Outliers in long–tailed discrete data

Mario Cortina Borja

MRC Centre of Epidemiology for Child Health
University College London, Institute of Child Health

m.cortina@ucl.ac.uk

LSHTM
16th November 2012

Outline

1. Defining outliers
2. Outlier detection methods
3. Long–tailed discrete distributions
4. Long–tailed discrete distributions
5. Outliers in long–tailed discrete distributions
   - Surprise Index

Definitions

- *Oxford English Dictionary*: Statistics. An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point
- *Grubbs (1950)*: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
- *Hawkins (1980)*: An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism
- *Bayarri & Morales (2003)*: A common approach consists in assuming that the (possible) outliers are generated by contaminating models different from the one generating the rest of the data

Definitions

- *Oxford English Dictionary*: Statistics. An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point
- *Grubbs (1950)*: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
- *Hawkins (1980)*: An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism
- *Bayarri & Morales (2003)*: A common approach consists in assuming that the (possible) outliers are generated by contaminating models different from the one generating the rest of the data

Definitions

- *Oxford English Dictionary*: Statistics. An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point
- *Grubbs (1950)*: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
- *Hawkins (1980)*: An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism
- *Bayarri & Morales (2003)*: A common approach consists in assuming that the (possible) outliers are generated by contaminating models different from the one generating the rest of the data

Definitions

- *Oxford English Dictionary*: Statistics. An observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point
- *Grubbs (1950)*: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs
- *Hawkins (1980)*: An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism
- *Bayarri & Morales (2003)*: A common approach consists in assuming that the (possible) outliers are generated by contaminating models different from the one generating the rest of the data

Defining outliers

Everybody knows what it is, but nobody can define it

St Augustine of Hippo, 401AD, (on Time)

Outliers

- Detection and handling (outlying, atypical, spurious) observations is an important part of any statistical analysis
- These observations may simply refer to observational noise or a data processing error – **spurious data**
- Or they may be a special part of the population, and should be treated differently to most of the observations – **outlying data**

Outliers

- Detection and handling (outlying, atypical, spurious) observations is an important part of any statistical analysis
- These observations may simply refer to observational noise or a data processing error – **spurious data**
- Or they may be a special part of the population, and should be treated differently to most of the observations – **outlying data**

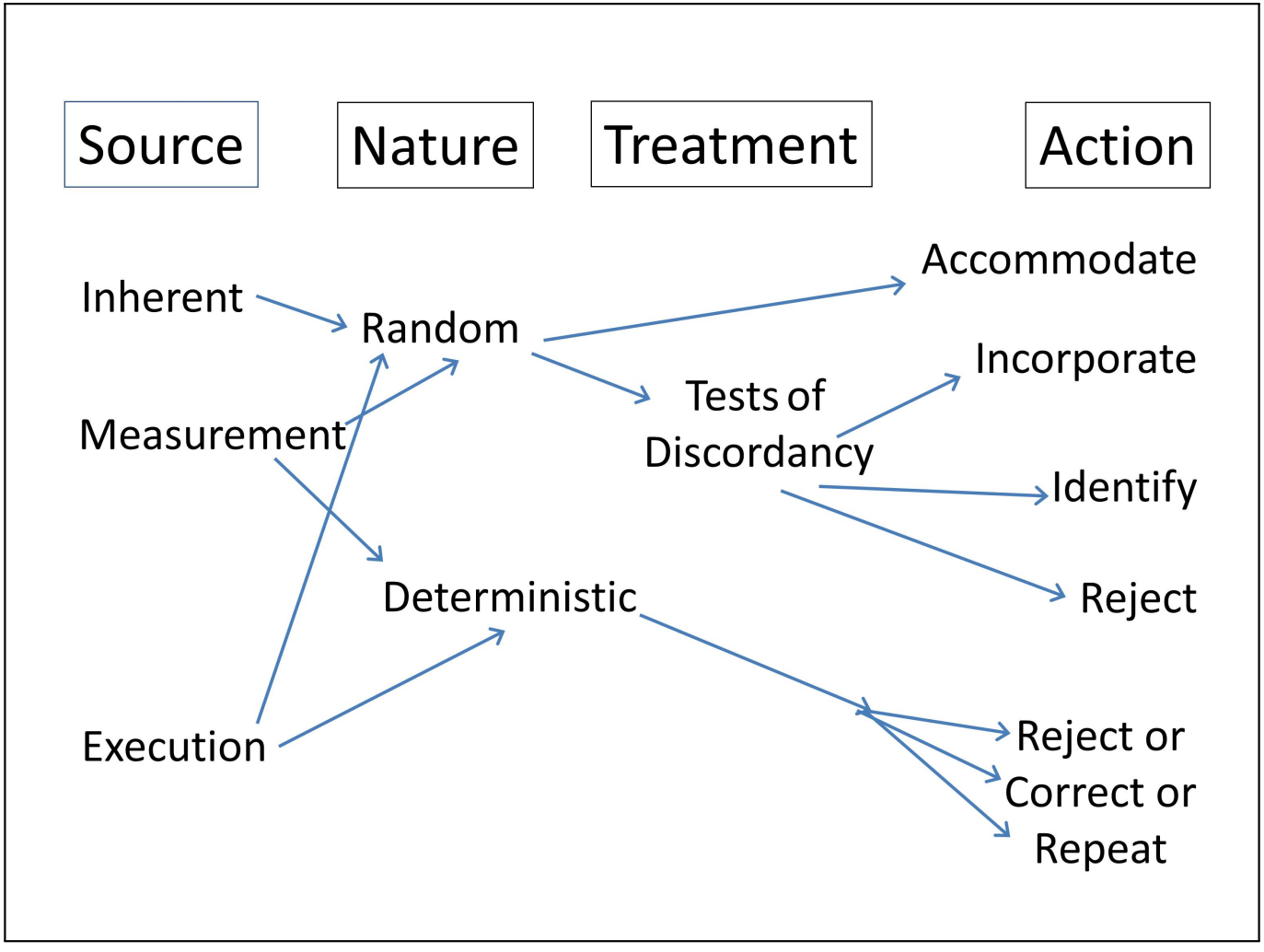
Outliers

- Including outliers in the analyses might lead to model misspecification, biased estimates... and incorrect results
- What do we do with outlying observations?

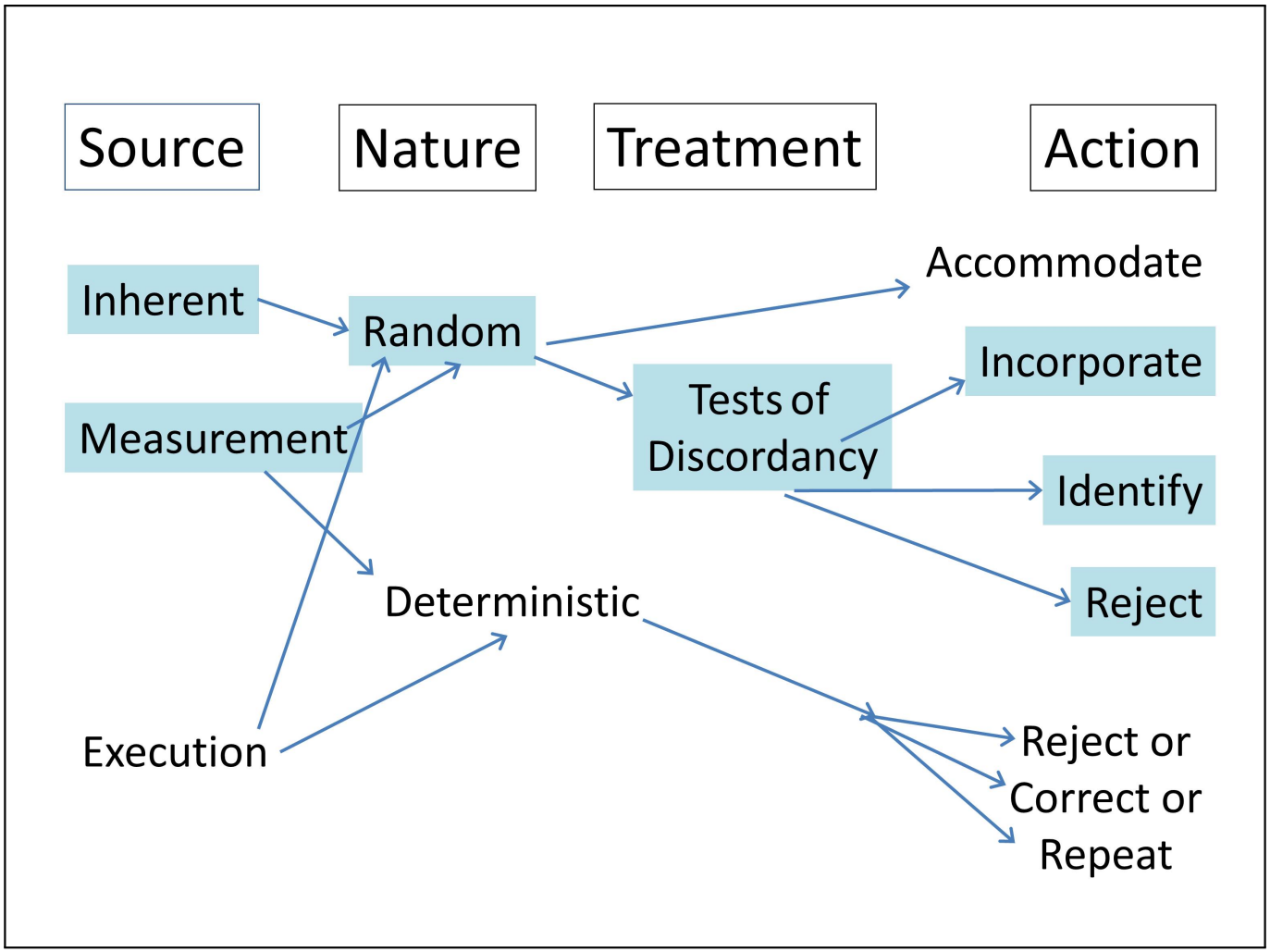
Outliers

- Including outliers in the analyses might lead to model misspecification, biased estimates... and incorrect results
- What do we do with outlying observations?

Treatment of outliers - Barnett & Lewis (1994)

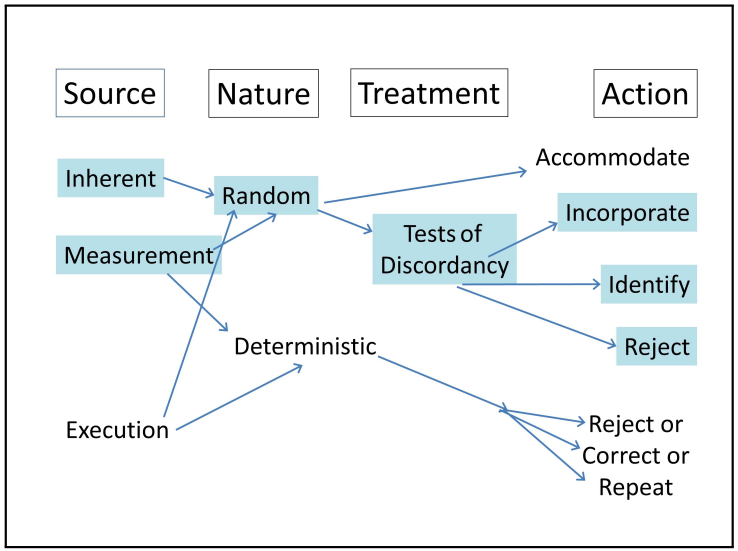


Treatment of outliers - Barnett & Lewis (1994)



Treatment of outliers - Barnett & Lewis (1994)

- Random outliers might be due to inherent variability or, unwittingly, to measurement errors
- Tests of discordancy might be based on assumed initial model
- Outliers might be:
  - Incoporated in revised model
  - Identified for a separate study of origin and form
  - Rejected, if the initial model is inviolable

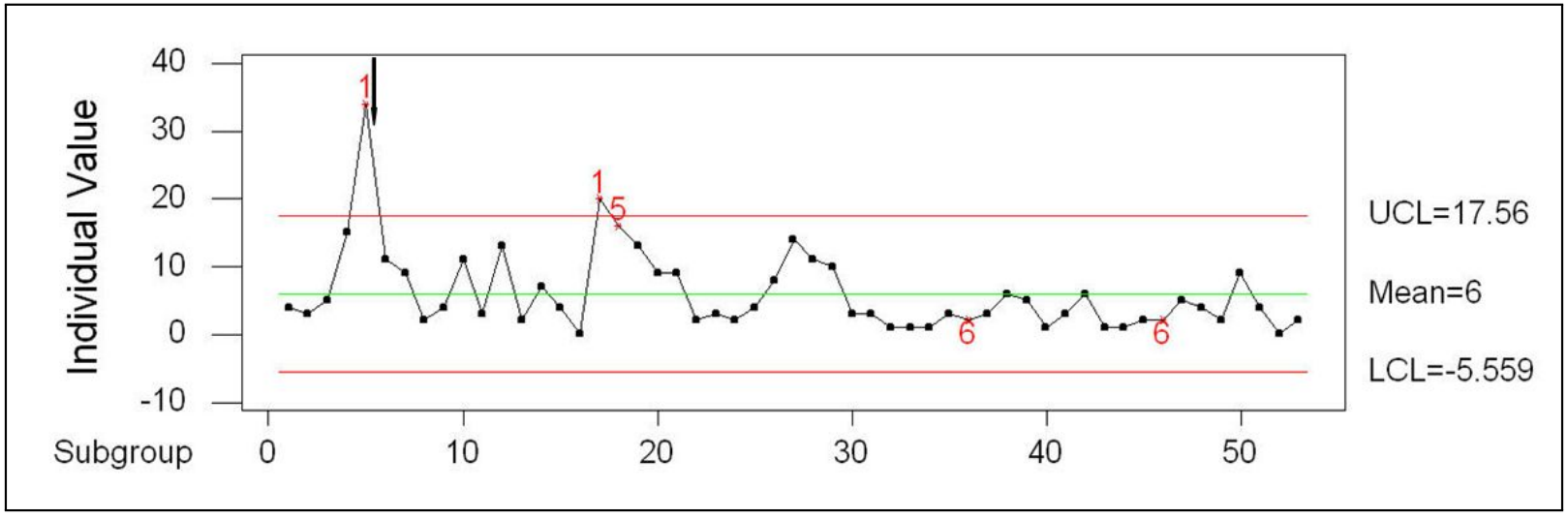


Outlier detection methods

1. Methods for dependent / independent observations
2. Parametric / distribution–free methods
3. Univariate / multivariate methods

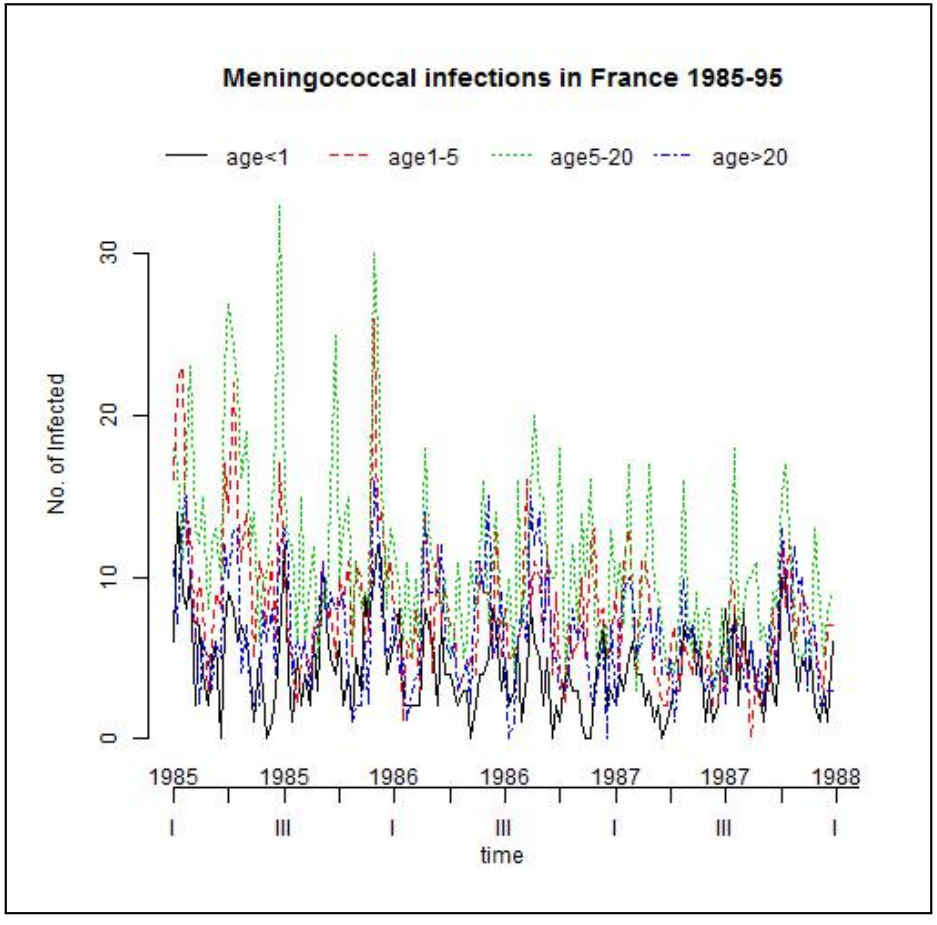
Outlier detection methods - dependent observations

- Dependent: closely related to Statistical Process Control (SPC) methods
- There are nine (Nelson's) criteria to characterise special cases



Outlier detection methods - infectious disease surveillance

- `package(surveillance)` implements the methods of Farrington et al (JRSS–A, 1996) *A statistical algorithm for the early detection of outbreaks of infectious disease*



EDA methods

- *Exploratory Data Analysis (EDA) – Tukey's (1977):*
  - Data samples contain outliers, and the larger the sample size the higher the probability of getting at least one outlier
  - To detect outliers, use empirical quartiles, which are insensitive to large deviations, and define the central part of the data
  - Completely data–driven
  - Problem: quantiles may not be uniquely defined for discrete data

Non–parametric methods: Boxplots

- Defined by Tukey (*Exploratory Data Analysis, 1977)*
- Observation Y is an outlier if
- Box is defined by hinges $[Q1, Q3]$, with $Q2$ in the middle
- Any observation, Y, such that

$$Y < Q1 - k \times \text{ IQR} \quad \text{or} \quad Y > Q3 + k \times \text{ IQR}$$

where IQR = $Q3 - Q1$

is considered as an outlier and marked on the plot

- Whiskers go from the limits of the box to the most distant points that are no outliers
- $k = 1.5$ flags ”‘out”’ values; $k = 3$ flags ”‘far out”’

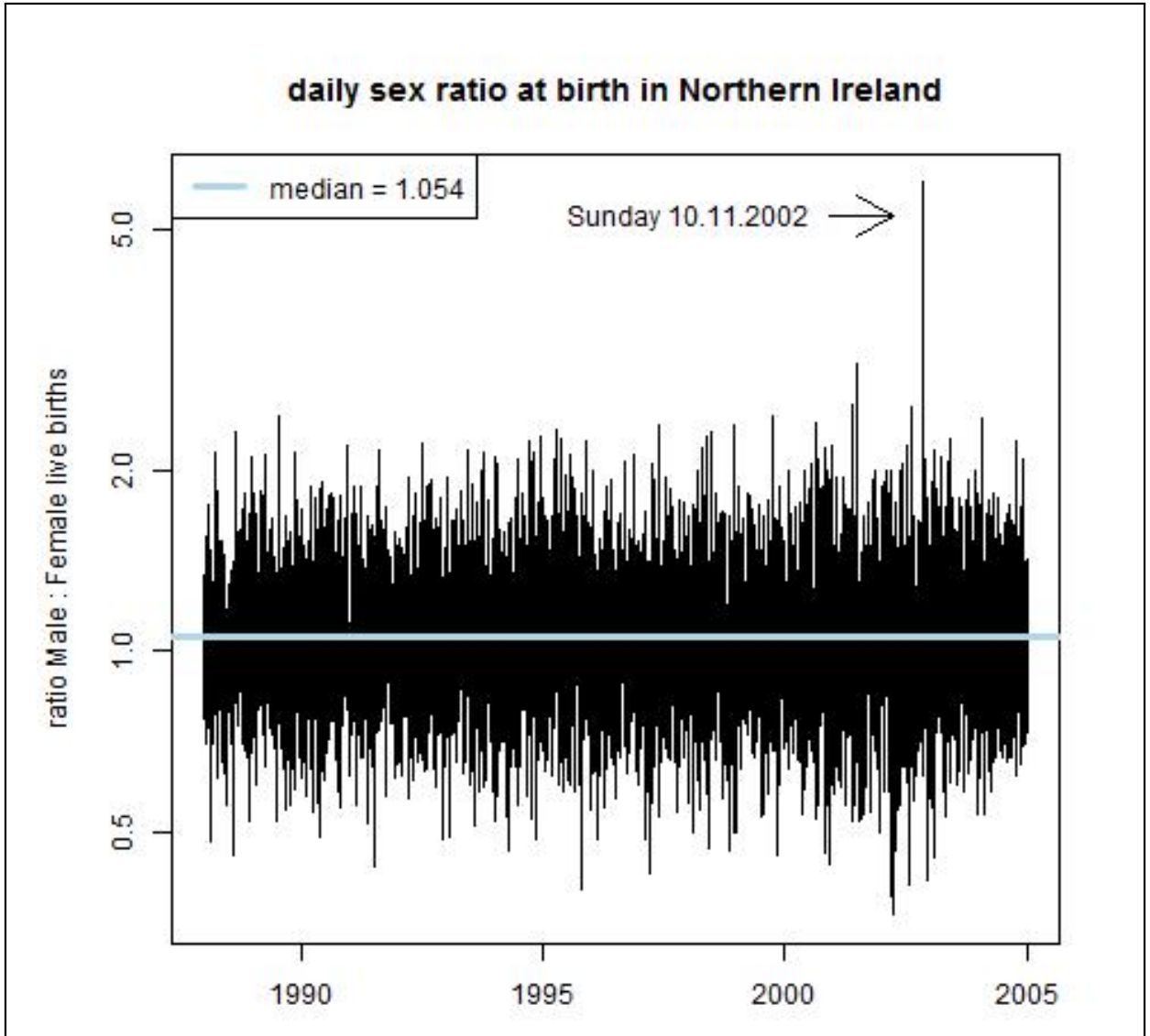
Non–parametric methods: Boxplots – why 1.5?

- Paul Velleman, a student of John Tukey, asked "'Why 1.5?'", Tukey answered, "'Because 1 is too small and 2 is too large.'"
- This factor is appropriate for identifying outliers in symmetric, continuous distributions
- The information given about the tails – outliers – is often not reliable for skewed data
- Typically the upper whisker is too short resulting in too many outliers

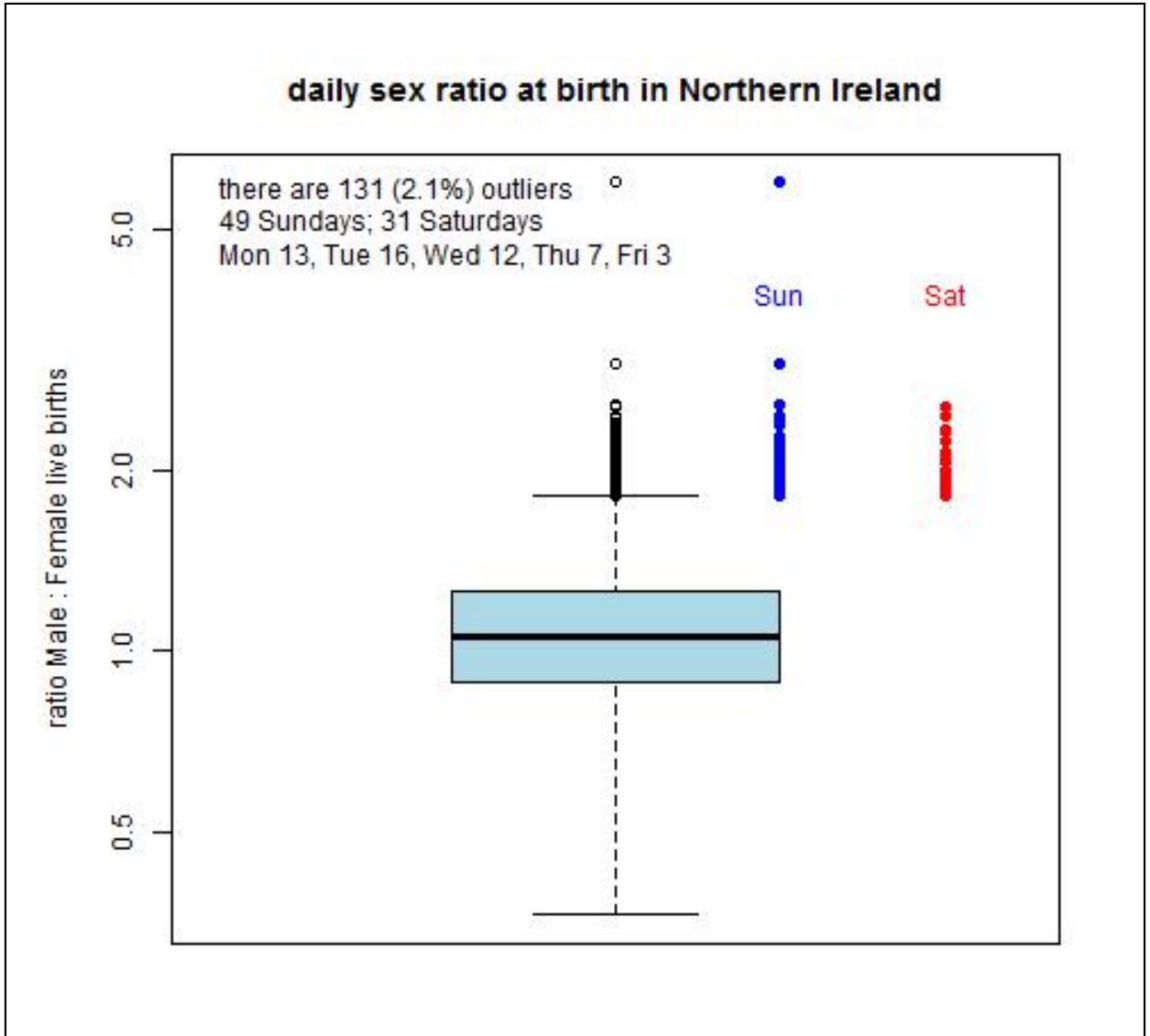
Non–parametric methods: Boxplots – why 1.5?

- Tukey's rules are not sample–size dependent and the probability of labelling outliers when none exists changes with n
- Hoaglin et al (*JASA*, 1986) *Performance of some resistant rules for outlier labeling* found $k = 1.5$ too liberal and $k = 3$ too conservative for moderate n from $Y \sim N$
- Several modifications have been suggested... but the 1.5 rule is very widely used

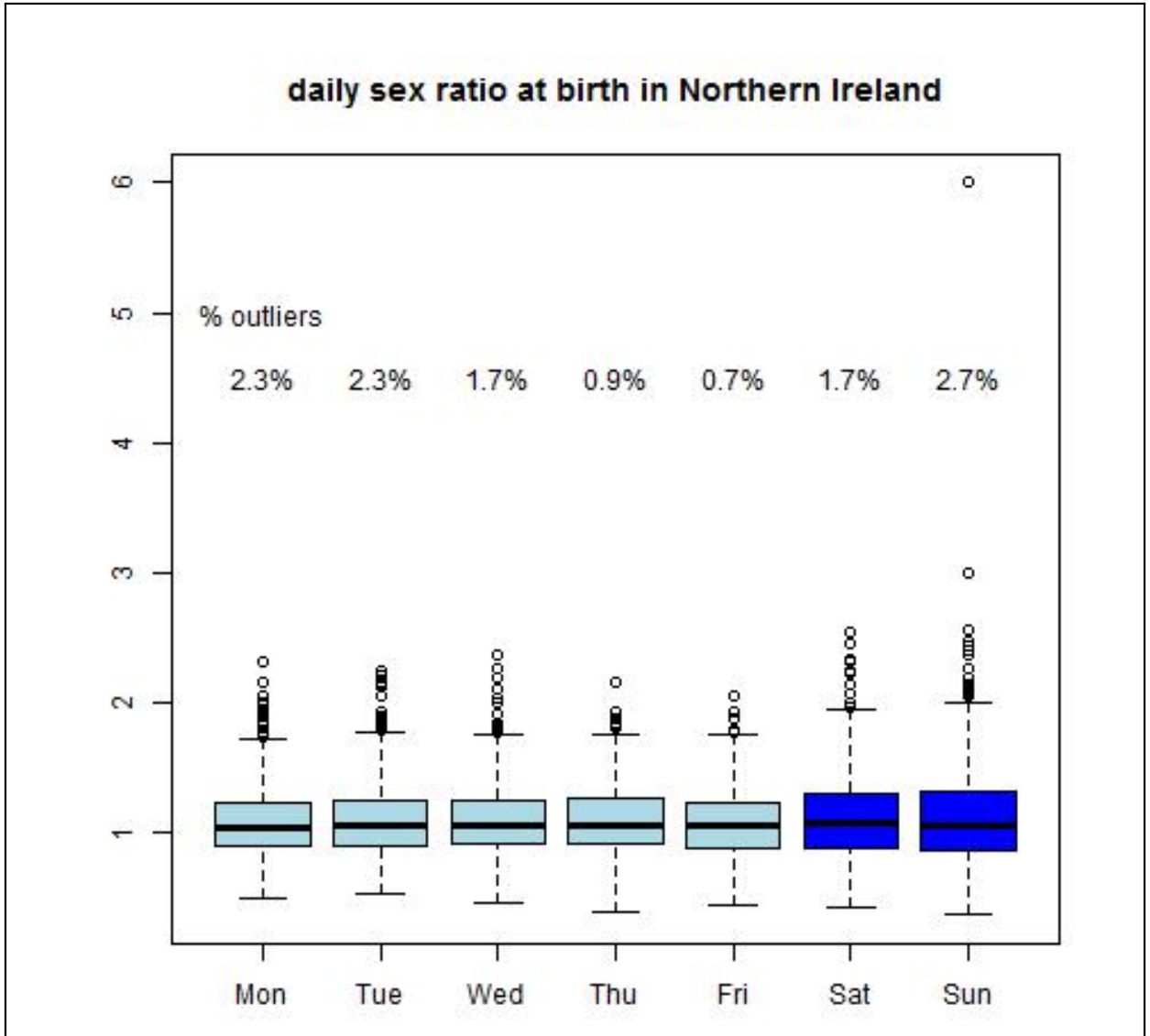
Labelling outliers



Labelling outliers



Labelling outliers



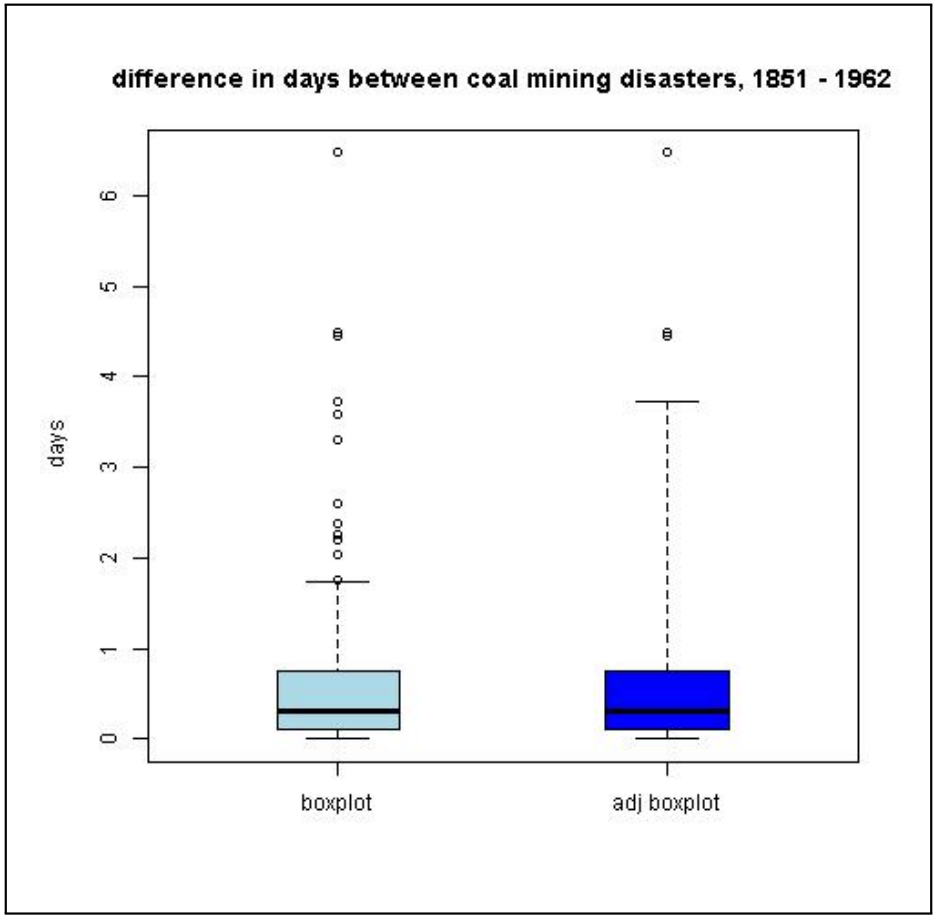
Non–parametric methods: Boxplots – which quantile definition?

- Boxplots provide a simple way to label outliers
- There are many ways to define sample quantiles, cf Hyndman & Fan, (*TAS* 1996) *Sample quantiles in statistical packages*
- For discrete r.v. the pdf F is a right–continuous step function, with the height of the step being Pr $[Y = y]$
- In the extremes, and for finer partitions the quantiles may not be unique

Non–parametric methods: adjusted boxplots

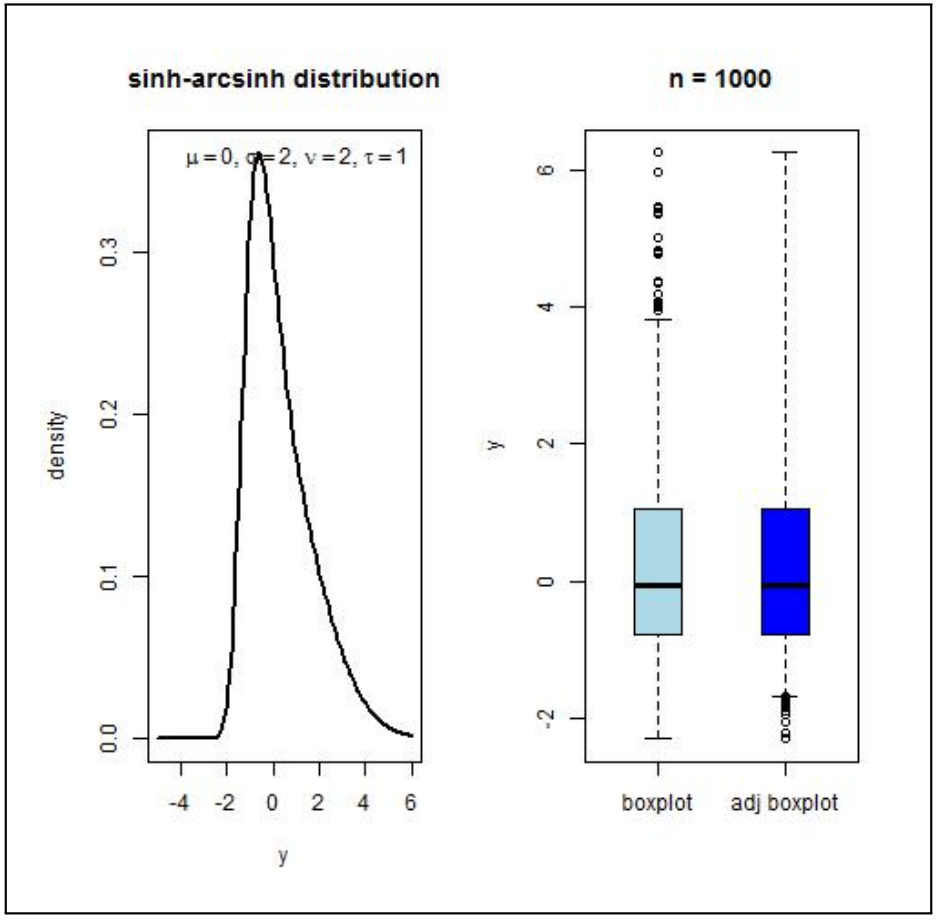
- For right–skewed distributions the boxplot labels too many large outliers and too few small outliers
- A simple modification, based on a robust estimate of skewness is in Hubert and Vandervieren (*Comp Stats Data An* 2008) *An adjusted boxplot for skewed distributions*
- It's available in `package (robustbase)`

Non–parametric methods: adjusted boxplots



Non–parametric methods: adjusted boxplots

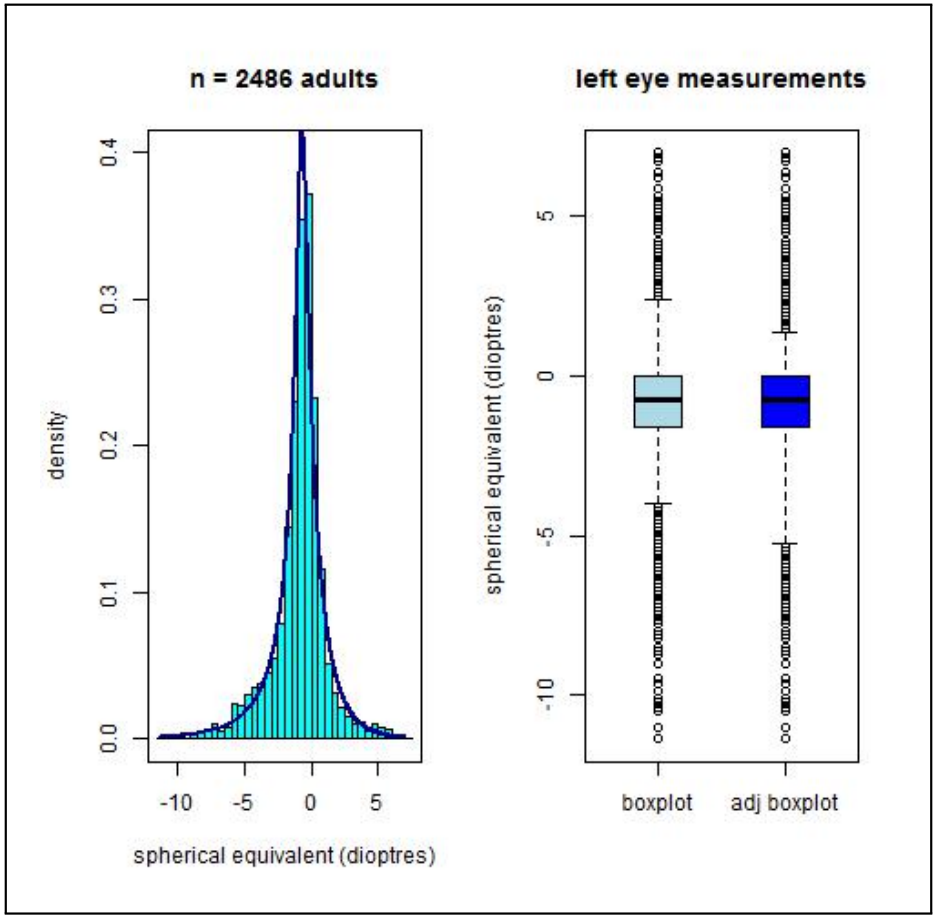
- This adjusted boxplot doesn't correct for kurtosis, and it may be more liberal than the boxplot
- An example is from the four–parameter distribution proposed by Jones & Pewsey (*Biometrika* (2009) *Sinh-arcsinh distributions*



Non–parametric methods: adjusted boxplots

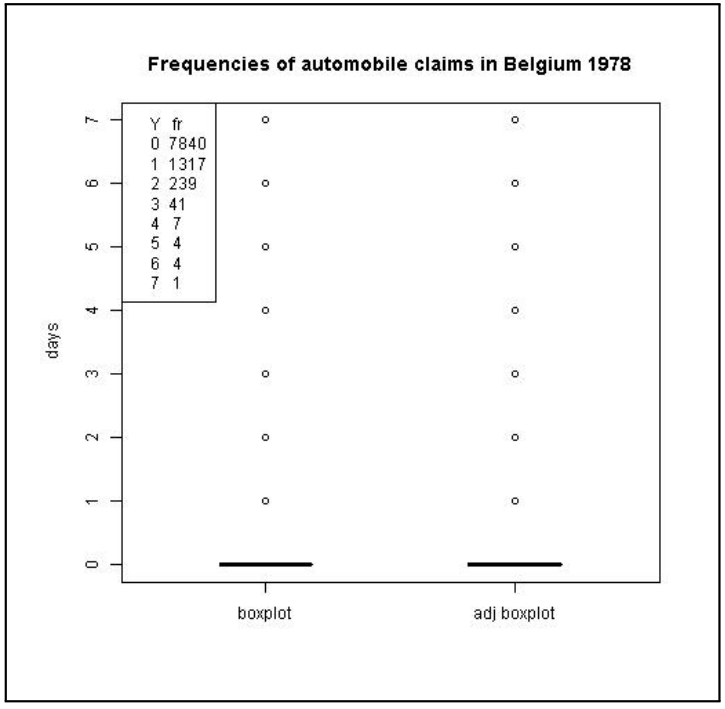
- The SHASH model describes well kurtotic and asymmetric datasets
- `package (gamlss)` Rigby & Stasinoupoulos, *JRSS-C* (2005) fits this and other 3– and 4–parameter densities
- For example the spherical equivalent (dioptres) in adults
- In this example boxplot identifies 261 outliers vs 285 from adjusted boxplot

Spherical equivalent example



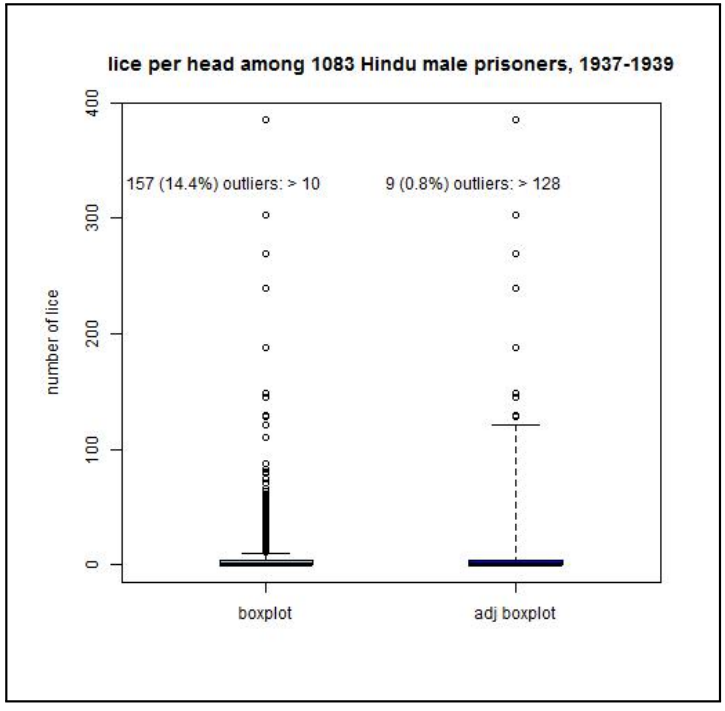
Boxplots and discrete data

- Often, for zero–inflated discrete data neither boxplot work
- This is a not particularly overdispersed dataset but $Q1 = Q2 = Q3 = 0$ so no outliers can be identified



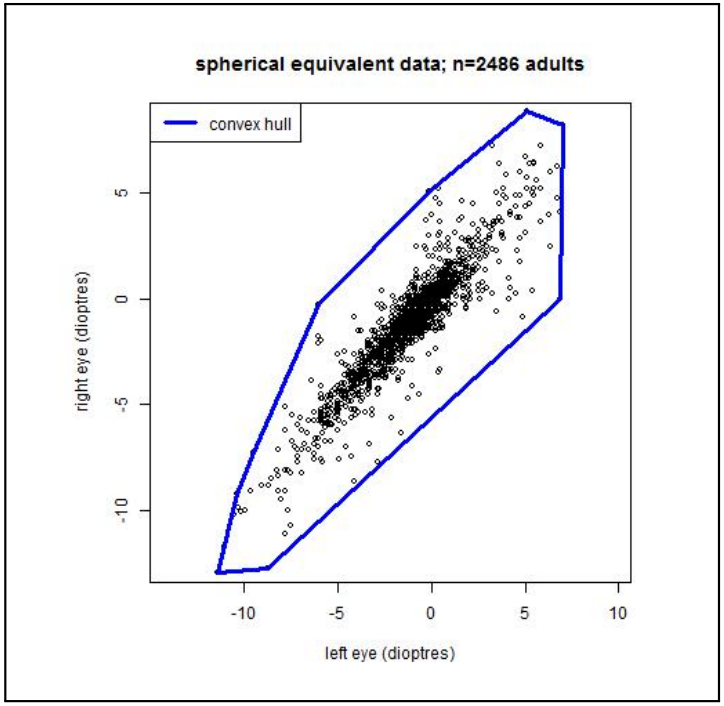
Boxplots and discrete data

- Often, for overdispersed discrete data neither boxplot work – they're arguably too liberal or too conservative in outlier labelling
- This is a zero–inflated (57%) and very overdispersed (OD=85) count dataset
- The data come from Williams, (*J of Ecology* (1944), *Some applications of the logarithmic series and diversity index to ecological problems*)



Bivariate outliers

- Joint outliers might not be outliers in the marginals
- Outliers tend to appear at the extremes of the data space; typical observations occur at its centre – if there is one!
- Convex hull–based: data depth, e.g. Rousseeuw & Ruts, *Comp Stats & Data Analysis* (1996) *Computing Depth Contours of Bivariate Point Clouds*

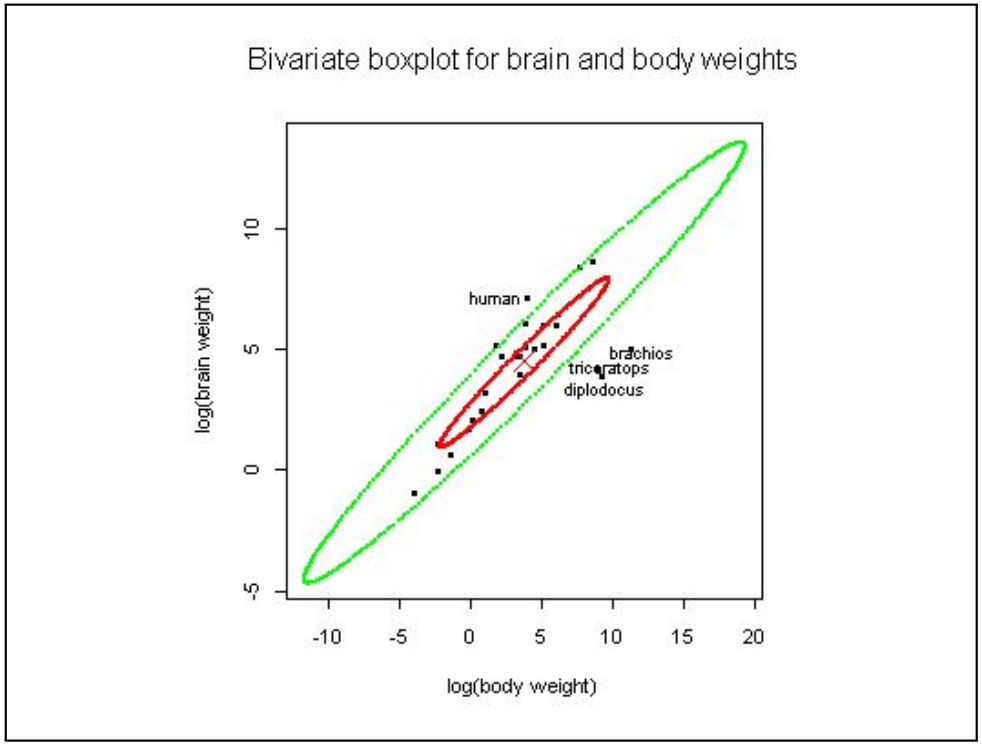


Bivariate extensions of boxplots - replot and quelplot

- Both proposed by Goldberg & Iglewicz, *Technometrics* (1992) *Bivariate extensions of the boxplot*
- The relplot is based on robustly fitting a bivariate Gaussian pdf, and drawing 50% (box) and 99% (whiskers) confidence ellipses
- The quelplot adds two degrees of asymmetry, accounting for residuals on both the major and minor axes of the ellipse

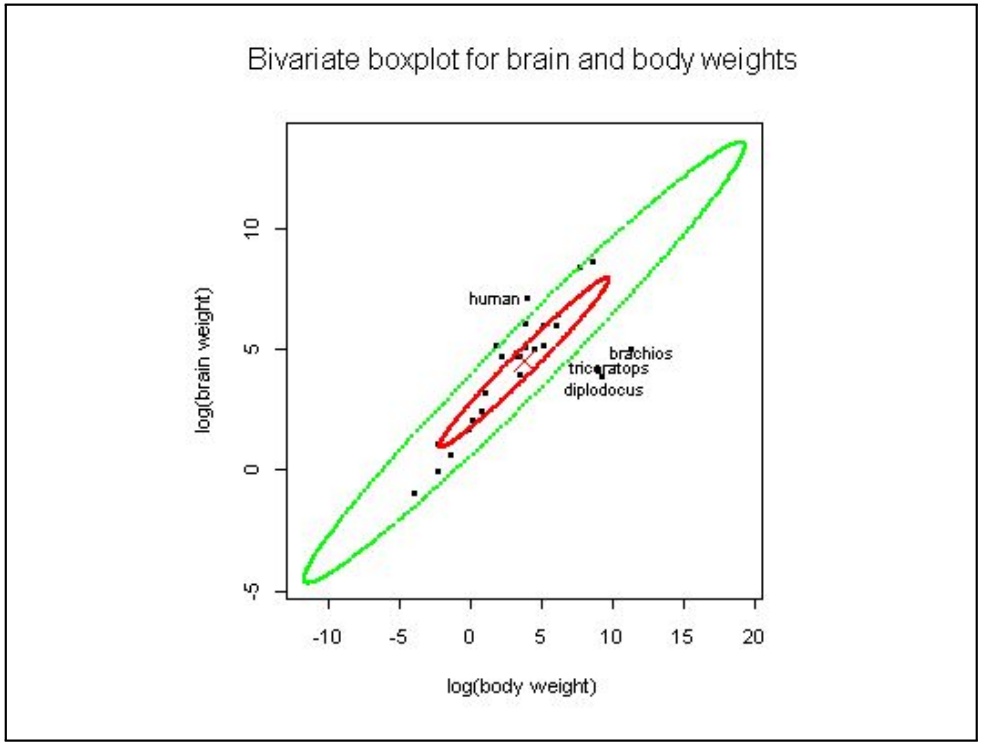
Bivariate boxplots – replot and quelplot

- Relplot = Robust elliptical plot
- Note the assumption of bivariate normality
- The estimate of the variance–covariance matrix is robust vs outliers



Bivariate boxplots – replot and quelplot

- Relplot = Robust elliptical plot
- Note the assumption of bivariate normality
- The estimate of the variance–covariance matrix is robust vs outliers



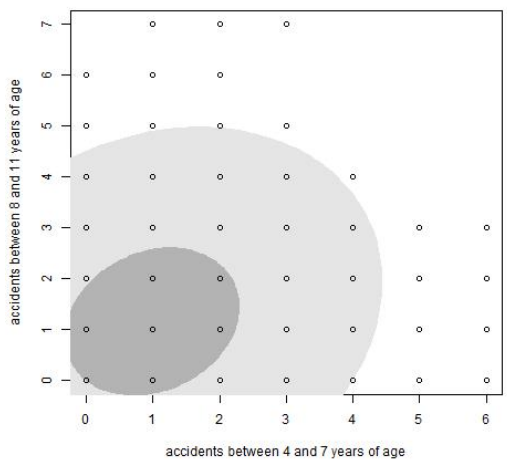
Bivariate boxplots – replot and quelplot

- quelplot = quarterly robust elliptical plot
- Accidents in 621 children - two periods, (4,7), (8,11) years of age

	0	1	2	3	4	5	6	7
0	101	76	35	15	7	3	3	0
1	67	61	32	14	12	4	1	1
2	24	36	22	15	6	1	2	1
3	10	19	10	5	2	4	0	2
4	1	7	3	4	2	0	0	0
5	2	1	4	2	0	0	0	0
6	1	1	1	1	0	0	0	0

Example: relplot of chidren's accidents in two periods

- Data from Mellinger et al *JASA* (1965), *A mathematical model with applications to a study of accident repeatedness among children*



Model–based methods

- *Model–based outliers, e.g. Ronan M Conroy (allstat, 15.07.1999)*: I'm not an outlier; I just haven't found my distribution yet
- Significance tests depend on pivotal quantities and optimal criteria
- Tests are available for continuous distributions, mostly for the Gaussian

Model–based methods

- *Model–based outliers, e.g. Ronan M Conroy (allstat, 15.07.1999)*: I'm not an outlier; I just haven't found my distribution yet
- Significance tests depend on pivotal quantities and optimal criteria
- Tests are available for continuous distributions, mostly for the Gaussian

Model–based methods

- There are very few model–based outlier tests for discrete r.v.'s
- The classic procedure (Hawkins, 1980), is:
  - Find a sufficient statistic T for the parameters
  - Find a suitable outlier test statistic $h(Y, T)$ whose distribution is independent of the unknown parameters
- This pivotal quantity, h exists for only very few discrete distributions
- There are procedures based on the exact distribution of the n–th order statistic but this assumes knowing the parameters

Long–tailed discrete distributions

- Parameter–mix distributions
- Usually constucted by assuming a continuous pdf for the parameter of a Poisson distribution
- This can be seen as an indicator of variable frailty, or as a random effect
- The best known model is: Negative Binomial = (Poisson, Gamma) mix
- Two related models:
  - Holla = (Poisson, Inverse Gaussian) mix
  - Sichel = (Poisson, Generalised Inverse Gaussian) mix

Surprise

- *Oxford English Dictionary*: an unexpected occurrence or event; anything un-expected or astonishing
- Applied statisticians evaluate how surprising the observed value of a random variable is with respect to a probability model
- A *p*-value may be defined as the probability of observing values of the test statistic at least as extreme as the observed one assuming the null model is correct
- No explicit formulation of alternative models is required so this measure has appealed to statisticians over the years.

Surprise

- Good *Some Logic and History of Hypothesis Testing* (1981): The evolutionary value of surprise is that it causes us to check our assumptions. Hence if an experiment gives rise to a surprising result given some null hypothesis H it might cause us to wonder whether H is true even in the absence of a vague alternative to$H$. It is therefore natural to consider whether a statistical test of H might be to depend upon some index of surprise

Surprise

- A surprise measure other than the *p*-value was first proposed by Weaver, *Probability, Rarity, Interest and Surprise* (1948), *The Scientific Monthly*
- Generalized by Good, *The Surprise Index for the Multivariate Normal Distribution* (1956) *Ann of Math Stats*
- Compares the expected value of the random variable whose possible values are the probabilities of the observations and the individual probabilities of such observation
- Very few applications: expressions for the surprise index available only for very few pdf's
- Well–suited for analysing discrete long–tailed distributions

Surprise Index

- Weaver's surprise index: an empirical measure of how unexpected an observed value of a random variable is
- Low probability implies rarity but not necessarily surprise; a surprising event is always rare
- E.g. winning the lottery is certainly a rare event but it's not surprising that somebody wins the lottery as each combination has an equal probability of occurring
- E.g. tossing a coin: $\Omega = \{\text{head, tail, edge}\}$ with probabilities $\left\{\frac{1-\epsilon}{2}, \frac{1-\epsilon}{2}, \epsilon\right\}$, where ϵ is the probability a coin lands on its edge

Surprise Index

- Let a discrete r.v. V take values in Ω of $\{V_1, V_2, \ldots\}$ with probabilities $\{p_1, p_2, \ldots\}$
- The expected value of these random quantities is $\mathrm{E}(V) = p_1\, V_1 + p_2\, V_2 + \ldots$
- This assumes we know the true model H defining the p_i's
- The SI for each V_i, $\lambda^{(i)}$ is:

$$\lambda_i = \frac{\mathrm{E}\,(p\,|H)}{p_i} = \frac{1}{p_i} \sum_{j=1} p_j^2$$

Surprise Index

- For the lottery example $\lambda = 1$ for all combinations of lottery numbers, as each combination has an equal probability of occurring
- For the coin tossing experiment, $\lambda \sim 1$ for head or tail, and $\lambda = \frac{3\epsilon}{2} + \frac{1}{2\epsilon} - 1$ for the edge event

Surprise Index – Weaver's interpretation

- Weaver suggested the following categories to determine if a value of SI may be considered as large enough to correspond to a surprising event
- We follow Weaver's conventional scale and consider an observation occurring with probability p_i as an outlier if $\lambda_i \geq 1000$

<5	Not surprising
10	Begins to be surprising
10^3	Definitely surprising
10^6	Very surprising
10^{12}	Miracle!

Surprise Index for discrete pdf's

- The only analytical expressions published for Surprise Indices λ are those of the Binomial, and Poisson distributions, obtained by Redheffer, *A note on the surprise index*, (1951) *Ann Math Stats*
- Good, *The Surprise Index for the Multivariate Normal Distribution*, (1956) *Ann Math Stats* calculated λ for the Normal and multivariate Normal distributions

Examples of SI's

- Poisson (μ) : $\lambda_i = \dfrac{\mathrm{I}_0(2\,\mu)\,i!}{e^{\mu}\,\mu^i}$
- Negative Binomial (r, p) :

$$\lambda_i = (1-p)^{-i}\,p^r\,{}_2\mathrm{F}_1\left[r, r, 1, (p-1)^2\right]\begin{pmatrix} r+i-1 \\ r-1 \end{pmatrix}^{-1}$$

- General zero–inflated with ω as the mixture parameter :

$$\lambda_{\mathrm{ZI},i} = \frac{(1-\omega)^2\,\lambda_i + 2\,p_0\,\left(\omega-\omega^2\right)+\omega}{\left[\omega\,+(1-\omega)\,p_0\,\right]\,U_0^{(i)} + \left[(1-\omega)\,p_i\,\right]\,\left(1-U_0^{(i)}\right)}$$

where $U_0^{(i)}$ is an indicator function being 1 for $i = 0$ and 0 otherwise

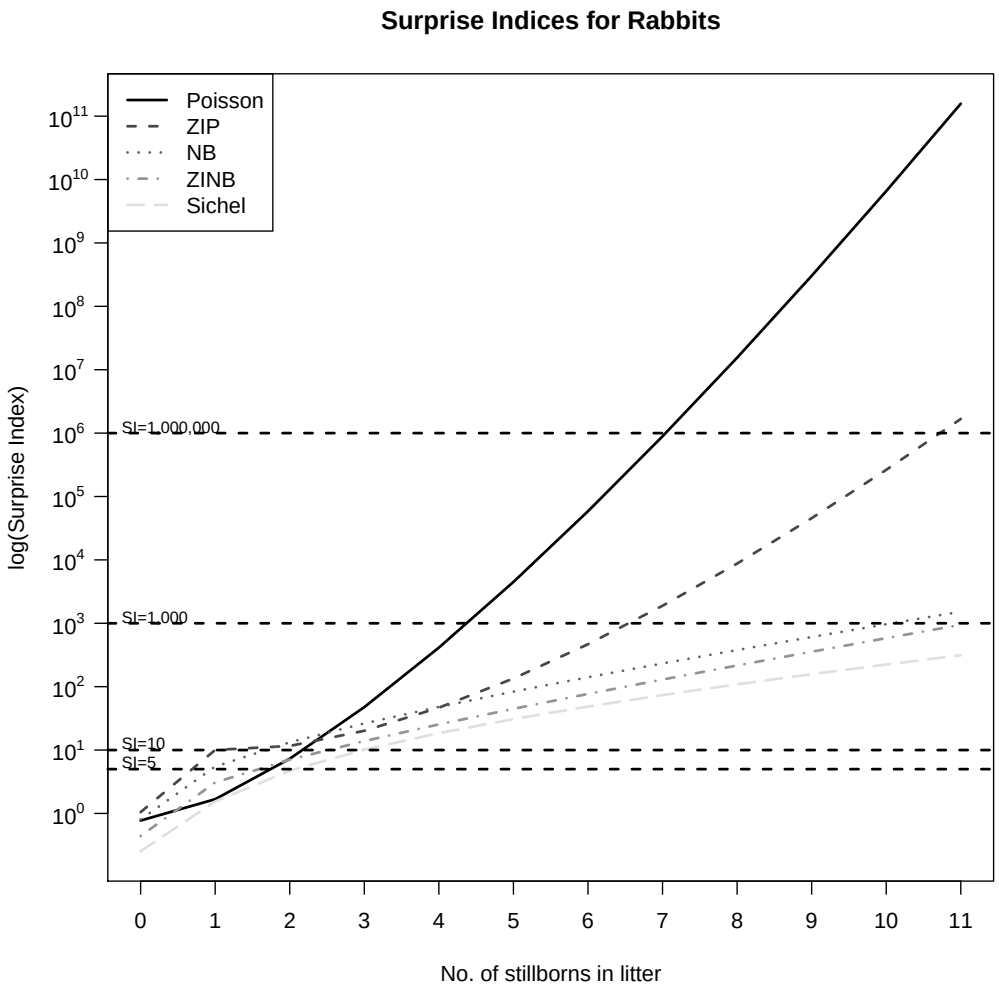
Surprise Index in practice

- The hypothesised data–generating mechanisms imply pdf's to be fitted
- Models fitted using maximum likelihood
- There are numerical issues with some densities
- Good initial estimates are often available via rapid estimation methods, e.g. matching moments or pgf values
- Model selection is based on $\mathrm{BIC} = -2\ell + d\ \ln(n)$ where ℓ is the maximized log–likelihood, d is the number of parameters in the model and n is the number of independent observations

Example: Stillborns in litters of NZ white rabbits; Morgan et al *Negative Score Test*, (2007) *Am Statsn*

Distribution	No. of Stillbirths												BIC
	0	1	2	3	4	5	6	7	8	9	10	11	
Observed	**314**	**48**	**20**	**7**	**5**	**2**	**2**	**1**	**2**	**0**	**0**	**1**	—
Poisson	254	117	27	4	·	·	·	·	·	·	·	·	887.7
ZIP	314	33	28	16	7	2	1	·	·	·	·	·	726.4
NB	314	46	19	10	5	3	2	1	1	·	·	·	686.3
ZINB	314	46	19	10	5	3	2	1	1	·	·	·	692.3
Sichel	314	49	18	9	5	3	2	1	1	·	·	·	691.9
ZI Sichel	314	48	18	9	5	3	2	1	1	·	·	·	697.9

Surprise Index: NZ white rabbits litters

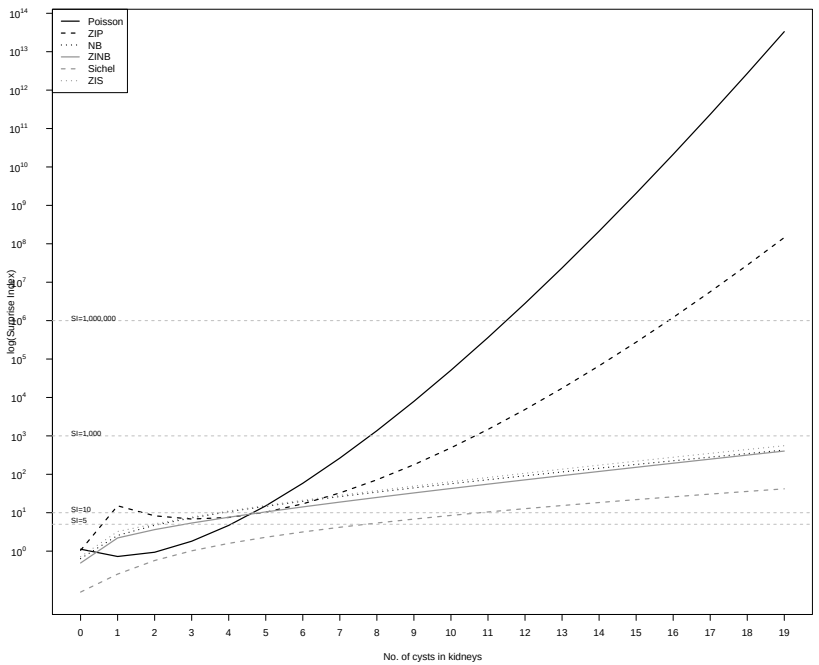


Example: Cysts in embryonic mice kidneys; Chan et al, (2009) *Am J Physiol*

- A group of 111 kidneys subjected to a particular steroid (n=111)
- High proportion of zeroes (58.6%) and OD = 5.7

Dist.	**No. of Cysts in kidneys**																			**BIC**
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
Obs.	**65**	**14**	**10**	**6**	**4**	**2**	**2**	**2**	**1**	**1**	**1**	**2**	·	·	·	·	·	·	·	**1**	—
Poisson	24	37	28	15	6	2	·	·	·	·	·	·	·	·	·	·	·	·	·	·	564.1
ZIP	65	5	8	10	9	7	4	2	1	·	·	·	·	·	·	·	·	·	·	·	418.8
NB	65	16	9	6	4	3	2	2	1	1	1	1	·	·	·	·	·	·	·	·	359.0
ZINB	65	15	9	6	4	3	2	2	1	1	1	1	·	·	·	·	·	·	·	·	414.1
Sich	64	16	9	6	4	3	2	2	1	1	1	1	·	·	·	·	·	·	·	·	364.0
ZI Sich	65	14	9	6	4	3	2	2	1	1	1	1	·	·	·	·	·	·	·	·	368.1

Surprise Index: Cysts in embryonic mice kidneys



Example: Accidents in Belgian drivers 1978; Nikololoupoulos and Karlis (2008) *Comp Stats Data An*

Distribution	No. of Claims								BIC
	0	1	2	3	4	5	6	7	
Observed	**7840**	**1317**	**239**	**42**	**14**	**4**	**4**	**1**	—
Poisson	7635	1637	175	13	1	0	0	0	10990.7
ZIP	7840	1274	296	46	5	0	0	0	10769.5
NB	7847	1288	257	54	12	3	1	0	10714.4
Holla	7844	1306	238	53	14	4	1	0	10705.3
Sichel	7842	1310	236	53	14	4	1	0	10713.7
Waring	7848	1281	253	58	15	4	1	0	10707.7

- For this data set all models gave $SI \geq 1000$ for $Y \geq 7$
- Characterise this driver as an outlier

Example: lice in heads of prisoners

Distribution	BIC
Poisson	29179.8
ZIP	17951.9
NB	4663.6
ZINB	4715.2
Holla	4766.4
ZIHolla	4822.1
Sichel	4661.3
ZISichel	4710.3

- For this dataset the Sichel model gave $SI \geq 1000$ for $Y \geq 120$
- Not too far from the adjusted boxplot result

Conclusion

- Determining ouliers in discrete datasets ... is not easy!
- Extension of the SI to bivariate discrete distributions is possible... but not easy!

Acknowlegements

- This is work done in collaboration with Angie Wade and Fiona McElduff
- Thanks to Bob Rigby and Mikis Stasinopulos for help with `gamlss`