Statistics for data science: what are the essentials?

David Spiegelhalter

Chair of the Winton Centre for Risk & Evidence Communication, University of Cambridge
President, Royal Statistical Society (2017-2018)

ISM Consortium for Human Resource Development of Statistical Experts in Japan. February 2022
I used to do statistical methodology.... until I was philanthropically funded in 2007.....
At the heart of the climate change debate is a paradox - we need more information about our changing climate, yet surveys show the public are, if anything, getting less sure they understand what's what.
The Norm Chronicles
Stories and Numbers about Danger
The Art of Statistics
Learning from Data
David Spiegelhalter

This marvellous book will transform your relationship
with the numbers that swirl all around us

THE ART OF STATISTICS
HOW TO LEARN FROM DATA
DAVID SPIEGELHALTER
Interpreting data is not easy

INTRODUCTION

The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.
— Nate Silver, *The Signal and the Noise*¹
The traditional statistics course

• Describing data with summary statistics
  o dull

• Probability theory for drawing random observation from a population distribution
  o difficult and mathematical

• Probability theory for distributions of summary statistics
  o mathematical and incomprehensible

• Formulae for statistical tests
  o mathematical, unmotivated, just a bag of tools

• (If lucky) Examples of using statistical models in real life.
A ‘modern’ statistical course

• Motivate by problem solving
• Start with visualisation and exploring data
• Focus on what can be reasonably learned from data, biases in data, concluding causation, etc
• Models and algorithms
• Assessing uncertainty through re-sampling data (‘bootstrap’)
• Probability theory as neat way of turning random variation into uncertainty about what is true
• Hypothesis testing and its potential problems
• Bayesian methods
All these rather abstract, challenging, ideas are there to help answer real questions

• The ’data cycle’

• eg PPDAC (promoted in New Zealand)
Are You a Data Detective?

Problem
* understanding and defining the problem
* how do we go about answering this question

Plan
* what to measure & how?
* study design?
* recording?
* collecting?

Data
* collection
* management
* cleaning

Analysis
* sort data
* construct tables, graphs
* look for patterns
* hypothesis generation

Conclusion
* interpretation
* conclusions
* new ideas
* communication

Data detectives use PPDAC
Looking at data
What was the pattern of Harold Shipman’s murders?
‘I have nothing to hide’

Dr Harold Shipman, general practitioner, on his arrest in September 1998
Shipman Inquiry July 2002:
215 definite victims,
45 probable
Looking at data
What was the pattern of Harold Shipman’s murders?

• **Problem**: can more detail tell us more about what Shipman did?

• **Plan**: compare actual times at which his patients died with the times of deaths recorded by other local GPs

• **Data**: a huge exercise requiring examination of death certificates

• **Analysis**: simple plotting.....
People die at all hours
People die at all hours - but not Shipman’s victims.
Inference and bias

How many sexual partners have people in Britain had in their lifetime?

• **Problem**: cannot know this as a fact

• **Plan**: survey in which people are carefully asked about the sexual activity (Natsal)

• **Data**: reports of numbers of partners

• **Analysis**: plotting and summary statistics
How many sexual partners do people report?

![Bar chart showing the reported number of lifetime opposite-sex partners for women and men aged 35-44.](chart.png)
Inference and bias
How many sexual partners have people in Britain really had in their lifetime?

<table>
<thead>
<tr>
<th>Reported number of sexual partners in lifetime</th>
<th>Men aged 35–44</th>
<th>Women aged 35–44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>14.3</td>
<td>8.5</td>
</tr>
<tr>
<td>Median</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Mode</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Range</td>
<td>0 to 500</td>
<td>0 to 550</td>
</tr>
<tr>
<td>Inter-quartile range</td>
<td>4 to 18</td>
<td>3 to 10</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>24.2</td>
<td>19.7</td>
</tr>
</tbody>
</table>

• **Conclusions**: can we generalise this to the whole population??????
**Induction**: the stages in generalising from data

1 to 2. How reliable are the reports?
- Poor memory, social acceptability bias etc

2 to 3. How representative is the sample of those eligible for the study?
- Random sampling of families (soup), 66% response

3 to 4. How close does the study population match the target population?
- No people in institutions, etc
Causation (or correlation)
The power of the press release....
abstract:

- We observed consistent associations between higher socio-economic position and higher risk of glioma

press release

- High levels of education linked to heightened brain tumour risk

Daily Mirror...
Why going to university increases risk of getting a brain tumour

Highly educated people are more likely to suffer from brain tumours than those who do not progress as far in their education.
Regression, prediction and algorithms

Who was the luckiest person on the Titanic?
Ilfracombe, North Devon
Challenge: can we build an algorithm that will accurately predict who survives the Titanic?

Based on factors in data-base, produce either a yes/no judgement, or a probability of survival

Split the data-base of 1309 passengers at random into a **training set** (70%) on which to build algorithms, and a **test set** (30%) to assess how good it is.

Currently over 59,000 entries in a similar online Kaggle competition
Unsurprising factors predict survival.
A simple classification tree

- **Title = Mr?**
  - Yes: Estimated chance of survival 16%
  - No: 3rd Class?
    - Yes: Rare title?
      - Yes: Estimated chance of survival 93%
      - No: Estimated chance of survival 37%
    - No: At least 5 in family?
      - Yes: Estimated chance of survival 60%
      - No: Estimated chance of survival 3%
How good is my algorithm?

• ‘Accuracy’ is a very crude way of judging an algorithmic prediction

• Better to use the probabilities provided

• If probability $p$ is given to an event $X$ (0,1), then the Brier score is $(X - p)^2$
## Performance of a range of methods on the test set

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (high is good)</th>
<th>Brier score (low is good)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone has a 39% chance of surviving</td>
<td>0.639</td>
<td>0.232</td>
</tr>
<tr>
<td>All females survive, all males do not</td>
<td>0.786</td>
<td>0.214</td>
</tr>
<tr>
<td>Simple classification tree</td>
<td>0.806</td>
<td><strong>0.139</strong></td>
</tr>
<tr>
<td>Classification tree (over-fitted)</td>
<td>0.806</td>
<td>0.150</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.789</td>
<td>0.146</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.799</td>
<td>0.148</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td>0.782</td>
<td>0.153</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.794</td>
<td>0.146</td>
</tr>
<tr>
<td>Averaged neural network</td>
<td>0.794</td>
<td>0.142</td>
</tr>
<tr>
<td>K-nearest-neighbour</td>
<td>0.774</td>
<td>0.180</td>
</tr>
</tbody>
</table>
• Potentially a very misleading graphic!
• When comparing, need to acknowledge that tested on same cases
• Calculate differences and their standard error

• How confident can we be that simple CART is best algorithm?
Ranking of algorithms

• Bootstrap sample from test set (ie sample of same size, drawn with replacement)
• Rank algorithms by performance on the bootstrap sample
• Repeat ‘000s of times

• (ranks actual *algorithm* – if want to rank *methods*, need to bootstrap training data too, and reconstruct algorithm each time)
Distribution of true rank of each algorithm

Probability of ‘best’:

63% simpleCART
23% ANN
8% randomforest
Who was the luckiest person on the Titanic?

- Karl Dahl, a 45-year-old Norwegian/Australian joiner travelling on his own in third class, paid the same fare as Francis Somerton
- Had the lowest average Brier score among survivors – a very surprising survivor
- He apparently dived into the freezing water and clambered into Lifeboat 15, in spite of some on the lifeboat trying to push him back.
- Hannah Somerton was left just £5, less than Francis spent on his ticket.
Hypothesis testing

Could Harold Shipman have been caught earlier?

• Using mortality rates from local GPs, calculate how many deaths he would have been expected to observe each year, under the null hypothesis that his mortality rates were normal.

• Subtract expected from observed number to get excess mortality
(NB: Shipman Inquiry total of definite or probable victims: 189 female > 65, 55 male over 65)
Hypothesis testing
Could Harold Shipman have been caught earlier?

• But when to ‘blow the whistle’?
• This are two possible types of error -
  • **Type I error**: falsely accuse an innocent person
  • **Type II error**: miss someone with true increased risk
• This is an example of a *hypothesis test* used throughout science. Again, two possible types of error
  • **Type I error**: falsely claim an ‘effect’ when nothing is there (ie the *null hypothesis* is true)
  • **Type II error**: miss a true effect
• Generally, we want to
  • control the probability of a Type I error at a low value (α)
  • make experiments large enough to make Type II errors rare (β)
Shipman: “Sequential probability ratio test” (SPRT) older females would have set off ‘alarm’ in 1985, after only 40 deaths
Probability and Bayes
• 1000 experiments
• Assume that in 10%, there is really an effect
• Size ($\alpha$) = 5%
• Power (1-$\beta$) = 80%
• ‘reverse the tree’

• If reject the null, only $\frac{80}{125} = 64\%$ chance that null is False
Bayes theorem

- Initial odds that null hypothesis False = 10 / 90
- After ‘significant’ results, final odds that null hypothesis False = 80 / 45
- Likelihood ratio = \[ \frac{\Pr(\text{significant result} \mid \text{null hypothesis False})}{\Pr(\text{significant result} \mid \text{null hypothesis True})} \]
- \[ = \frac{\text{Power}}{\text{Size}} = \frac{1-\beta}{\alpha} = \frac{0.80}{0.05} = 16 \]
- Bayes theorem:
  the initial odds for the hypothesis \( \times \) the likelihood ratio
  = the final odds for a hypothesis.
  \[ \frac{10}{90} \times \frac{80}{5} = \frac{80}{45} \]
What is the probability that the skeleton in a Leicester car park was really Richard III?
A recent case

• On Saturday 25 August 2012, archeologists started digging in a car park in Leicester – the site of Grey Friars friary

• In a few hours they found their first skeleton

• This was later claimed to be Richard III
Likelihood ratio = \[
\frac{\text{probability of evidence, if skeleton were Richard III}}{\text{probability of evidence, if someone else}}
\]
Suggested ‘verbal equivalents’ for bands of likelihood ratios

<table>
<thead>
<tr>
<th>Value of likelihood ratio</th>
<th>Verbal equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1–10</td>
<td>Weak support for proposition</td>
</tr>
<tr>
<td>10–100</td>
<td>Moderate support</td>
</tr>
<tr>
<td>100–1000</td>
<td>Moderately strong support</td>
</tr>
<tr>
<td>1000–10,000</td>
<td>Strong support</td>
</tr>
<tr>
<td>10,000–1,000,000</td>
<td>Very strong</td>
</tr>
<tr>
<td>&gt;1,000,000</td>
<td>Extremely strong</td>
</tr>
</tbody>
</table>

Standards for the formulation of evaluative forensic science expert opinion

Association of Forensic Science Providers
<table>
<thead>
<tr>
<th>Evidence</th>
<th>Likelihood ratio (conservative estimate)</th>
<th>Verbal equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiocarbon dating AD 1456–1530</td>
<td>2</td>
<td>Weak support</td>
</tr>
<tr>
<td>Age and sex of skeleton</td>
<td>5</td>
<td>Weak support</td>
</tr>
<tr>
<td>Scoliosis</td>
<td>212</td>
<td>Moderately strong support</td>
</tr>
<tr>
<td>Post-mortem wounds</td>
<td>42</td>
<td>Moderate support</td>
</tr>
<tr>
<td>mtDNA match</td>
<td>478</td>
<td>Moderately strong support</td>
</tr>
<tr>
<td>Y chromosome not matching</td>
<td>0.2</td>
<td>Weak evidence against</td>
</tr>
<tr>
<td>Combined evidence</td>
<td>6.5 million</td>
<td>More than extremely strong support</td>
</tr>
</tbody>
</table>
Conclusions – statistics for data science

• Motivate by problem solving
• Start with visualisation and exploring data
• Focus on concepts: what can be reasonably learned from data, biases, causation, etc
• Models and algorithms
• Probability can come much later
• Conditional probability / Bayes theorem taught through ‘expected frequency trees/