Multivariate analysis and data mining
-
statistics in the computer age

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

Part 1: Background

Definitions and history

Statistics, machine learning, pattern recognition, data mining, ...

All intersecting disciplines

- common areas of interest
- differences in emphasis
  - understanding vs prediction
  - large data sets
  - sequential methods
  - data quality, data collection

Greater statistics:

"*everything related to learning from data, from the first planning or collection to the last presentation or report*"

(John Chambers)

Not static

"*The first version of [Cooley and Lohnes, 1962] demonstrates how the usual [Multivariate Analysis] techniques must be implemented on a computer. This means, of course, that it is now completely out of date.*"

Gifi (p5):

Revolution driven by the computer

- Large data sets
  - particle physics, bioinformatics, retail
- Complex models
- Vast model space
- Replace careful thought by computer power

Models versus algorithms

Modern stats (mathematics heritage): ***models***

$$f\left(y \mid x_1, \ldots, x_p\right) = \beta_0 + \sum_i \beta_i x_i + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

Data mining (computer heritage): ***algorithms***

A ***recipe*** for calculating *y* from ***x*** to *y*

A mapping is both a ***structure*** and a ***process***

Recursive partioning in supervised classification

Breiman *et al* (1984):
Classification and Regression ***Trees***

Quinlan (1993)
Programs *for Machine Learning*

To summarise, describe, and predict, no distributional assumptions are necessary

But for inference - answering the question '*is this how the underlying truth is?*' - we need to make distributional assumptions

e.g.

a regression line is simply the *best fitting* line in the sense that it minimises the residual sum of squares - no distributional assumptions

but to answer the question 'could the population regression slope really be zero' I need to know something about distributions.

Classical multivariate statistics

Albert Gifi (1990) *Nonlinear Multivariate Analysis*, Ch.1 contained a *content analysis* of earlier MVA books:

Roy (1957), Kendall (1957), Kendall (1975), Anderson (1958), Cooley and Lohnes (1962), Cooley and Lohnes (1971), Morrison (1967), Morrison (1976), Van de Geer (1967), Van de Geer (1971), Dempster (1969), Tatsuoka (1971), Harris (1975), Dagnelie (1975), Green and Carroll (1976), Cailliez and Pages(1976), Giri (1977), Gnadadesikan (1977), Kshirsagar (1978), Thorndike (1978)

Many books on these topics have appeared since then, taking advantage of modern computer power to considerably enhance the tools. On my bookshelves alone I have

Mardia, Kent, and Bibby (1979), Tabachnik and Fidell (1983), Flury and Riedwyl (1988), Krzanowski (1990), Krzanowski and Marriott (1995), Rencher (1995), Everitt and Dunn (2001), Bartholomew, Steele, Moustaki, and Galbraith (2008)

The Gifi multivariate analysis of the content of multivariate analysis books described each book in terms of the number of pages devoted to 7 topics:

- Mathematics other than statistics. i.e. linear algebra, matrices, transformation groups, sets, relations
- Correlation and regression, including path analysis, linear structural and functional equations
- Factor analysis and principal components analysis
- Canonical correlations analysis
- Discriminant analysis, classification, cluster analysis
- Statistics, including distributional theory, hypothesis testing, and estimation; and analysis of categorical data
- MANOVA, and the general multivariate linear model

Inference largely based on the *multivariate normal distribution*

"following on the brilliant work of R. A. Fisher who showed that, when universal normality could be assumed, inferences of the widest practical usefulness could be drawn from samples of any size. Prejudice in favour of normality returned in full force. . . and the importance of the underlying assumptions was almost forgotten.

(Geary, 1947, p. 241)"

Modern attitude is more skeptical about normality

Less need to assume normality now because less need for mathematical tractability

Standard classical multivariate techniques

Data matrix $\mathbf{X}_{n\times p}$ with mean centred columns

SINGULAR VALUE DECOMPOSITION

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\mathbf{D}_{r\times r}\mathbf{A}^{T}_{r\times p} \qquad \mathbf{U},\mathbf{A} \text{ orthonormal cols}$$

$$x_{ij} = \sum_{k=1}^{r} d_k\, u_{ik} a_{jk}$$

Sequence $x_{ij} = \sum_{k=1}^{s} d_k\, u_{ik} a_{jk}, \quad s = 1,\ldots,r$ minimises the Euclidean norm of the approximation to **X**

PRINCIPAL COMPONENTS ANALYSIS

Linear combination $\mathbf{a}^T\mathbf{x}$ has variance $\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$

Choose that which maximises the variance (s.t. $\mathbf{a}^T\mathbf{a}=1$)

Leads to $(\boldsymbol{\Sigma}-\lambda\mathbf{I})\mathbf{a}=0$

$$\boldsymbol{\Sigma}=\mathbf{X}^T\mathbf{X}=(\mathbf{A}\mathbf{D}\mathbf{U}^T)(\mathbf{U}\mathbf{D}\mathbf{A}^T)=\mathbf{A}\mathbf{D}\mathbf{D}\mathbf{A}^T=\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T$$

$$\boldsymbol{\Sigma}=\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T=\sum_{i=1}^{p}\lambda_i\mathbf{a}_i\mathbf{a}_i^T \qquad \mathbf{a}_i^T\mathbf{a}_j=\delta_{ij} \qquad \mathbf{z}=\mathbf{A}^T\mathbf{x}$$

(PCs are invariant under orthogonal transformations, but not other transformations: eigenvalues/vectors of covariance and correlation matrix are not simply related; Pearson, 1901; Hotelling, 1933)

FACTOR ANALYSIS

$$\mathbf{x} = \mathbf{W}\mathbf{f} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$$

Contrast with PCA:

PCA $\quad \boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T = \left(\mathbf{A}\sqrt{\boldsymbol{\Lambda}}\right)\left(\mathbf{A}\sqrt{\boldsymbol{\Lambda}}\right)^T$

FA $\quad \boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$

PCA describes variance (diagonal elements of $\boldsymbol{\Sigma}$)
FA describes covariance (off-diagonal elements of $\boldsymbol{\Sigma}$)

PCA is a *transformation*
FA is a *model*

Suppose *p* independent variables, all but one with same variance, and other has a much larger variance

PCA: one PC FA: no factors

- PCs are exact linear combinations of **x**
- **x** linear function of **f** + error, inversion doesn't lead to an exact relation between **f** and **x**

- increasing # of PCs doesn't affect earlier PCs
- increasing # of factors can completely change others

PCA often a good starting point for iterative FA model-fitting (but not always)

('*Principal Factor Analysis*': PCA on $\boldsymbol{\Sigma} - \boldsymbol{\Psi} \left(= \mathbf{W}\mathbf{W}^T \right)$)

LINEAR DISCRIMINANT ANALYSIS

$$\max_{\mathbf{a}} \left\{ \mathbf{a}^T \mathbf{B} \mathbf{a} / \mathbf{a}^T \mathbf{W} \mathbf{a} \right\} \quad \Rightarrow$$

$$(\mathbf{B} - d\mathbf{W})\mathbf{a} = (\mathbf{W}^{-1}\mathbf{B} - d\mathbf{I})\mathbf{a} = 0$$

(c.f. PCA $(\boldsymbol{\Sigma} - \lambda \mathbf{I})\mathbf{a} = 0$)

Scale invariant, in contrast to PCA

Note:

- optimal for MVN distributions (with common CV matrix)
- optimal for elliptical distributions (with common CV matrix)
- but distributional assumptions are not necessary (Fisher's original derivation)

CANONICAL CORRELATIONS ANALYSIS

Given two vector random variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, find the linear combinations $\mathbf{a}_1^T\mathbf{x}_1$ and $\mathbf{a}_2^T\mathbf{x}_2$, which have maximum correlation:

$$C(\mathbf{x}_1,\mathbf{x}_2)=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$$

➔ $\rho\left(\mathbf{a}_1^T\mathbf{x}_1,\mathbf{a}_2^T\mathbf{x}_2\right)=\mathbf{a}_1^T\boldsymbol{\Sigma}_{12}\mathbf{a}_2\Big/\sqrt{\mathbf{a}_1^T\boldsymbol{\Sigma}_{11}\mathbf{a}_1\mathbf{a}_2^T\boldsymbol{\Sigma}_{22}\mathbf{a}_2}$

➔ $\max_{\mathbf{a}_1,\mathbf{a}_2}\left(\mathbf{a}_1^T\boldsymbol{\Sigma}_{12}\mathbf{a}_2-\lambda\mathbf{a}_1^T\boldsymbol{\Sigma}_{11}\mathbf{a}_1-\mu\mathbf{a}_2^T\boldsymbol{\Sigma}_{22}\mathbf{a}_2\right)$

➔ $\left(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}-\lambda\mu\mathbf{I}\right)\mathbf{a}=0$

MANY OTHER TOOLS

- structural equation models, path analysis

$$\mathbf{y}_{q\times 1} = \mathbf{B}_{q\times q}\mathbf{y}_{q\times 1} + \mathbf{\Gamma}_{q\times p}\mathbf{z}_{p\times 1} + \mathbf{u}_{q\times 1}$$

(precursor to graphical models and Bayesian belief networks)

- repeated measures analysis, structured CV matrices
- projection pursuit
- cluster analysis --- and segmentation analysis
- independent components analysis
- partial least squares, $\max Corr^2(y, \mathbf{Xa})Var(\mathbf{Xa})$
- manova
- nonlinear MVA, Gifi
- . . . . .

Part 2: General issues

Linear → nonlinear: complicated estimation (e.g. ANNs)

Search over vast model space (e.g. trees, basis functions,...)

Parametric → nonparametric (e.g. kNN, kernel, tree, ...)

Generalise: fit *future* data not *current* data (bias/variance tradeoff)

Curse of dimensionality

Regularisation, smoothing, shrinking

Ensemble methods: bagging, boosting, model average

The bootstrap principle

Part 3: Data Mining

Data mining:

"*the technology of extracting interesting, unexpected, or valuable structures from large data sets*"

Electronic data capture
+ huge data stores
+ fast data processing

➔ massive searches
through data space
and model space

‘Data mining’ principles not new:

data ***dredging***, ***trawling***, ***fishing*** through data

But the ***modern discipline is new***:

Statistics + ideas, tools, and methods from ***computer science, machine learning, database technology***, etc.

Classical statistical perspective: ‘***with a large enough data set one is bound to find structure***’

Others’ perspectives: ‘***true, but it is also certain that there is valuable information in the data***’

‘*All one has to do*’ is to find it

History:

Originally driven by computer scientists:
“extract interesting configurations from databases”

But this ignores
- issues of data quality
- issues of inference (underlying reality vs chance)

➔ shift towards statistical perspective

Evolution of data mining

Now also ***experimental*** as well as ***observational***

Google, Amazon experiments

Tesco Pointscard

Capital One

Two aspects to data mining

Model building

- characterising large scale features of data sets
  - cluster analysis
  - time series decomposition
  - belief networks

Anomaly detection and discovery

- outlier detection
- detecting sudden changes
- bump hunting

(Coal deposits vs gold nuggets)

Modelling issues in data mining

- data sets too large for standard tools
    - scalability of algorithms is essential
    - favour simple methods over complex
    - automatic analysis is essential

- data won't fit into memory
    - adaptive/sequential algorithms (ML)

'Why don't you just take a sample?'

Sometimes you should

But

- random sampling may be difficult or impossible
- may be impossible to define the sampling frame
    - much data are dynamic
- data may be complex
- storage may be complex
    - no single flat file, distributed (e.g. the web)
- access may be time consuming
- the *questions* may be dynamic:
    - collaborative filtering
    - customer modelling in call centres

Anomaly detection and discovery

Examples

- Detecting anomalous astronomical objects
- Detecting the Higgs boson
- Pharmacovigilance
- Bioinformatics: microarray, genomics, etc
- Disease outbreaks
- Customer relationship management
- Credit card fraud detection
- ...

The top ten algorithms in data mining (Wu and Kumar)

C4.5: supervised classification
Adaboost: supervised classification
k-NN: supervised classification
Naive Bayes: supervised classification
CART: supervised classification
SVMs: chiefly supervised classification

k-means: unsupervised classification, clustering
A priori: unsupervised pattern detection

EM: model fitting
PageRank: ranking webpages relative to query

Part 4: Data quality

- central but underestimated, aspect of data mining

Data are never perfect

Poor data

$\Rightarrow$ mistaken conclusions

$\Rightarrow$ wrong decisions

$\Rightarrow$ incorrect actions

Data can be right in only one way

(a faithful record of the underlying reality)

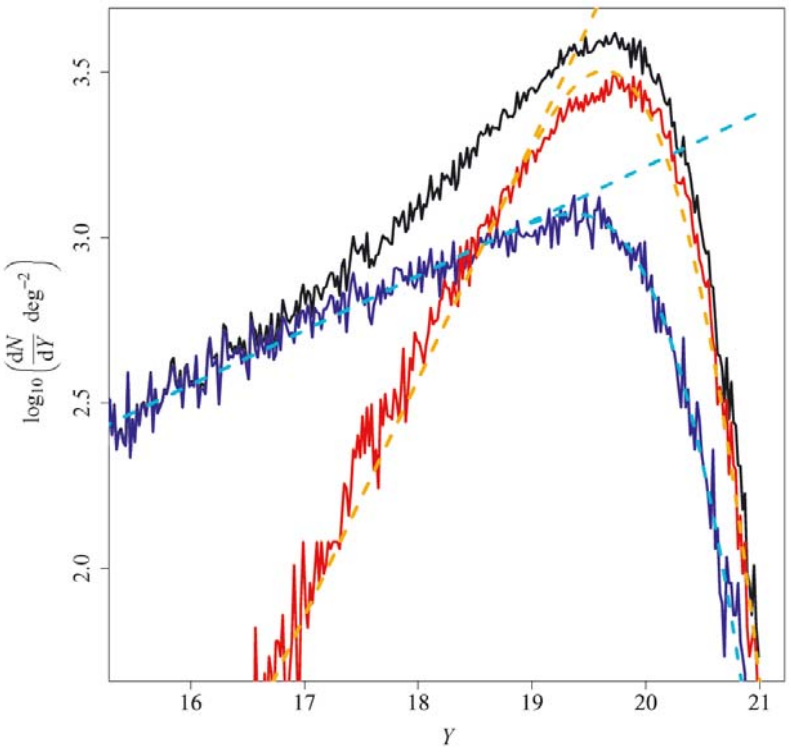
but

Data can be wrong in an infinity of ways

Data quality is not an absolute:

Data may be good for one purpose, but not for another

Data quality 1: Selection bias



(Henrion, 2011)

The anthropic principle
Millikan oil drop experiment
....

Example 1: Selection bias in screening customers for bank loans

Training set:

Existing customers, with known characteristics, and known ‘good/bad’ outcomes

But ‘existing customers’ are those we previously thought were likely to be good

They are not a random sample from the population of potential applicants

Extreme illustration:

Binary feature X: highly predictive

X = 1 ➔ will certainly default

X = 0 ➔ will certainly not default

All other predictors, Y, are poor

So we previously rejected all those applicants with X=1

Our training set contains *none* with X=1

So when we build a new classifier, variable X is not identified as a good predictor

and we are left only with the poor predictors Y to use in our new scorecard

Example 2: Selection bias in anomaly detection - banking fraud

True transaction state sequence

nnnnnnnnnnnnfnfnfffnfff

Detector D1 in place
Detector D2 proposed new detector

$D_i,\ i = 1, 2$ taking values 0 (no fraud suspected)
1 (fraud suspected)

$D_1 = 1$ and true state *n* means:

investigation and then sequence continues

$D_1 = 1$ and true state *f* means

investigation and then sequence ends

AND

true states of all previous transactions discovered

Define, for $j,k = 0,1$

$$p_{jk}^{(n)} = P\left(D_1 = j, D_2 = k \,|\, n\right)$$

and

$$p_{jk}^{(f)} = P\left(D_1 = j, D_2 = k \,|\, f\right)$$

Then the new detector, D2, is unequivocally better than the old one, D1, if both

(i) $P\left(D_2 = 1 | \, f\right) > P\left(D_1 = 1 | \, f\right)$

and

(ii) $P\left(D_2 = 1 | \, n\right) \leq P\left(D_1 = 1 | \, n\right)$

These are equivalent to

(i) $$p_{01}^{(f)} > p_{10}^{(f)}$$

and

(ii) $$p_{01}^{(n)} \leq p_{10}^{(n)}$$

		D2	
		0	1
D1	0	p_{00}	p_{01}
	1	p_{10}	p_{11}

Consider straightforward estimates of the $p_{jk}^{(f)}$ and $p_{jk}^{(n)}$ based on proportions of observations in

n		D2	
		0	1
D1	0		
	1		

f		D2	
		0	1
D1	0		
	1		

BUT: All observed sequences in which a fraud is detected end in either the $\left(D_1 = 1, D_2 = 0\right)$ ***cell or the*** $\left(D_1 = 1, D_2 = 1\right)$ ***cell of the f table***

Consider a single terminating account with *c* fraudulent transactions

The (*c*-1) undetected frauds contribute only to f_{00} or f_{01}

Hence $E(f_{0k} \mid c) = (c-1)\, p_{0k}^{(f)} / p_{0+}^{(f)}$

The one final detected fraud contributes to f_{1k}

Hence $E(f_{1k} \mid c) = p_{1k}^{(f)} / p_{1+}^{(f)}$

So the expectations of simple multinomial estimates are

$$E(\tilde{p}_{0k}^{(f)} \mid c) = (1-1/c)\, p_{0k}^{(f)} / p_{0+}^{(f)}$$
$$E(\tilde{p}_{1k}^{(f)} \mid c) = (1/c)\, p_{1k}^{(f)} / p_{1+}^{(f)}$$

e.g. suppose c = 1

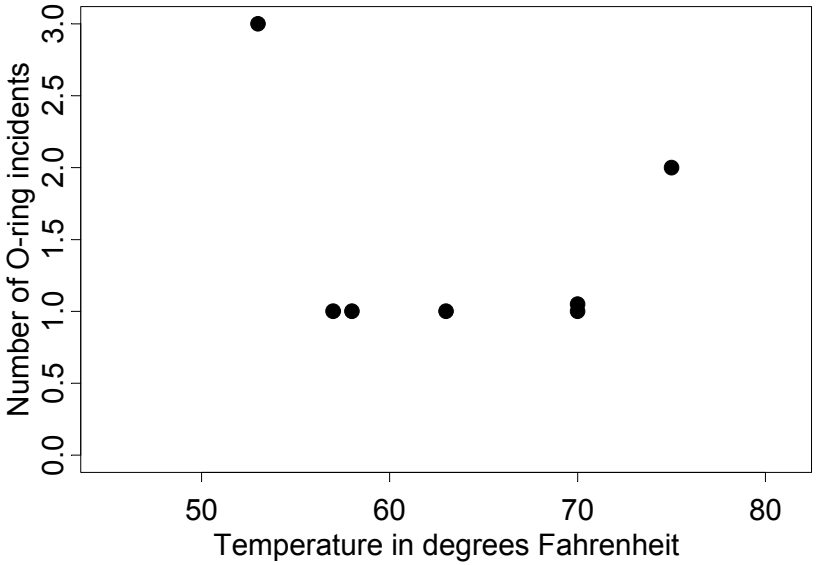
Then $\tilde{p}_{0k}^{(f)} = 0$

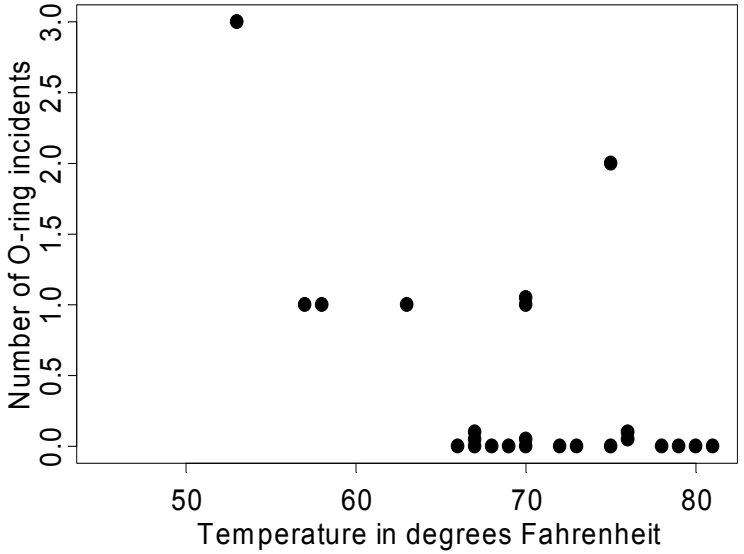
$$E\left(\tilde{p}_{1k}^{(f)} \,|\, 1\right) = p_{1k}^{(f)} \Big/ p_{1+}^{(f)} \geq p_{1k}^{(f)}$$

So the condition for D2 beating D1, that $p_{01}^{(f)} > p_{10}^{(f)}$ *cannot be met by these estimators*

If you use the simple multinomial estimators there is an intrinsic built-in bias favouring the existing detector

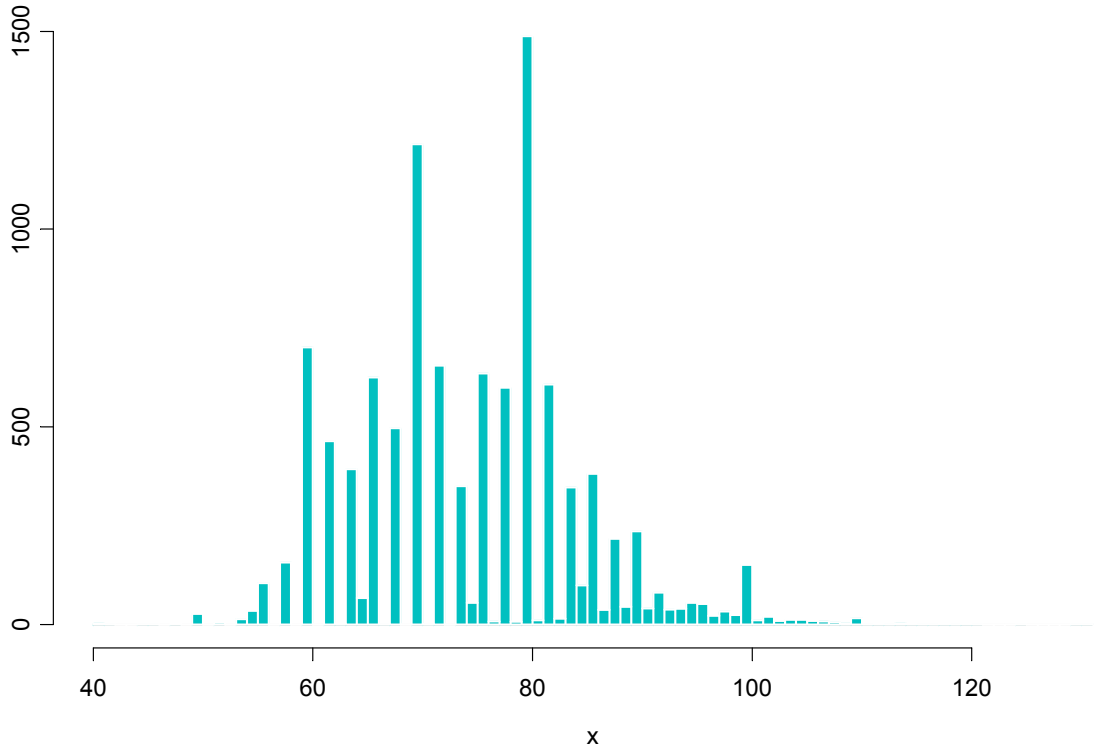
Example 3: Selection bias: a classic example



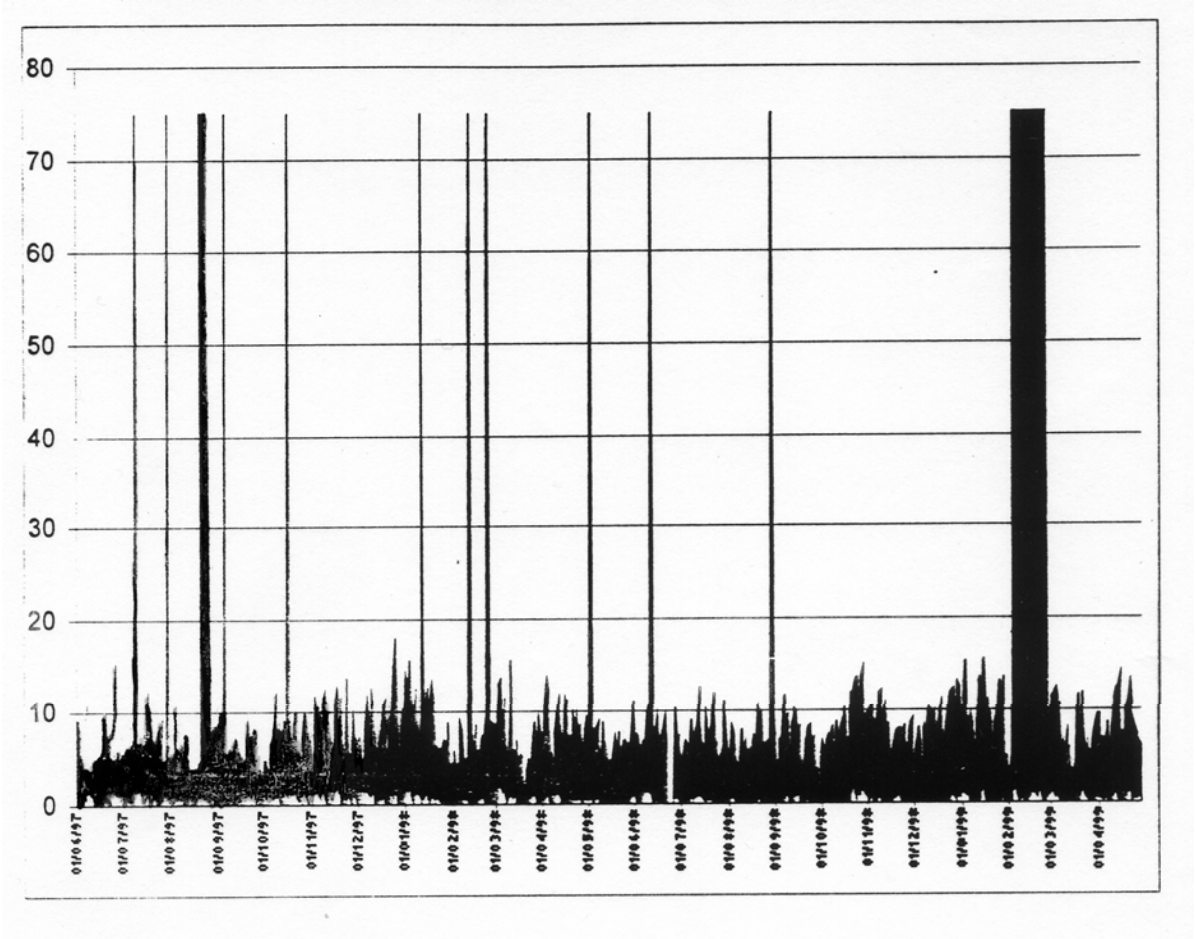


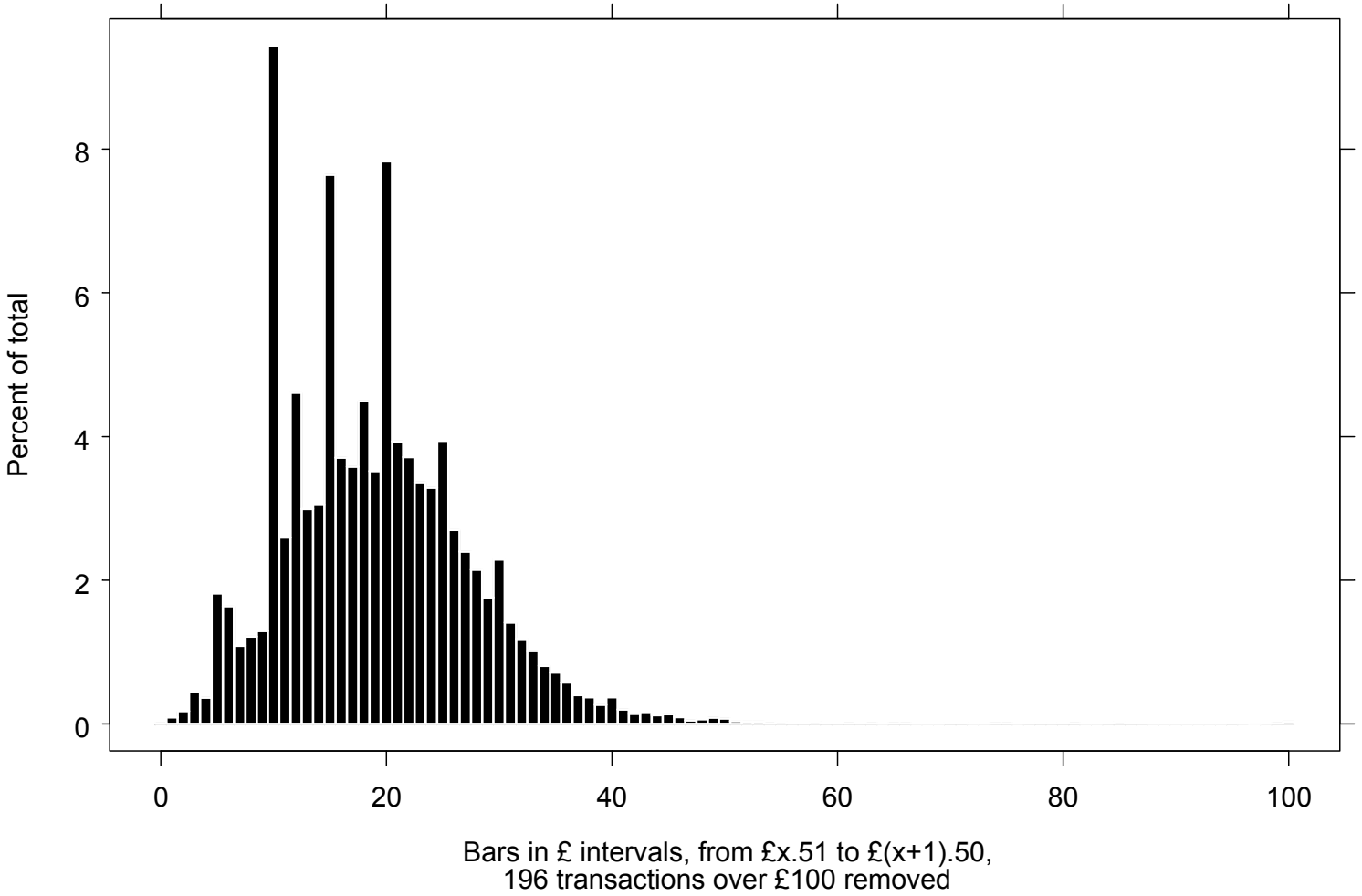
Data quality 2: digit preference

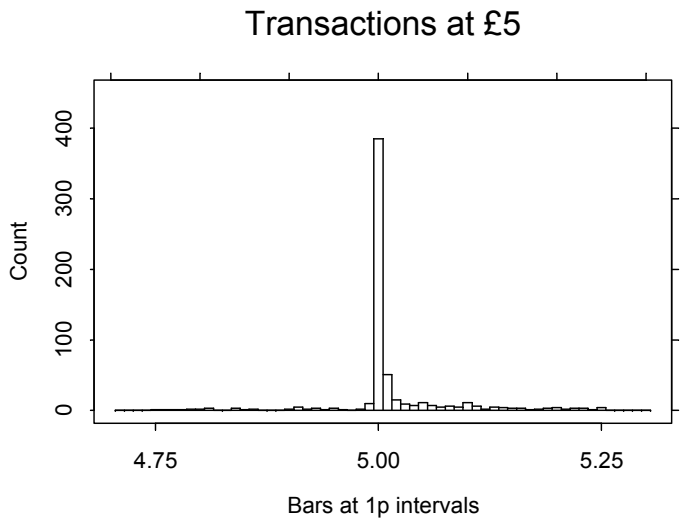
Diastolic blood pressure of 10,000 men

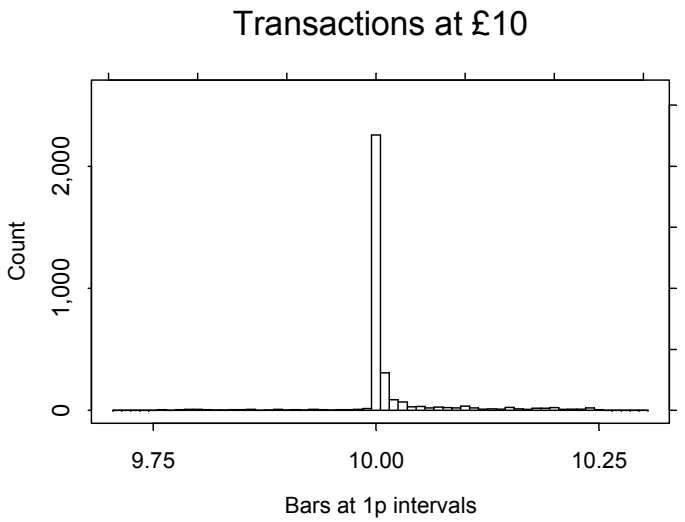


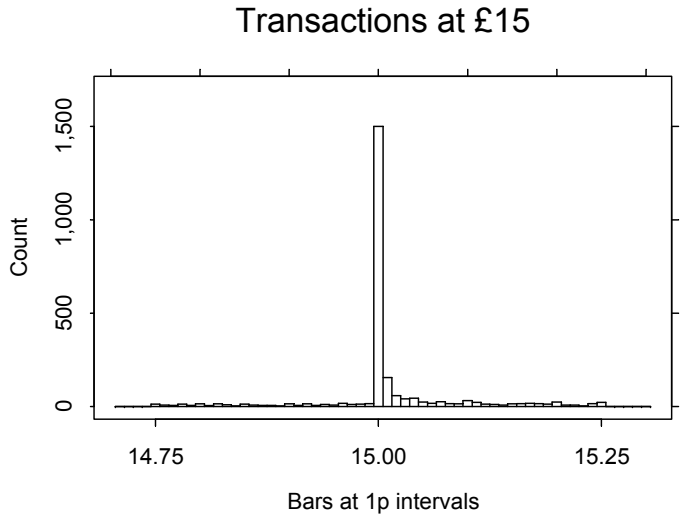
Wind speed

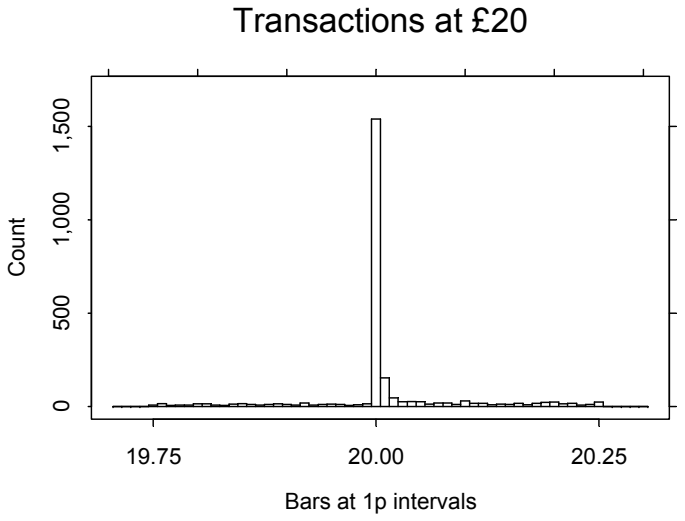


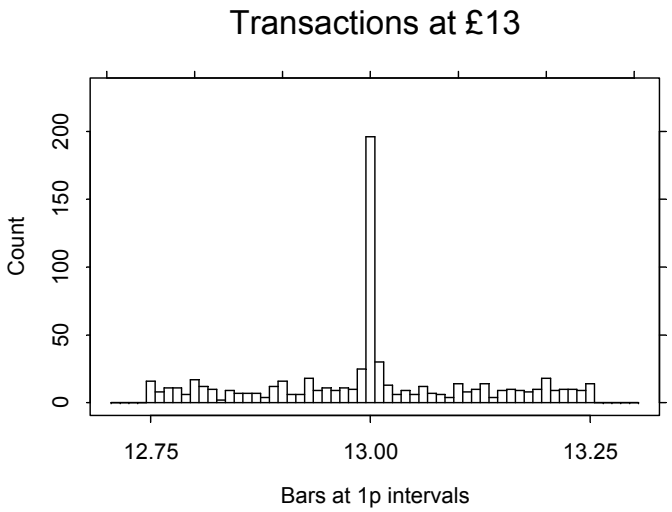

Percent of total
0
2
4
6
8
0
20
40
60
80
100
Bars in £ intervals, from £x.51 to £(x+1).50,
196 transactions over £100 removed

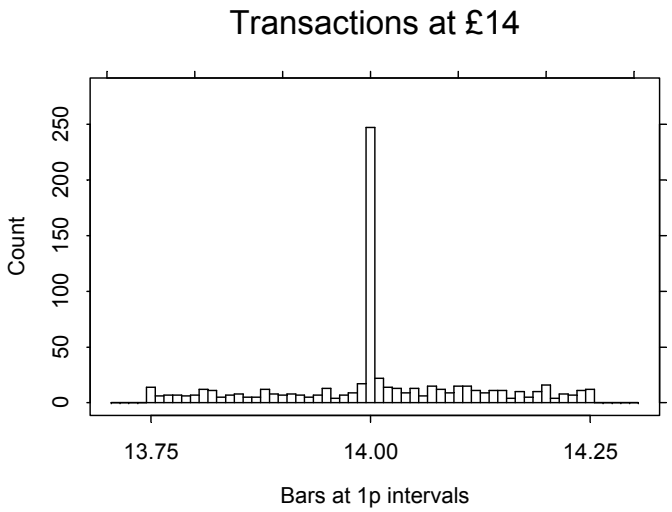

Transactions at £5
Count
0
100
200
300
400
4.75
5.00
5.25
Bars at 1p intervals

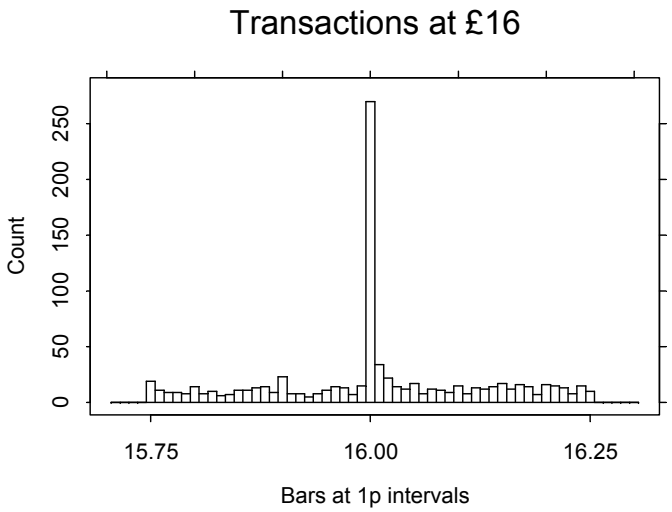

Transactions at £10
Count
0
1,000
2,000
9.75
10.00
10.25
Bars at 1p intervals

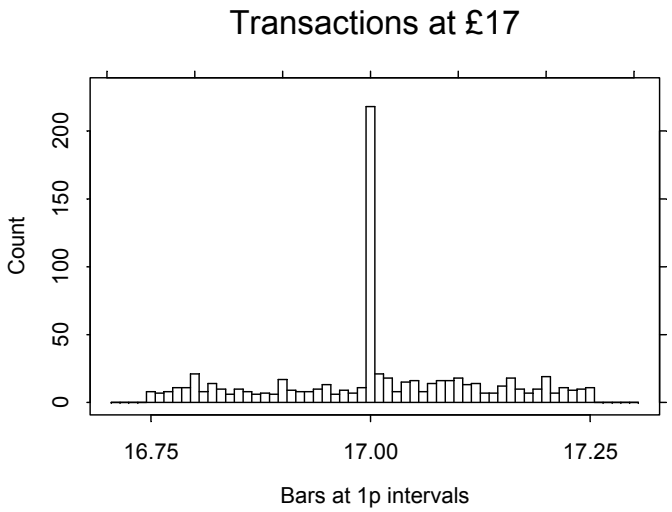

Transactions at £15
Count
0
500
1,000
1,500
14.75
15.00
15.25
Bars at 1p intervals

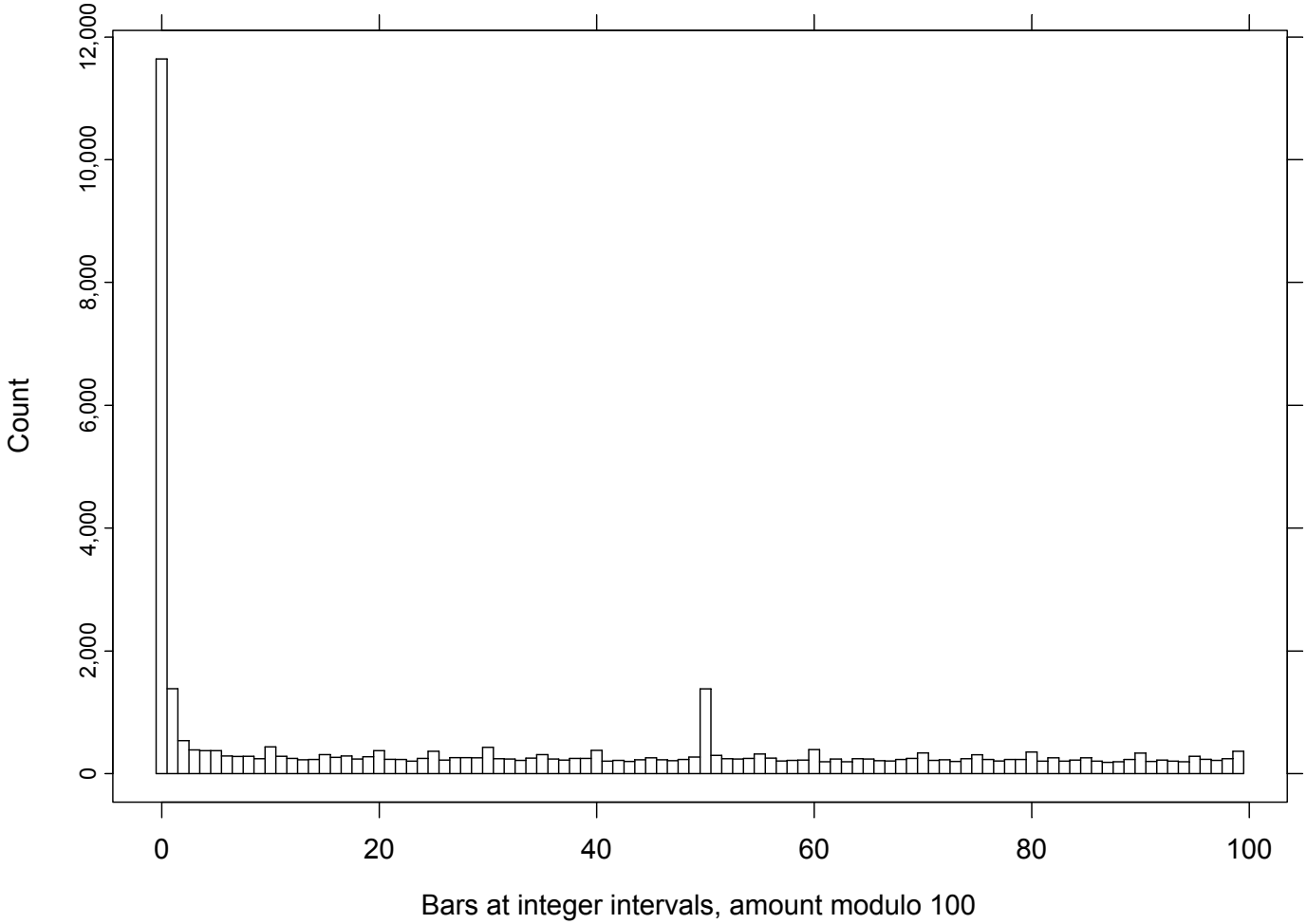

Transactions at £20
Count
0
500
1,000
1,500
19.75
20.00
20.25
Bars at 1p intervals

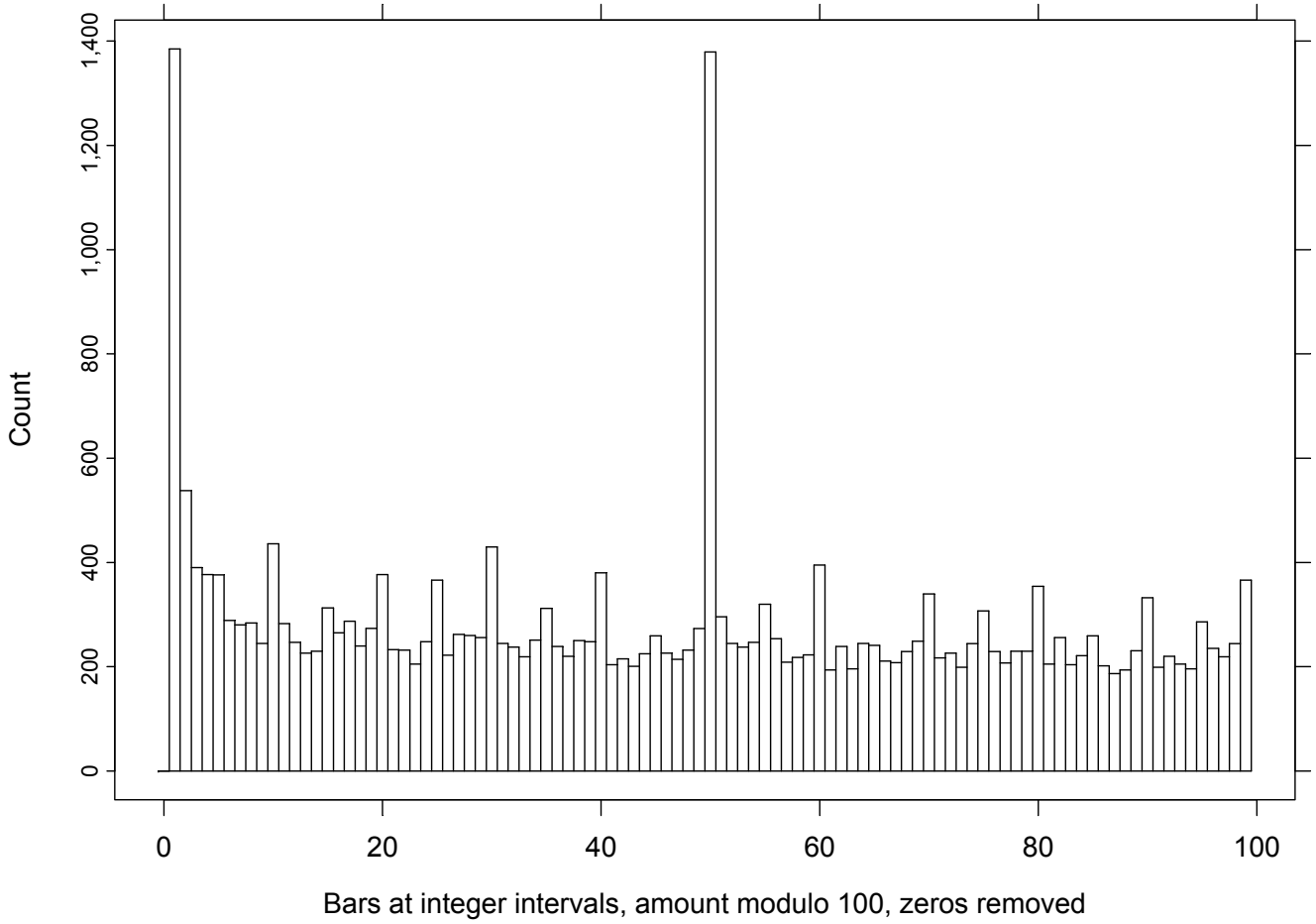

Transactions at £13
Count
0
50
100
150
200
12.75
13.00
13.25
Bars at 1p intervals


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

Part 5: The future

We are not at the *peak* of data analysis, but merely on the *slopes*

New problems present new challenges, meaning new theory and methods will be needed

- e.g. large *p* small *n* problems
- e.g. false discovery rate in anomaly detection

Some reading material

- Krzanowski WJ and Marriott FHC (1994) *Multivariate analysis (2 vols)*. Edward Arnold.
- Hand, Mannila, and Smyth (2001) *Principles of data mining*. MIT Press
- Hastie, Tibshirani, and Friedman (2009, 2nd ed.) *The elements of statistical learning*. Springer
- Webb (2002) *Statistical pattern recognition*. Wiley
- Wu and Kumar (2009) *The top ten algorithms in data mining*. CRC Press

thank you !