

Statistics: the science of the individual

*David J. Hand
Imperial College London*

25th February 2011

Structure

Part I: Measurement

Part II: From the individual to the population

Part III: From the population to the individual

Risks?

Part I:

Measurement

Statistics is the science of uncertainty and the technology of extracting understanding from data

Data arise from *measurements*

Notions of measurement

- have their origins in mists of time (Bible, Koran)
- are being formalised
- slowly encroach on all areas of endeavour
 - *the national well-being measurement*

Formalisation

All measurement is a mix of *representational* and *pragmatic* perspectives

Representational:

A homomorphism from an 'empirical relational system' to a 'numerical relational system'

e.g. two sticks, A , B

*Relationship (1) whether, when placed with one end against a wall, one stick projects beyond the other:
 $A \succ B$*

Not a unique mapping ...

Relationship (2) place stick A end-to-end with stick B, $A \circ B$, and find stick C which projects the same distance at this combination $A \circ B = C$

For sticks A , B , and C , choose numbers $x(A)$, $x(B)$, and $x(C)$ such that

$$A \succ B \rightarrow x(A) > x(B) \quad \text{and}$$
$$x(C) = x(A \circ B) = x(A) + x(B)$$

Very elaborate theory, pure mathematics, dimensional analysis, the natural sciences

Pragmatic:

Simultaneously *define* and describe *how to measure*

e.g. HRQoL:

- *decide to include* pain, feelings of worthlessness, lack of sleep, lack of close relationships, etc.
- *decide overall score is a weighted sum*

e.g. factor analysis

- choose what variables to include
- decide on model constraints relating latent and manifest variables

A struggle:

Gradual recognition that more abstract things could be measured

'The method I employ to this end is not yet very common, for instead of simply using terms in the comparative and superlative and purely rational arguments, I've adopted the method ... which consists of expressing oneself in terms of numbers, weights, and measures.'

William Petty, 1690

Occasional outspoken proponents throughout

'The Reader may here observe the Force of Numbers, which can be successfully applied, even to those things, which one would imagine are subject to no Rules.'

'There are very few things which we know, which are not capable of being reduc'd to a Mathematical Reasoning; and when they cannot, it's a sign our Knowledge of them is very small and confused.'

John Arbuthnott, 1692

But strong resistance to this extension in some quarters

Goethe wrote that measurement could be employed in strictly physical science, but biologic, psychologic and social phenomena necessarily eluded the profane hands of those who would reduce them to quantitative abstractions.

'I submit that any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact meaningless'

British Association for the Advancement of Science committee, 1938.

‘When you can measure what you are speaking of and express it in terms of numbers, you know something about it. When you cannot express it in terms of numbers, your knowledge is of a meagre and unsatisfactory kind.’

Lord Kelvin, 1888

‘Measures are more than a creation of society, they create society.’

Ken Alder (2002) *The Measure of All Things*

Part II:

From the individual to the population

Measurement applies at all levels:

- to individual people
- to social structures

'There is no such thing as Society'

Margaret Thatcher, 1987

But

‘They never quoted the rest. ... My meaning, clear at the time but subsequently distorted beyond recognition, was that society was not an abstraction, separate from the men and women who composed it, but a living structure of individuals, families, neighbours and voluntary associations.’

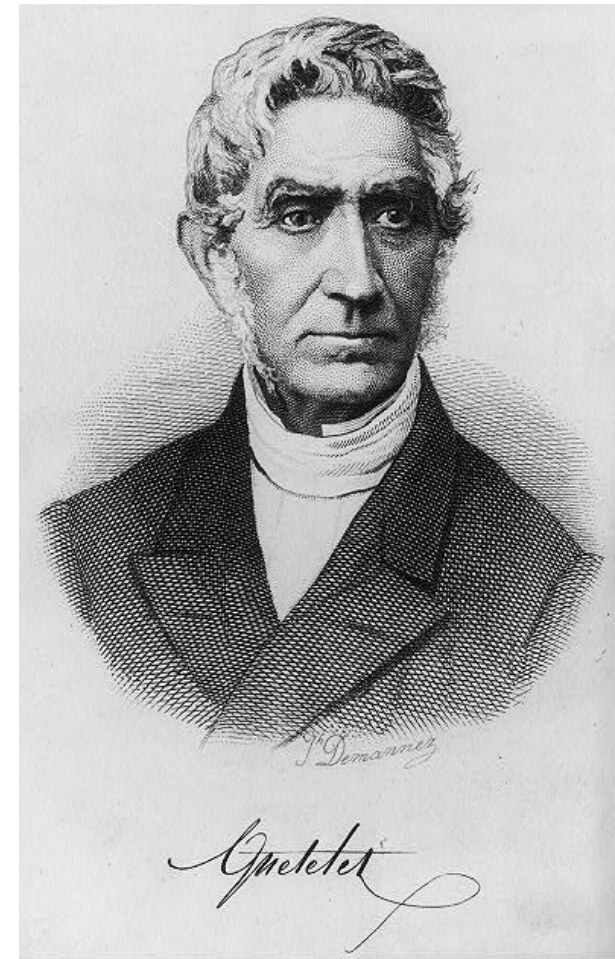
Adolphe Quetelet (1796-1874)

‘Sur l’homme et le développement de ses facultés, ou essai de physique sociale (1835)

= ‘social physics’

A classical view of physics:

*natural laws were deterministic
laws written as differential
equations*



He knew that individual behaviour could not be so represented, but believed that group averages could be

‘It is of primary importance to keep out of view man as he exists in an insulated, separate, or in an individual state, and to regard him only as a fraction of the species. In thus setting aside his individual nature, we get quit of all which is accidental, and the individual peculiarities, which exercise scarcely any influence over the mass, become effaced of their own accord, allowing the observer to seize the general results.’

Quetelet (1842, p5)

Quetelet's fundamental principle:

'The greater the number of individuals observed, the more do individual peculiarities, whether physical or moral, become effaced, and leave in a prominent point of view the general facts, by virtue of which society exists and is preserved.'

Quetelet (1842, p6)

An example of the constancy in human populations:
(Quetelet, 1842, p6)

* The following is the result of the reports of criminal justice in France, &c. :—

	1826.	1827.	1828.	1829.	1830.	1831.
Murders in general, -	241	234	227	231	205	266
Gun and pistol, - -	56	64	60	61	57	88
Sabre, sword, stiletto, poniard, dagger, &c.,	15	7	8	7	12	30
Knife, - - - - -	39	40	34	46	44	34
Cudgels, cane, &c., -	23	28	31	24	12	21
Stones, - - - - -	20	20	21	21	11	9
Cutting, stabbing, and bruising instruments,	35	40	42	45	46	49
Strangulations, - - -	2	5	2	2	2	4
By precipitating and drowning, - - - -	6	16	6	1	4	3
Kicks and blows with the fist, - - - - -	28	12	21	23	17	26
Fire, - - - - -	..	1	..	1
Unknown, - - - - -	17	1	2	..	2	2

'It would appear, then, that moral phenomena, when observed on a great scale, are found to resemble physical phenomena'

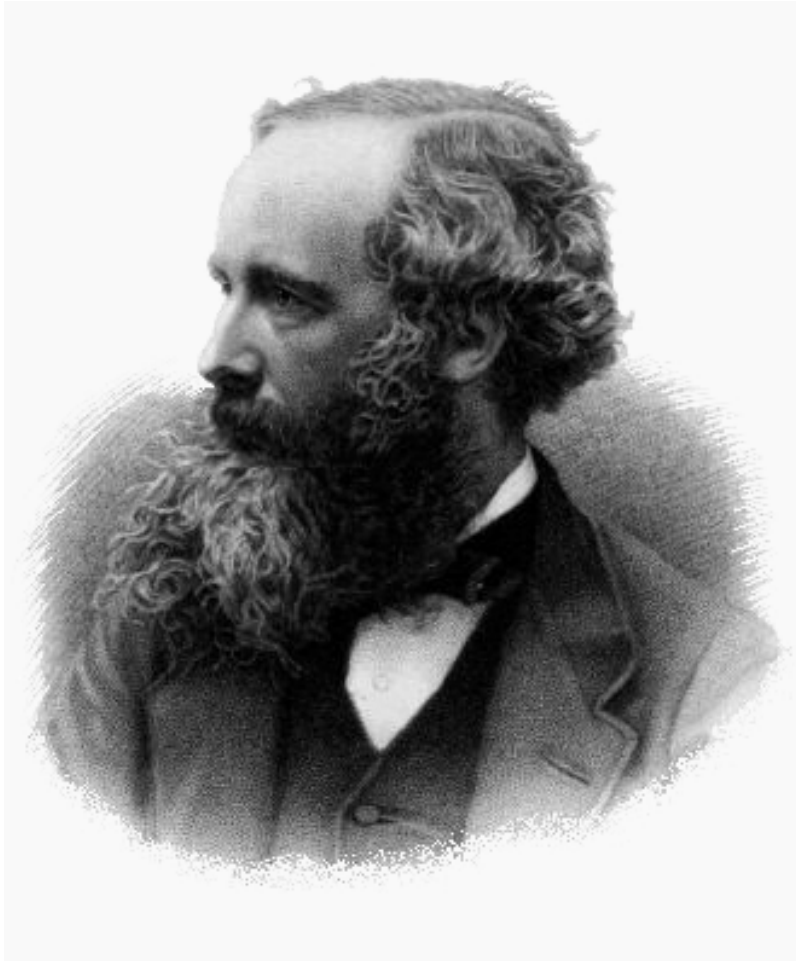
Quetelet (1842, p6)

Hence: *social physics*

But a deterministic, clockwork physics

A digression:

James Clerk Maxwell
1831-1879



Ludwig Boltzmann
1844-1906





Statistical physics was modelled on social statistics, *not the other way round*

Statistics is '*the study of the collective characters of populations*'

Egon Pearson

Statistics is '*the science of collectives and group properties*'

Maurice Kendall

Statistics is concerned with '*facts relating to communities of men which are capable of being expressed by numbers, and which promise when sufficiently multiplied to indicate general laws.*'

Statistical Section of the British Association,
when it was established in 1832

So:

Statistics is about ***aggregations***

Originally to describe ***mass social phenomena***

Now applied ***in all disciplines***

Statistics suffers from many misapprehensions

'Lies, damned lies, and

*Despite the Chief Economist at Google describing it as
'the sexy subject of the next ten years'*

*Despite RCTs being described as the greatest
advance in medical research of the twentieth century*

And

*Despite it being as much about the individual as
the mass*

Part III:

From the population to the individual

Data capture: the past

speak to respondents, record responses, enter into computer

Data capture: the present

automatic, no human intervention
accurate, tireless, instant
comprehensive: store *all* the data

e.g. credit card transactions, supermarket purchases, internet search traces, phone calls,

In many cases automatic data capture means the **entire population** of data is captured

*‘The transactional information produced by a loyalty scheme is enormously valuable if it is analysed and used well. As we’ve already said, these data are exact: they are not based on a small-scale study, a focus group or instinct – **they’re actually what is happening.**’*

(Clive Humby, *Scoring Points*, p17)

No sample inference problems

But perhaps other kinds of inference problems

Some information is collected ***explicitly***

- examination results and test scores
- application forms: education, finance, ...

But much more is collected ***implicitly***, secretly, covertly, casually

- supermarket purchases
- credit card transactions
- vehicle locations
- RFID scans
- Oyster cards
- phone calls
- CC TV

Databases

- (1) can be used for statistical modelling
- (2) but also to identify individuals

Match the two:

- (1) condense the information in a mass of people
 - to a *model* of the relationships
 - which is *simple* compared to the ?billions? of records
 - anonymising it in the process
- (2) and then apply that model to (new) individuals

Some examples:

Clinical trials:

Will drug A or B be better for patient X?

Credit scoring:

Should this applicant receive a loan?

Recommender systems:

e.g. Amazon book recommender

e.g. MeeMix music recommender

e.g. MovieLens film recommender

Predicting recidivism:

Which offenders are likely to reoffend?

Loyalty cards:

What products will ***this customer*** want to buy?

Fraud detection:

Is ***this transaction*** fraudulent?

Insolvency:

Is ***this company*** at risk?

Call centre response times:

How valuable are ***you*** to my company?

In each case

Step 1: *condense vast mass of data about large numbers of people to a summarising model*

Capture the relationship between characteristics of individuals and an outcome

A transformation from

- a description of the population in terms of the individuals comprising that population

to

- a description in terms of the characteristics of those individuals

Step 2: *match the characteristics of an individual to the model to deduce other characteristics of that individual*

Both steps necessary

Example: Capital One: A Case Study

Customer Relationship Management

Experiment with slightly different products to different customers

Constantly adapt to meet niche customer needs

Predictive selling to identify needs beforehand (50% success !)

“We record every customer interaction, every card purchase. We ... run experiment after experiment. For every action we have taken, we know what the reaction has been. If we have sent you a blue envelope or a pink one, we know which one you received and how you reacted to that. Every product that we develop is rolled out slowly and carefully. We track whether people buy something or not, and whether they ultimately use it.”

“200,000 times a day customers call our business to ask us a question, maybe about their balance, a recent transaction or about a change in their interest rate. This is then what happens before even the person calling hears the telephone ring -

The moment the last digit is punched, sophisticated networked software links with our Capital One computer to work out what is going on. Loaded with exhaustive amounts of data about our customers, our computers identify who is calling and predict why they are calling. There are then over 50 call routing options, and the best option is selected for each caller.

About two dozen items of data are analysed for each customer to help with this routing, ... The relevant data is then prepared for the customer service associate before the call even arrives in the head set.”

Example: national elections

Identify the specific individuals who can flip an unstable model

Elections often very close, and inherently unstable?

2004 US Presidential Election:

Electoral College: 53% for Bush; 47% for Kerry

⇒ 100% victory for Bush

2008 US Presidential Election:

Popular vote: 53% for Obama; 46% for McCain

⇒ 100% victory for Obama

Vote distribution across UK Parliamentary seats varies:

- some seats won by an 80:20 majority
- others won by a 51:49 majority

It is futile spending time and money campaigning in the 80:20 seat

It is futile spending time and money trying to persuade people who won't be persuaded

⇒ concentrate effort in those seats and on those voters which can be swayed

Statistical tools to detect these seats and voters

Learn the behaviour, opinions, likes and dislikes of *individuals*

A graduate who drives a BMW, shops at Waitrose, and plays golf is more likely to vote for party X than party Y

Identify characteristics of those who might be persuaded to switch

Applying market research and geodemographic tools to political campaigning

*Political policies have multiple aspects
Find a policy aspect on which you agree with the voter
and stress that aspect*

Previously the key was:
what the voter knew about the candidate

Now the key is:
what the candidate knows about the voter

No elections at national level in the UK or US are now fought without a back room of statisticians guiding actions

Without a team of statisticians (taking this individual perspective on statistics), you will lose the election

So statistics

- is about mass phenomena
summarising measurements of many individuals to produce descriptions of aggregate objects
- but it is also about the individual
using that aggregate description as a lever to make an inference about an individual

Risks?

Privacy

Data security

Data quality

Legislation: *lags behind technology*

Privacy

George Washington, 1787:

'it is necessary for individuals to give up a share of liberty to preserve the rest'

Changing in an internet world?

Data security

Loss

Theft

UK data loss

November 2007: “**HMRC** has lost computer discs containing confidential details of 25 million child benefit recipients”, containing names, addresses, birth dates, and bank accounts.

November 2007: “More than 15,000 **Standard Life** customers were put at risk of fraud after a courier lost a computer disc containing personal information”

October 2007: “A laptop computer holding sensitive information was stolen from the boot of a car belonging to an **HMRC** worker, putting hundreds of people at risk of fraud.”

Telegraph (7 Sept 2008): “A lost computer disc containing details about thousands of staff working in the **justice system** has not fallen into the wrong hands, ministers claimed last night” *[how could they know? unless....]*

Telegraph (25 May 2009): “The personal medical records of tens of thousands of people were lost by the **NHS**, the Department of Health has confirmed”

Telegraph (2 July 2009): “**HSBC Life** lost an unencrypted CD with details of 180,000 policy holders, while HSBC Actuaries lost a disc with data on almost 2000 pension scheme members.”

From *The Times*, Friday 26th November 2010

Data loss continues to harm the health of the nation

Carol Lewis

More than 15 million people worldwide have had personal data lost or stolen this year. In more than half the cases the data was lost from health, government or education organisations.

Overall there has been a decrease in the number of data-loss incidents in the past three years but in the past year the amount of data lost from healthcare organisations, including British health trusts, has risen by 12 per cent.

The figures, published today in the report *Data Loss Barometer* by the professional services company KPMG shows that while financial services and retail companies are tightening security, public sector organisations are lagging behind.

Malcolm Marshall, the author of the

report, said: "The number of data-loss incidents is declining. Companies do appear to have put in more and better controls but that doesn't seem to be the case in the healthcare sector."

He continued: "There are several problems in healthcare. One is to do with the cultural nature of the business, you look after patients first and paperwork second. Then there is the complexity of the cases, multiple people need to look at medical records. You have to give a lot of people access to the data. Also many NHS organisations are limited in the money they can spend on computers and training staff."

Almost four million people have had data lost from healthcare organisations this year so far and the loss of medical records has doubled since last year — representing a quarter of data-loss

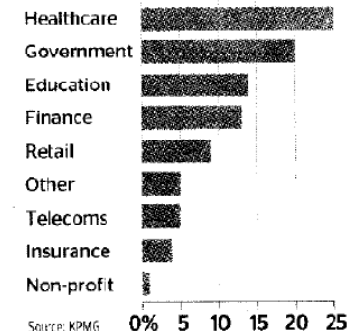
cases. Public sector cuts mean that a lack of resources for encryption and to train staff is likely to continue. The £100,000 fine handed out by the Information Commissioner to Hertfordshire County Council this week for data losses should be a wake up call for public sector organisations.

"The UK privacy regulator has given government organisations in its sights because there are concerns about the level of protection, particularly in the healthcare sector," Mr Marshall said.

Putting all patient details on medical records and holding them on external servers, as has been proposed, would not necessarily make the data less secure. Mr Marshall said: "If you were to ask me whether I'd trust a data protection centre more than a healthcare trust, then the answer is yes. But it also

Data loss by sector

Number of incidents as a percentage of total



Source: KPMG

depends on how secure your passwords are and who is responsible for them."

More than 10 million people worldwide lost national insurance or social security numbers through data losses or thefts in the past year and a further 10 million lost personally identifiable information. About a million people lost bank account details and 1.5 million credit or debt card information.

In healthcare the most common form of data loss was computer or portable media theft, while in government it was "malicious insiders". Although more people globally are affected by hacking than by any other cause of data loss, malicious insiders are a growing threat — cases attributable to them have increased from 4 per cent to 20 per cent in the past three years.

"More than 15 million people worldwide have had personal data lost or stolen this year"

(KPMG's Data Loss Barometer: <http://www.datalossbarometer.com/>)

Data in wrong hands....

Identity theft

"More than £1.6 million worth of card fraud occurs on UK plastic cards every day. A fraudulent transaction takes place every eight seconds"

The Gartner Group

Victims of identity theft spend six months to two years recovering

Up to 18 million households in the UK regularly throw away sensitive financial documents without shredding them

MEL research

Data quality: mistaken conclusions about individuals

Of data errors there is no end:

- mistyped entries
- transposed digits
- misused data
- missing values
- selection bias
- mismatched files
-

e.g.1: the 1970 US Census result showing that 2,926 males aged 25-29 were enrolled in the first grade of school, and that 289 boys had been both widowed and divorced by the age of 14.

Mismatched files

Friends of retired bus driver Frank Hughes were shocked to meet him in the street after attending his funeral

My reversible tonisellectomy

David Hand = David J. Hand = D.J.Hand = D.Hand ?

Conclusion

This view of statistics as '*the science of the individual*'

- enables the *unique*, the *idiosyncratic*, the *special* to be captured
- and *personalises* the discipline of statistics
- but no advanced technology is free of risk, and statistics is no exception

Statistics enables knowledge of the many to be focused down to improve the lot of the one

Some reading

Hand D.J. (2004) *Measurement Theory and Practice: The World Through Quantification*. Arnold (and Wiley, 2009)

Hand D.J. (2007) *Information Generation: How Data Rule Our World*. Oneworld Publications.

Hand D.J. (2009) Modern statistics: the myth and the magic (RSS Presidential Address). *Journal of the Royal Statistical Society, Series A*, **172**, 287-306.