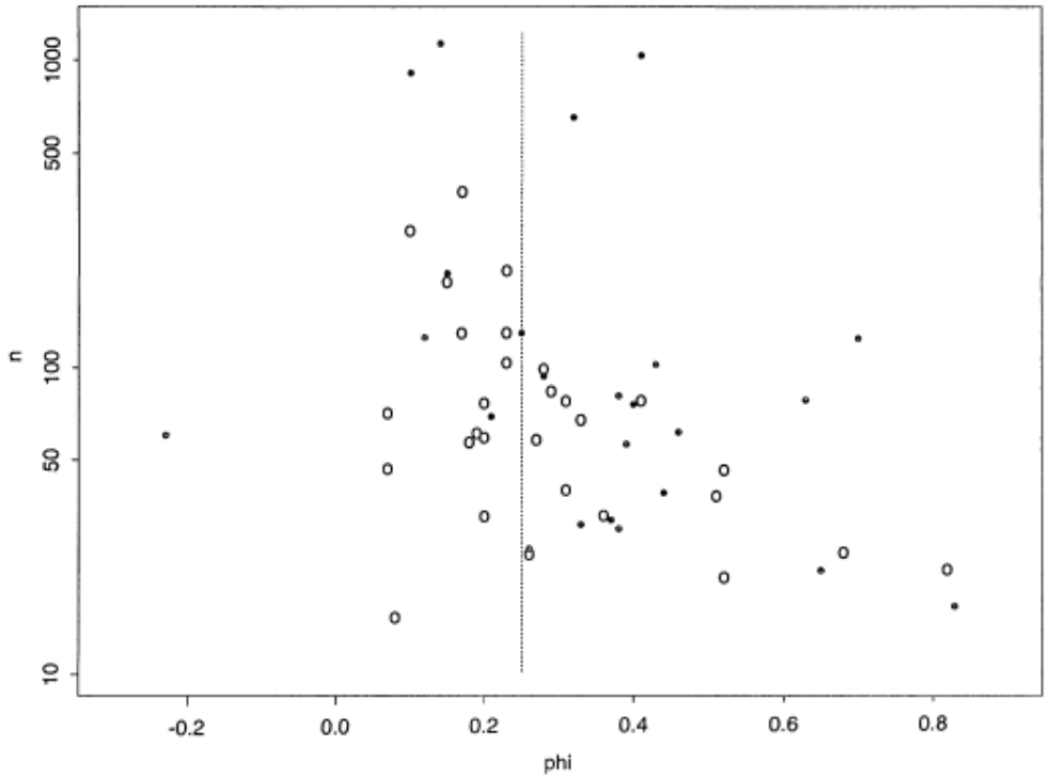
3) Detecting, e.g. publication bias

Caliper tests: ratio of reported results just above and just below the critical value associated with (e.g.) p= 0.05

Funnel plots (and tests derived from them) are based on the law of small numbers

- large studies are likely to be published regardless of results
- small studies are likely to be published only if the results are "interesting", i.e. significant

A relationship between sample size and effect size is suspicious
Hence the overabundance of plots in the bottom right of the funnel and the dearth in the left



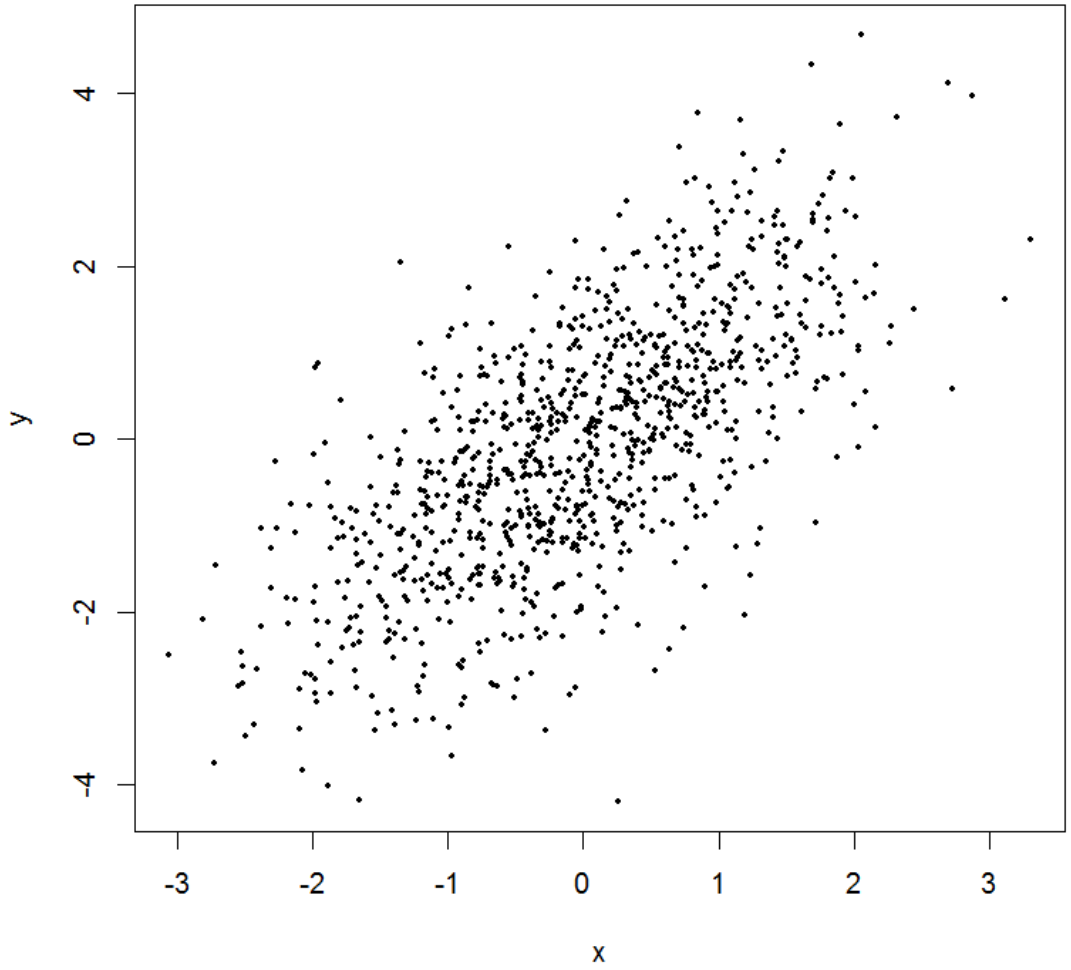
Copas 1999

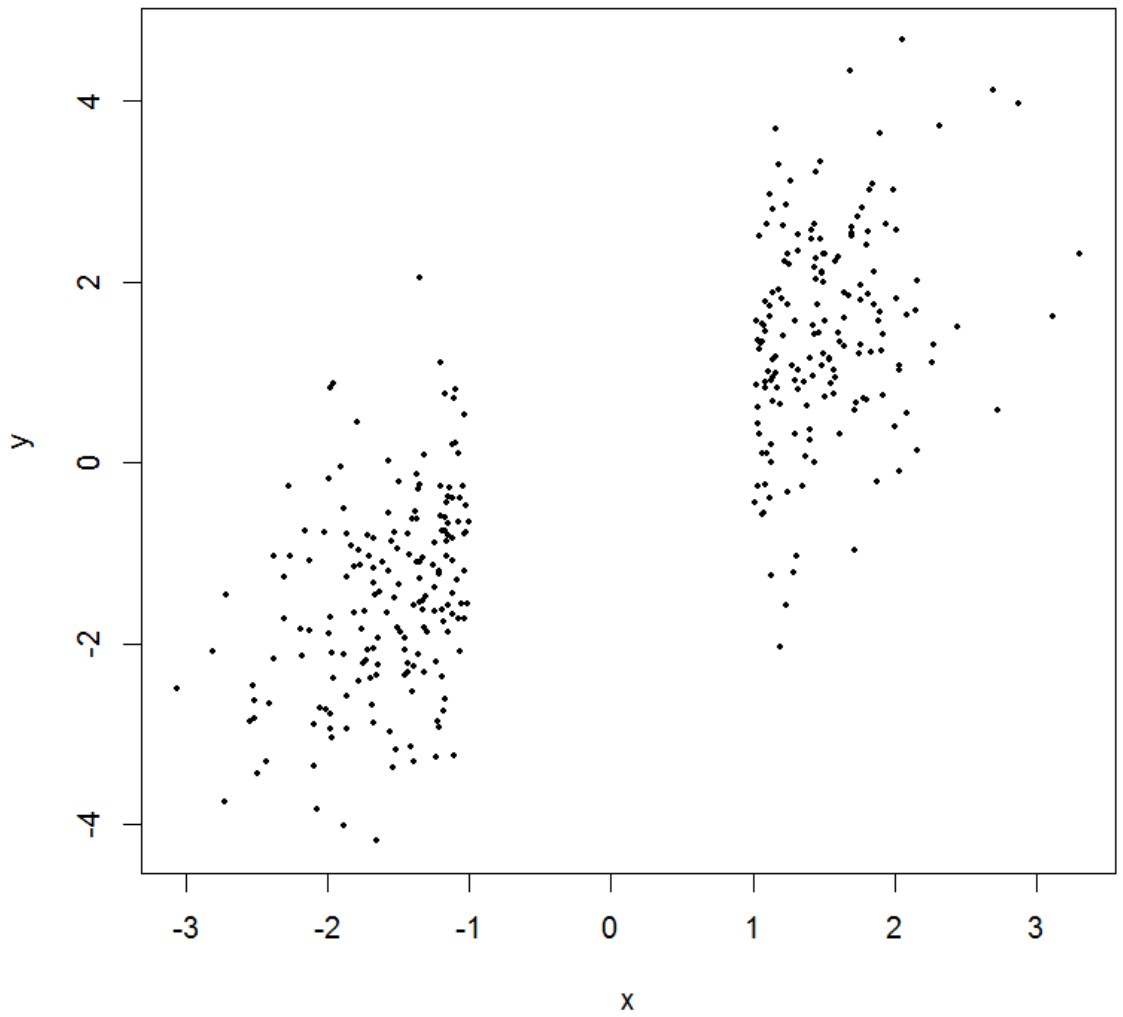
4) Model the selection mechanism

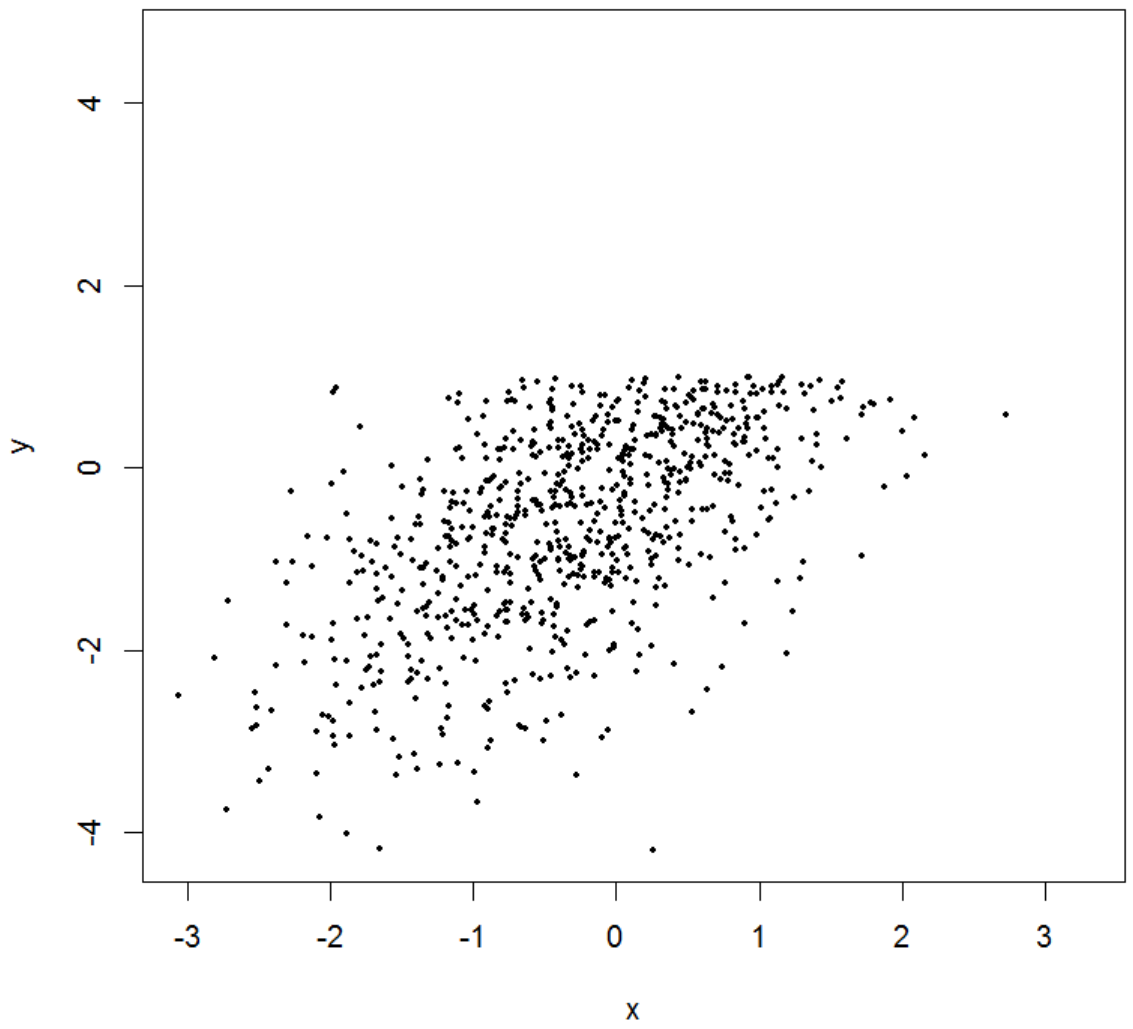
Heckman selection models (Nobel Prize)

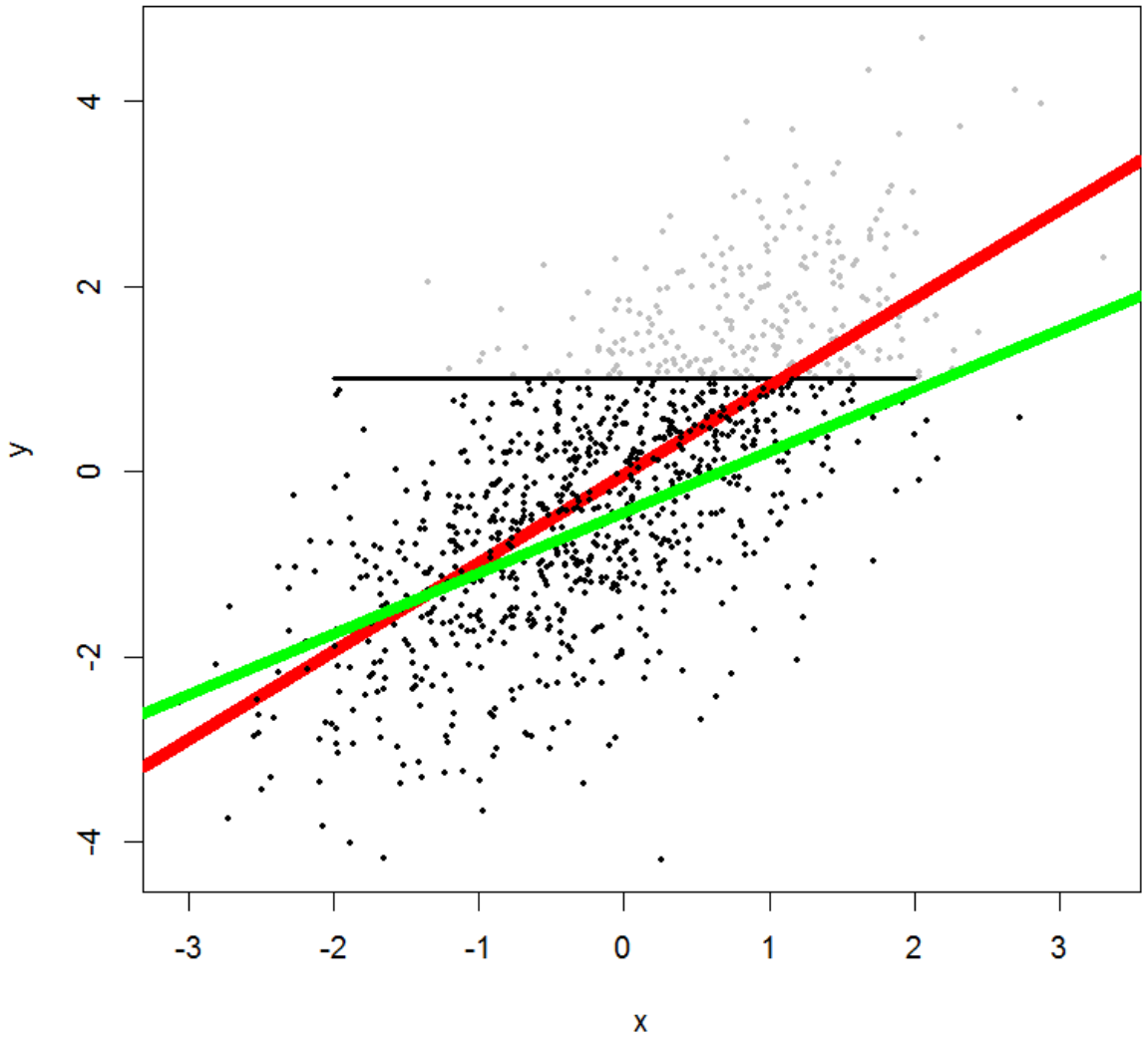
Copas publication bias correction models

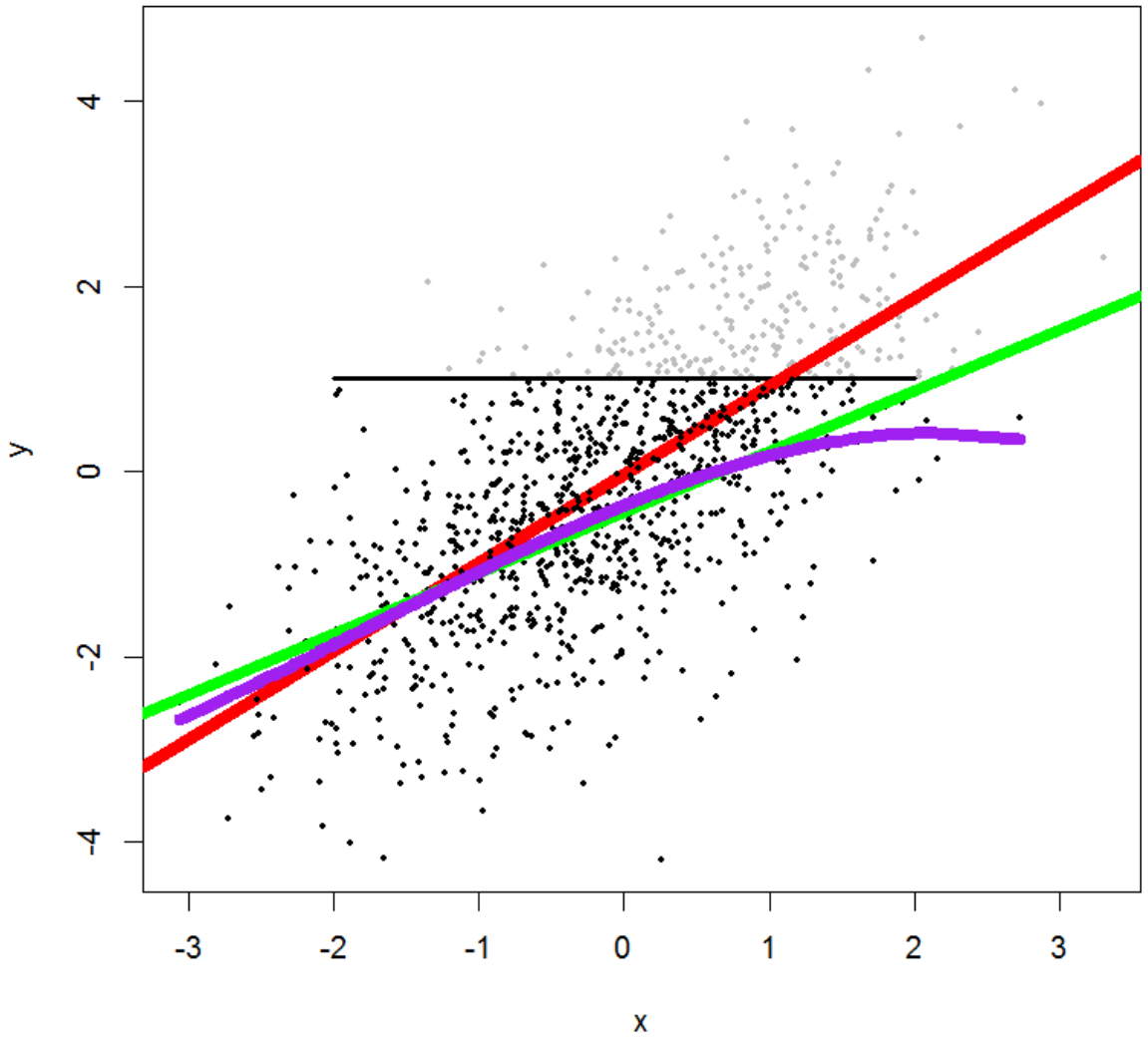
Rubin's taxonomy of missing data mechanisms


y
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
x


y
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
x


y
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
x


y
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
x


y
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
x

SOME MORE EXAMPLES:

Example 1: Comparing scorecards

Scorecards need to be monitored

- are they still performing well (populations of applicants change; economic conditions change; competitive environment changes)
- new potential scorecard suppliers appear

Leading to the **question**:

Is scorecard C (challenger) better than scorecard I (incumbent)?

The **data**: accept applicants if $S_I > 0$

Which means that I and C are treated asymmetrically

Example:

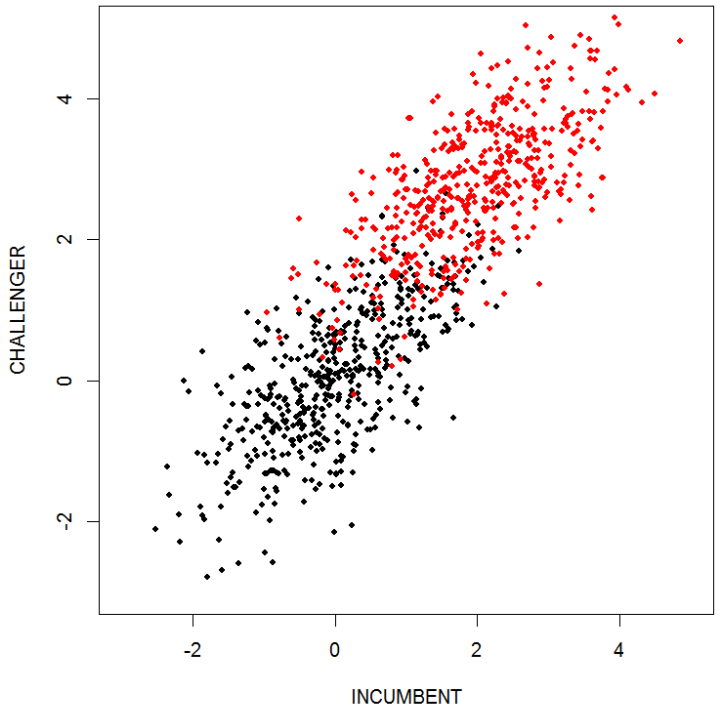
Two classes, scores (*I*,*C*)

Goods distribution: bivariate normal means (2,2), ρ=0.73, σ^2=1

Bads distribution: bivariate normal means (0,0), ρ=0.73, σ^2=1

Incumbent and ***C***hallenger have identical marginal distributions

$\Rightarrow$ ***I***ncumbent and ***C***hallenger have identical performance in separating the goods from the bads

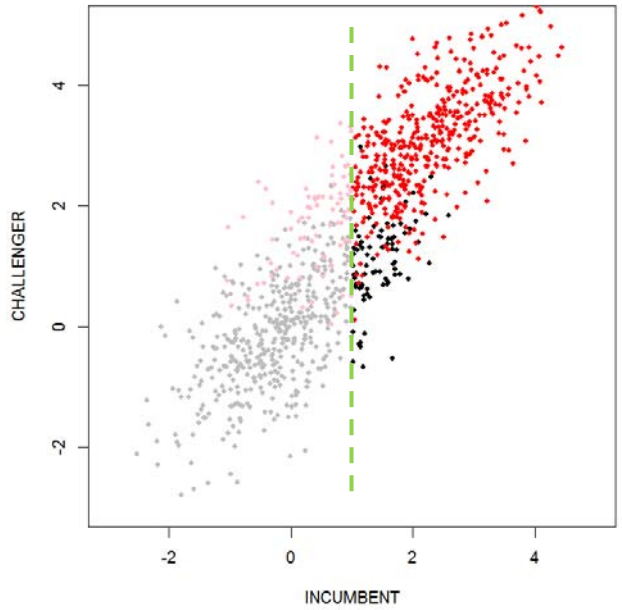


But now suppose that applicants are accepted using the ***I***ncumbent

i.e. truncate ***I***ncumbent at score 1

⇒ Incumbent distributions are ***truncated normal***

⇒ Challenger distributions are ***skew normal***



If the selection process is not included in the modelling, this means the separation of the marginal distributions is greater for the ***C***hallenger than for the ***I***ncumbent

An analysis ignoring the selection mechanism favours the ***C***hallenger

Example 2: Credit card fraud detection

Transaction stream terminated when incumbent detects a fraudulent transaction, not when the challenger does

→ data asymmetry

Standard fraud detection measures, ignoring the asymmetric data selection, favour the incumbent

But perhaps the challenger is quicker or cheaper ...

CONCLUSION:

When people say

the numbers speak for themselves

CONCLUSION:

When people say

the numbers speak for themselves

The real danger is that

the numbers might be lying

Expect *selection distortion*

Model *selection distortion*

Or risk mistaken conclusions

- *loss of money*
- *risk to life*
- *....*

thank you !