David J. Hand
Imperial College, London

38th Conference on Stochastic Processes and their Applications

Spa2015@oxford-man.ox.ac.uk

Discovery vs distortion

-

the importance of quality in learning from data

David J. Hand
Imperial College, London
and
Winton Capital Management

14 July 2015

Mathematical models are idealisations

1: On the veracity of models
2: What are we trying to do?
3: What's the question?
4: Data quality and useful models

1: On the veracity of models

George Box:

All models are wrong; some are useful;

1: On the veracity of models

George Box:

All models are wrong; some are useful;

but to be useful they must be wrong in the right way

1: On the veracity of models

George Box:

All models are wrong; some are useful;

but to be useful they must be wrong in the right way

Jorge Luis Borges *On Exactitude in Science*

The genius of scientific discovery often lies in deciding what matters and what is irrelevant

e.g. air impedes the rate of acceleration of falling bodies, light exerts a force, invisible electrostatic and magnetic fields have an influence, distant bodies have their own gravitational effect, the mass of the two players is enough to completely change the angle of deflection of colliding balls on a pool table after just nine collisions

Newton recognised that from one perspective, these are sources of error: they detract from the elegant simplicity of the underlying laws.

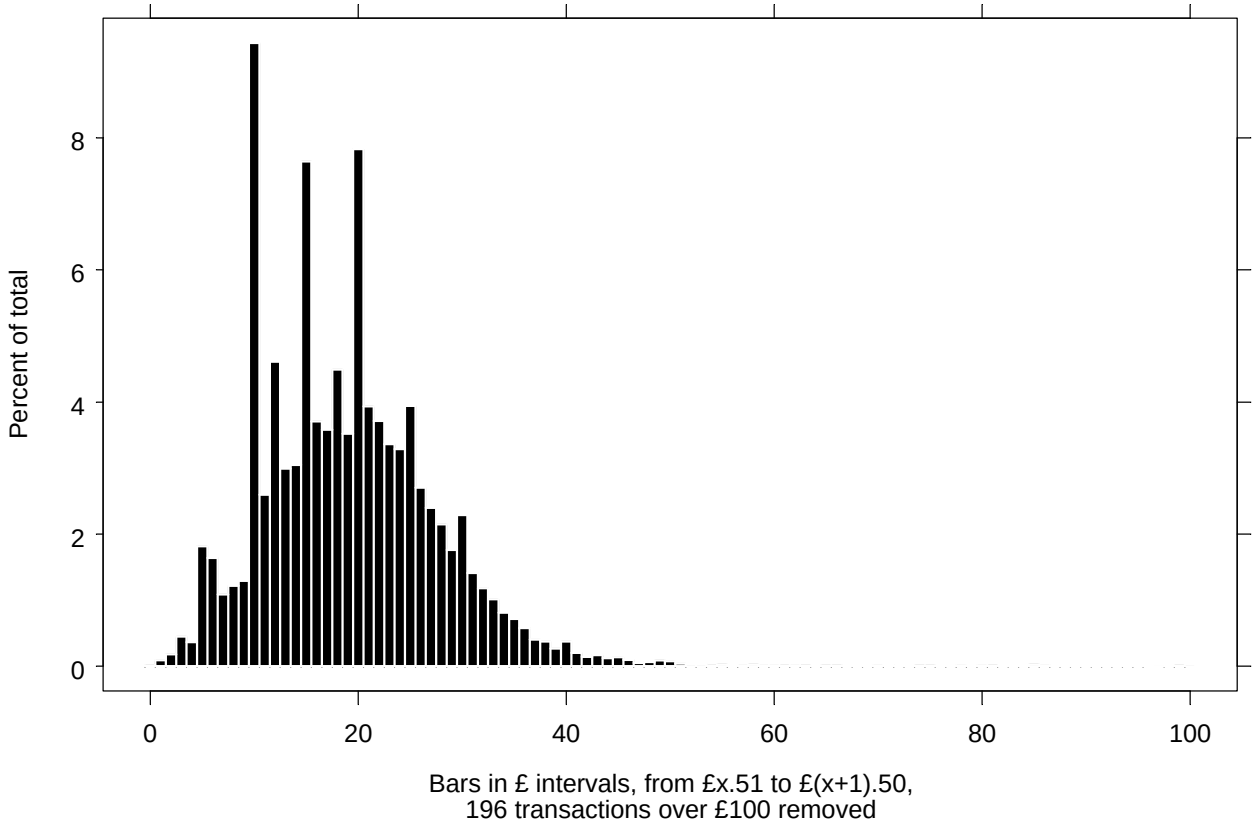
e.g. there is no chance of 10 billion customers arriving at the bank tomorrow; customers sometimes enter in small groups; customer arrival time is influenced by bank opening hours

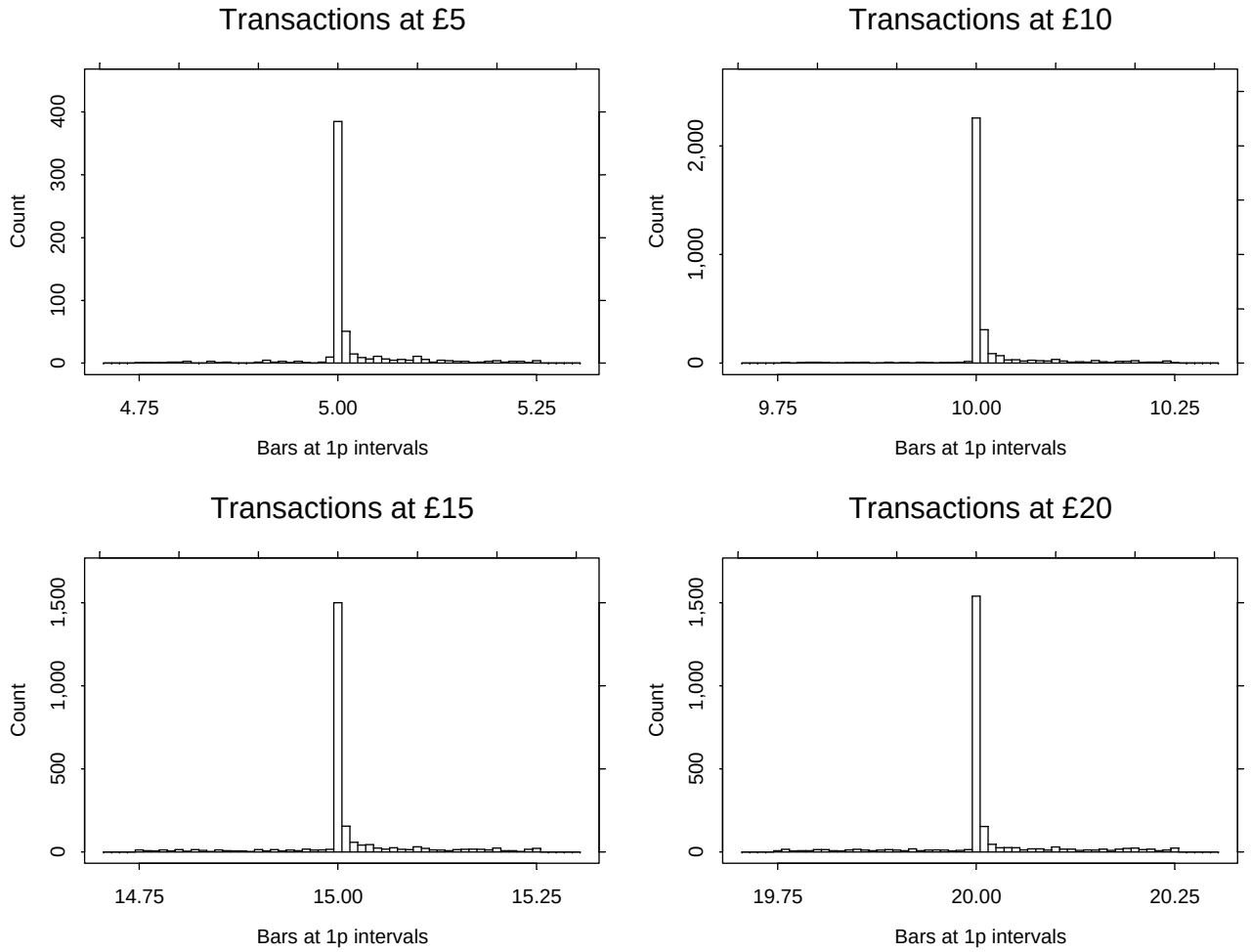
But it is often adequate to model the arrival as a Poisson process

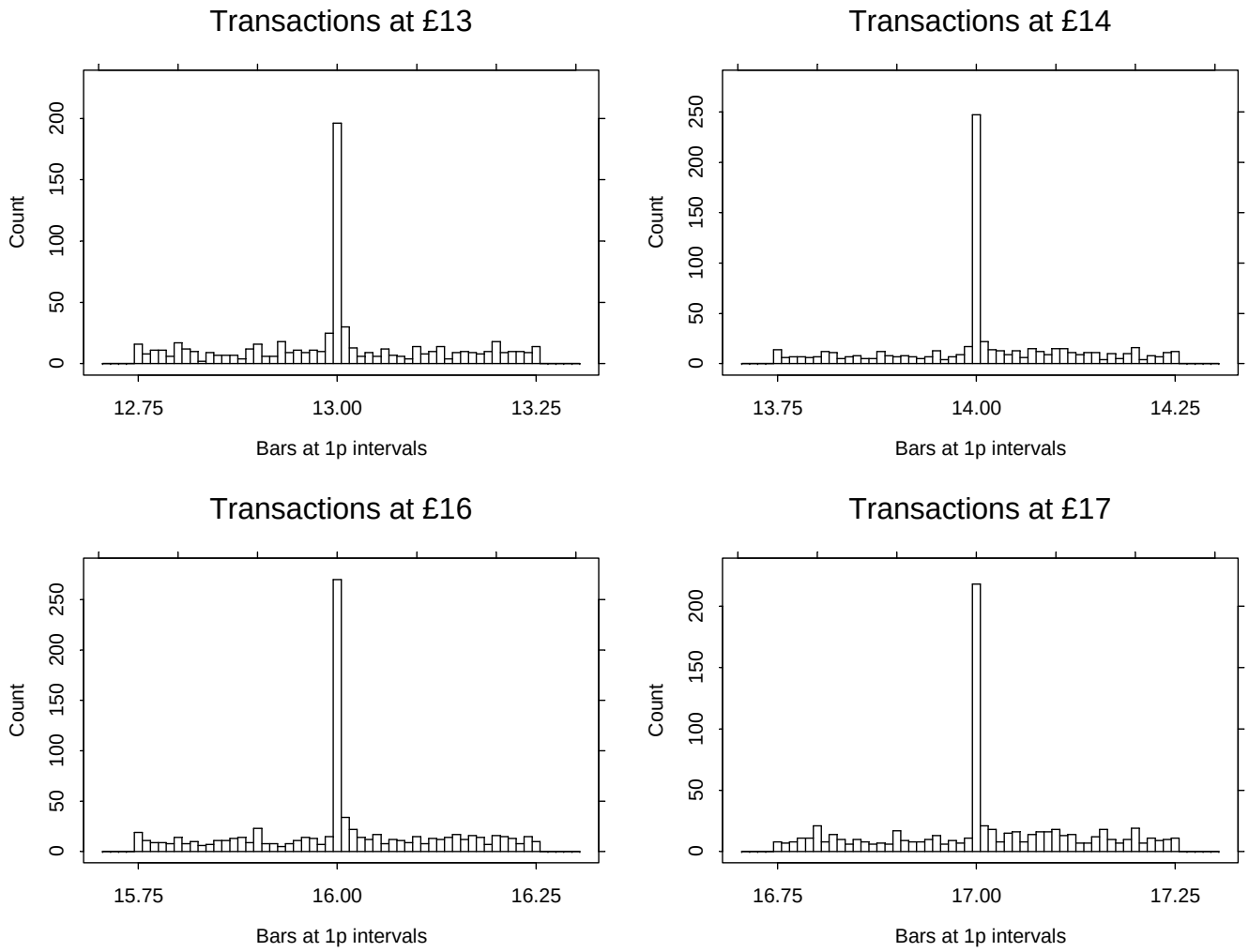
So one fundamental challenge is deciding what is ***relevant***, and what is ***error, distortion, contamination***, ...

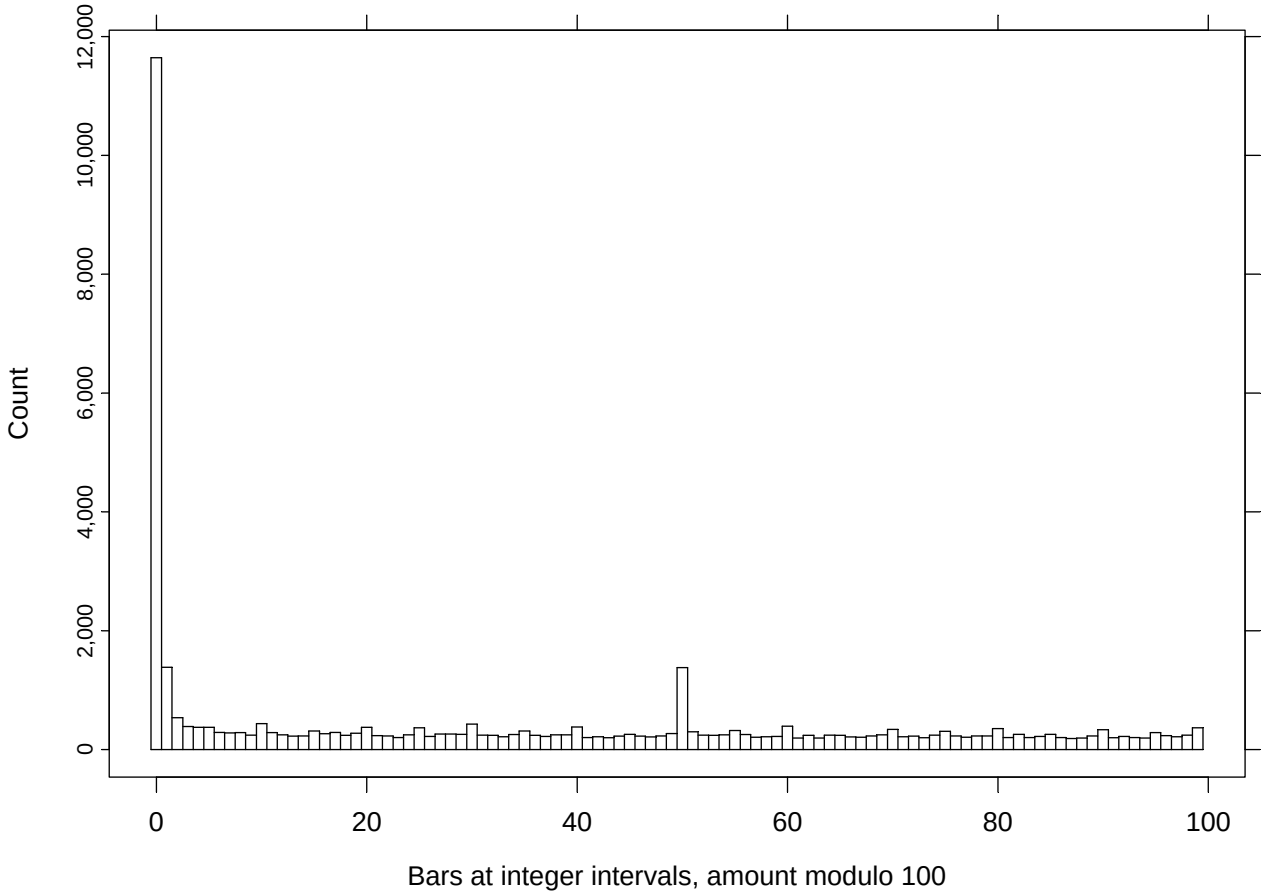
e.g. should an outlier be deleted, or is it an indicator of an important aspect of the process?

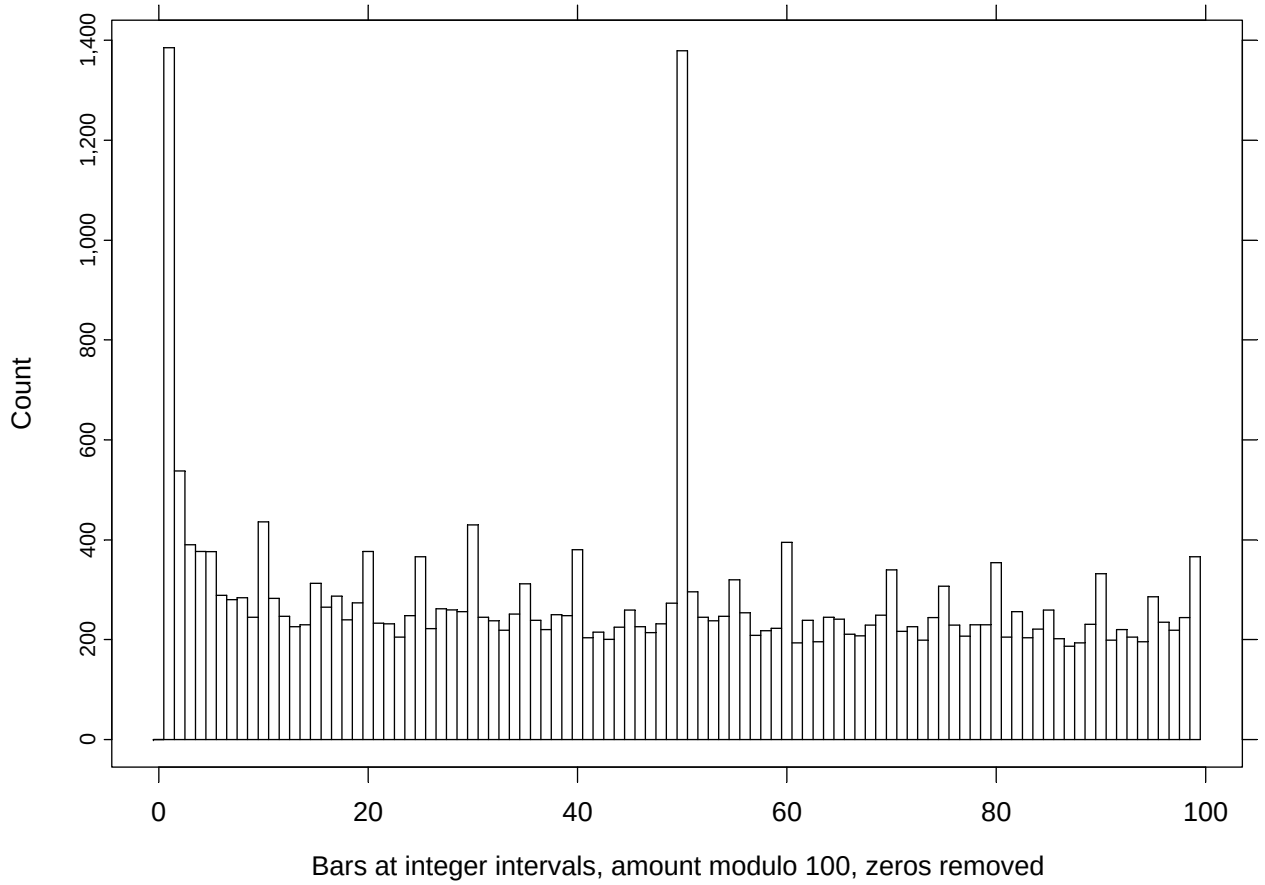
e.g. credit card usage in petrol stations




Transactions at £5
Count
0
100
200
300
400
4.75
5.00
5.25
Bars at 1p intervals
Transactions at £10
Count
0
1,000
2,000
9.75
10.00
10.25
Bars at 1p intervals
Transactions at £15
Count
0
500
1,000
1,500
14.75
15.00
15.25
Bars at 1p intervals
Transactions at £20
Count
0
500
1,000
1,500
19.75
20.00
20.25
Bars at 1p intervals


Transactions at £13
Count
0
50
100
150
200
12.75
13.00
13.25
Bars at 1p intervals
Transactions at £14
Count
0
50
100
150
200
250
13.75
14.00
14.25
Bars at 1p intervals
Transactions at £16
Count
0
50
100
150
200
250
15.75
16.00
16.25
Bars at 1p intervals
Transactions at £17
Count
0
50
100
150
200
16.75
17.00
17.25
Bars at 1p intervals


Count
12,000
10,000
8,000
6,000
4,000
2,000
0
0
20
40
60
80
100
Bars at integer intervals, amount modulo 100


Count
0
200
400
600
800
1,000
1,200
1,400
0
20
40
60
80
100
Bars at integer intervals, amount modulo 100, zeros removed

2: What are we trying to do?

Two important ***distinctions***

i: Theory-driven modelling vs data-driven modelling

ii: Populations vs elements

Underlying both of these is the distinction between *understanding* and *prediction*

One does not need to understand a mechanism to predict the outcome (e.g. driving a car)

Theory-driven models

Are essentially *theories* requiring parameters to be estimated

e.g. Newton's Laws of Motion $H = at^2/2 + \varepsilon$

- necessary for understanding

e.g. to detect dark matter from galaxy rotation

Social policy research speaks of logic models:

"*Logic models are* ***narrative or graphical depictions of processes*** *in real life that communicate the underlying assumptions upon which an activity is expected to lead to a specific result.* ***Logic models illustrate a sequence of cause-and-effect relationships****—a systems approach to communicate the path toward a desired result*" (McCawley)

Data-driven models

Based purely on empirical relationships in the data

e.g.in credit scoring the model of choice is a logistic regression tree

- the population is partitioned into segments on empirical grounds
- different logistic regression models built in each segment
- choice of variables in each segment purely empirical

No underlying theory

No psychology, prospect theory, behavioural finance, etc.

Data-driven models work because they are

- based on relationships observed in the data
    - assumption of historical continuity

Data driven models are often regarded as synonymous with "big data"

But the ideas are far from new

e.g. segmented regression in credit scoring in ***1960s***

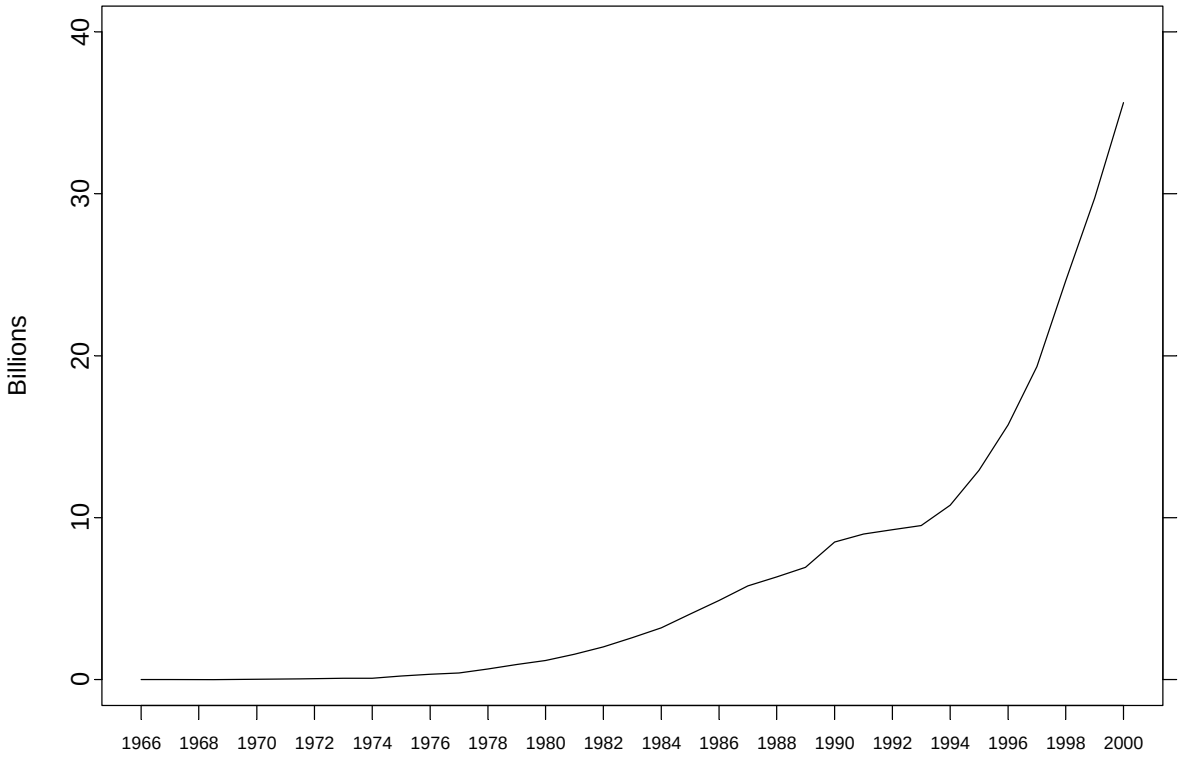
e.g. medical screening

Data-driven models are good for ***prediction*** and ***anomaly detection***

which is why they are so heavily used in some domains

Data driven models don't allow *counterfactuals*

Data-driven models don’t provide insight



Sources: BBA, CCRG (Access cards only added from 1974, Building Societies from 1996)

"There are two cultures in the use of statistical modelling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown."

Leo Breiman, 2001

"Models vs algorithms"

Distinction ii: Populations vs elements

A: Large scale, population analysis
summarise data, identify the main features, *understand*

B: Element level
make a statement about individuals, *predict*

Common misconception is that statistics is solely about A
e.g. Egon Pearson *"[statistics is] the study of the collective characters of populations"*

True that statistics is about A,
but ***false*** that it is solely about A:

Often statistics is about the ***combination*** of A and B

e.g. personal credit decisions

e.g. clinical trials

e.g. recommender systems

3: What's the question?

Three examples:

Example 1: mean or median ?

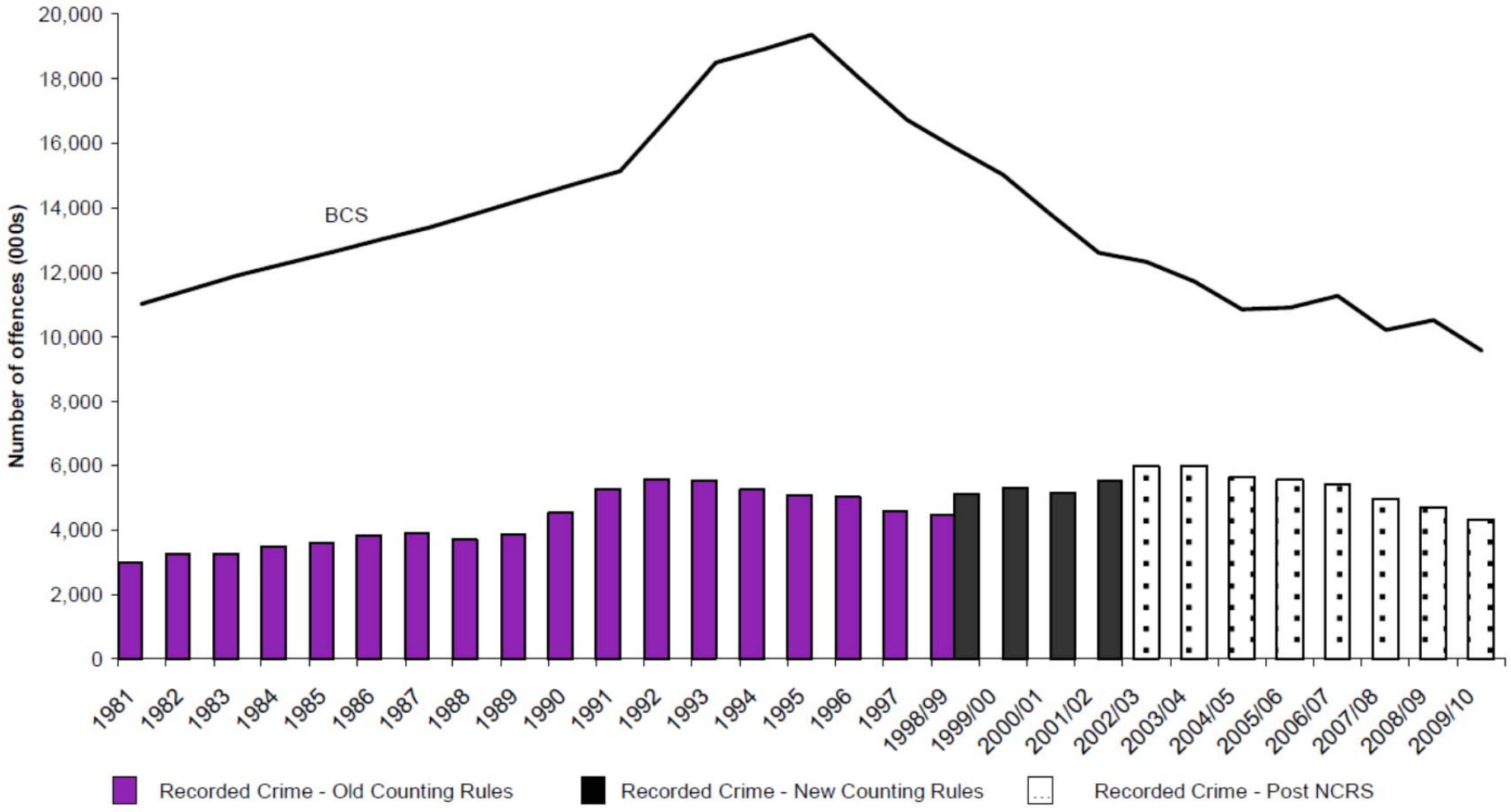
"Is it better to use the mean or the median? ... to know which you should use, you must know how your data is distributed. The mean is the one to use with symmetrically distributed data; otherwise, use the median. If you follow this rule, you will get a more accurate reflection of an 'average' value."

http://www.conceptstew.co.uk/PAGES/mean_or_median.html

An employer would like to know the mean salary, since the total wage bill is the product of that and the number of employees;

A potential new recruit will be interested in the median salary. To her the mean is of little interest, since she is very likely to receive substantially less than that.

Example 2: Crime statistics: BCS and PRC



Example 3: Which of treatments A or B is better?

Pragmatic versus explanatory studies

Pragmatic: Will the new method be more effective than the standard method, as used in practice?

Explanatory: Will the new method be more effective, all other things being equal?
(lab. conditions, physiological, biological, ...)

e.g. Treatment of cancer by radiotherapy

T1: Radiotherapy alone
T2: Radiotherapy, preceded for 30 days by sensitising drugs

Pragmatic: In T1, start radiotherapy immediately

Explanatory: In T1, precede radiotherapy with 30 days placebo

$\Rightarrow$ *different design*

Pragmatic: Guard against identifying poorer as better

'Type I error' (concluding a difference when there is none) does not matter

Explanatory: Avoid type I error

$\Rightarrow$ *different error structures*

$\Rightarrow$ *different sample sizes*

Pragmatic: Include side-effects withdrawals as treatment failures

Explanatory: Analyse only those who stick to protocol

$\Rightarrow$ *different samples analysed*

Pragmatic: Subjects should be a sample of those to be encountered in practice – heterogeneous

Explanatory: Eliminate superfluous variability – homogeneous

$\Rightarrow$ *different populations*

So to answer the question

Which of treatments A or B is better?

we need to know if the study is pragmatic or explanatory

- *different designs*
- *different error structures*
- *different sample sizes*
- *different samples analysed*
- *different populations*

4: Data quality and useful models

Claus Moser, 1979:

"Any figure that looks interesting is probably wrong"

Kruskal, 1981:

US 1960 census: 62 women aged 15-19 with 12 or more children
US 1970 census: 2,926 males aged 25-29 enrolled in first grade
US 1970 census: 2983 14-year-old widowers

"a reasonably perceptive person ... can sit down with almost any structured and substantial data set or statistical compilation and find strange-looking numbers in less than an hour."

"most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis"

Lazer et al 2014

The notion that we escape sample errors by using "all" the data is misleading:

Data typically arrive in the database by a complex socio/political/measurement process

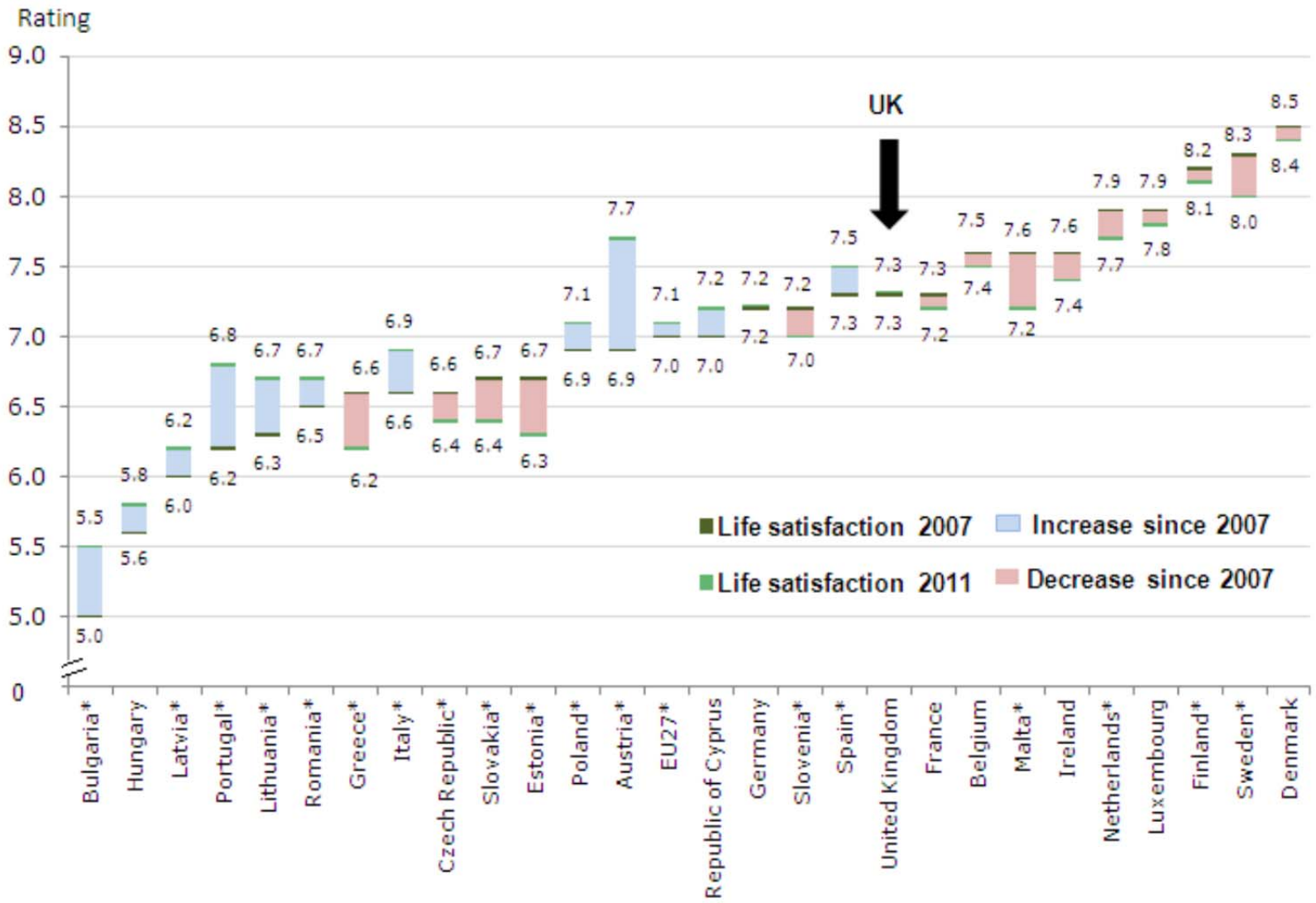
e.g. the database of all credit card transactions in a bank's database

- only those with this bank's card: how were they chosen?
- how/why did the customer decide to use this card?

Selection bias

e.g.1: Astronomy: dimmer (further) objects less likely to be detected, so underrepresented in the data

e.g.2: Credit scoring reject inference



http://www.ons.gov.uk/ons/dcp171778_319478.pdf

Feedback and gaming example

e.g.1: crime maps

> *"For many years, all forces have mapped crimes and incidents to help them focus investigations, analyse hot spots and tackle crime vigorously. The information now on the forces' websites has a different more community-focused perspective and means the public can now look at crime levels in their community simply by putting their postcode into their local police force's website."*

www.crimemaps.org.uk

“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”

Big data makes things worse

- the computer is a window between you and the data
    - one can see through windows
    - but windows get dirty

5: Conclusion

Learning from data

- theory vs data driven models
- data are generally the result of a complex socio-measurement process: what are the relevant aspects?
- data quality
- what's the question? What is the purpose of the model?