The Data Cleaning Problem:

Some Key Issues & Practical Approaches

Ronald K. Pearson

Daniel Baugh Institute for Functional Genomics and Computational Biology

Department of Pathology, Anatomy, and Cell Biology

Thomas Jefferson University
Philadelphia, PA

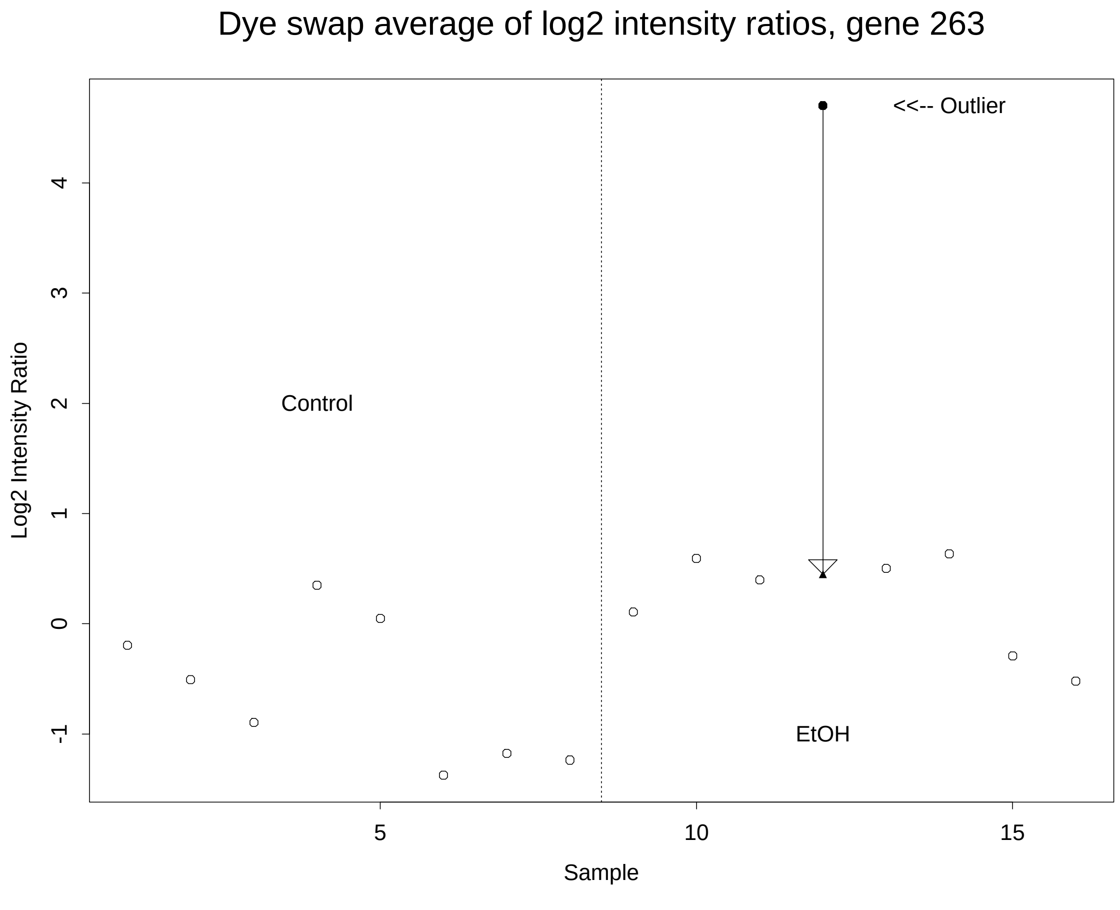
DIMACS Workshop on Data Quality, Data Cleaning and Treatment of Noisy Data

November 3-4, 2003

Topics

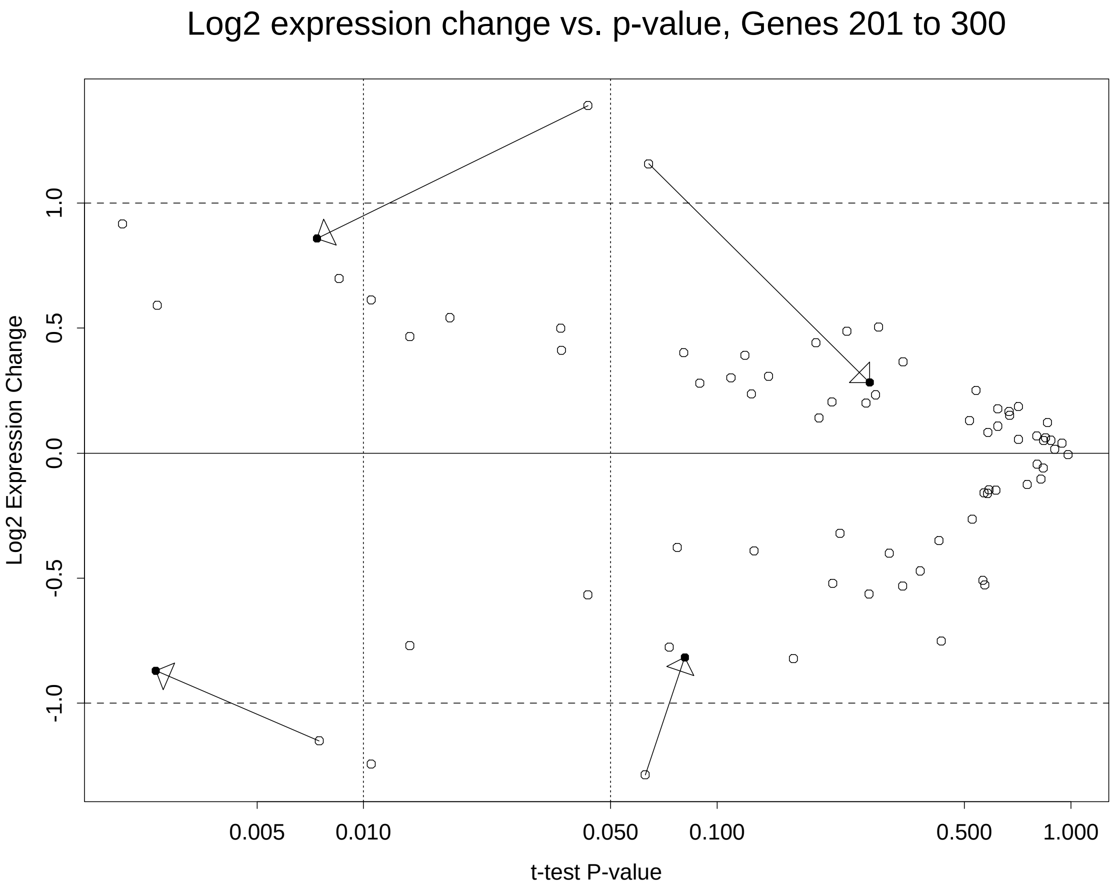
1. Outliers: an important data anomaly
   - types and working assumptions
   - some real data examples
2. Detecting outliers
   - the popular 3σ edit rule
   - order-statistics vs. moments
   - some alternative approaches
3. Other data anomalies
   - missing data
   - misalignments
   - noninformative variables
   - comparing performance

Example 1:
Outlier in a microarray data sequence



Example 2:

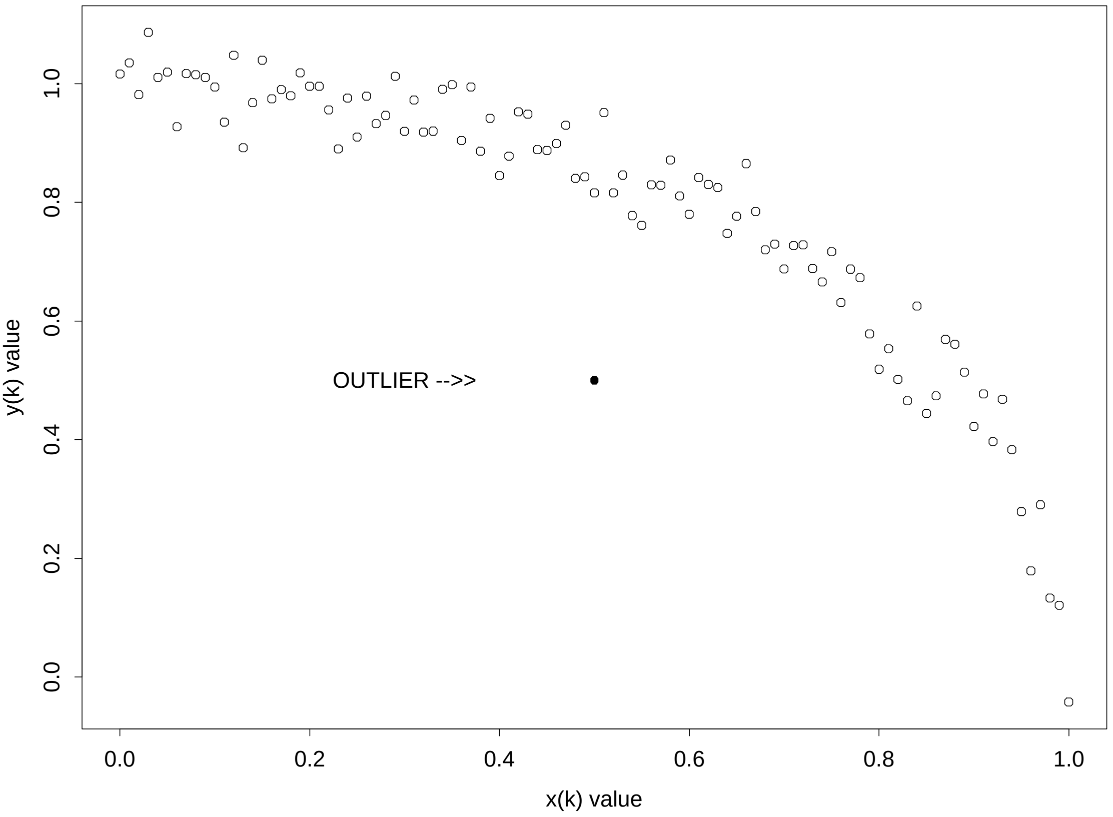
Influence of outliers on a volcano plot



Example 3:
Bivariate outlier in a simulated dataset

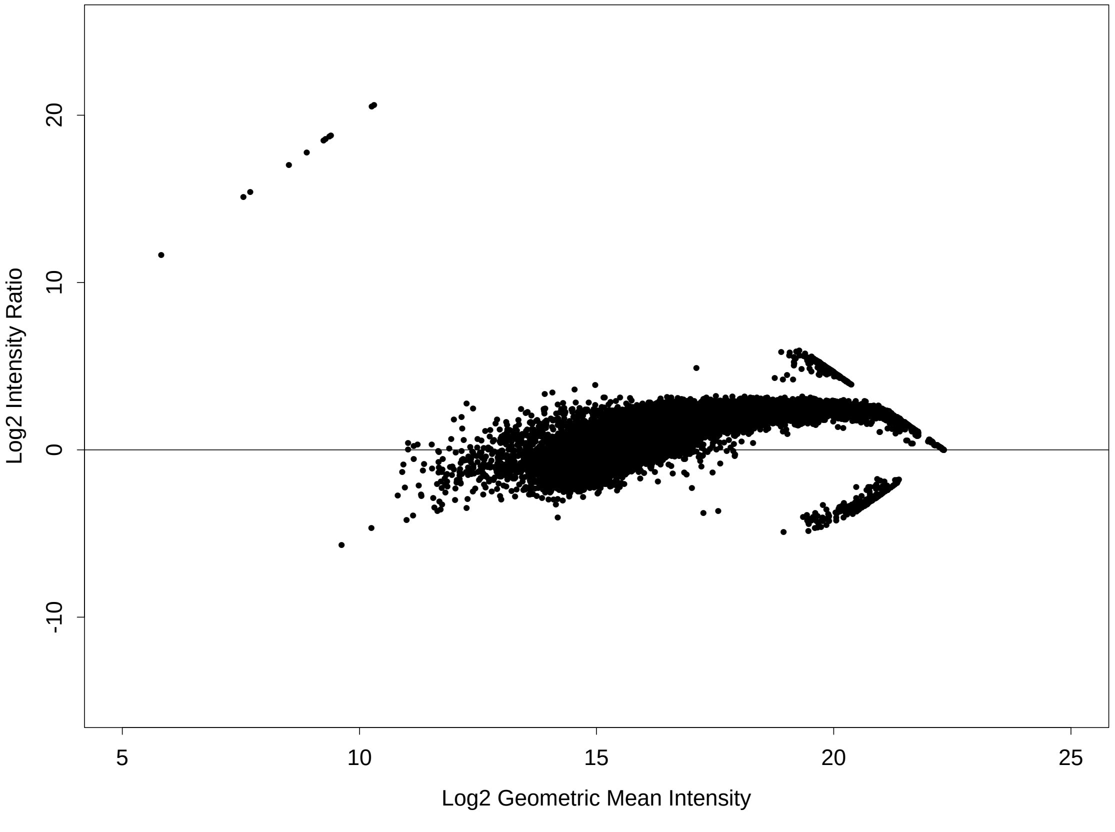
⇝ **NOTE:**

Outlier is not extreme with respect to either x or y individually



EXAMPLE 4:
Bivariate outliers in a real dataset

MA plot constructed from an uncorrected microarray dataset

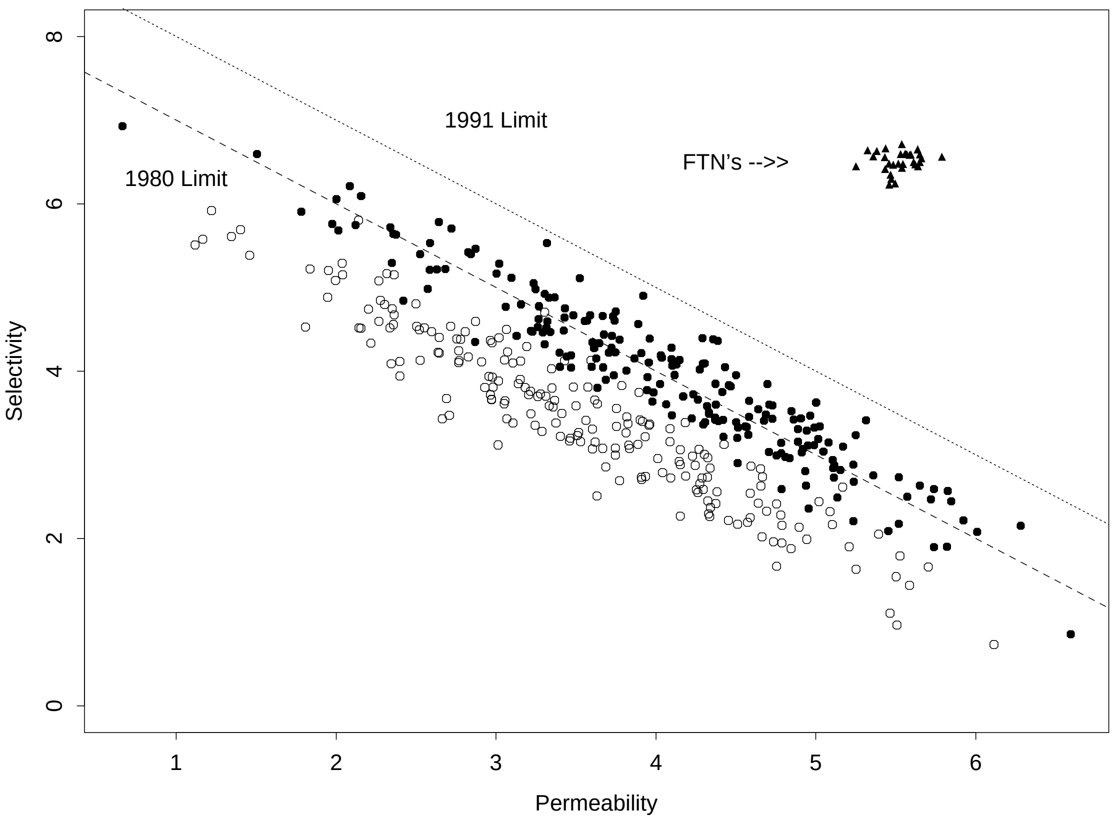


EXAMPLE 5:

Multivariate outliers in material property relationships

⇝ NOTE:

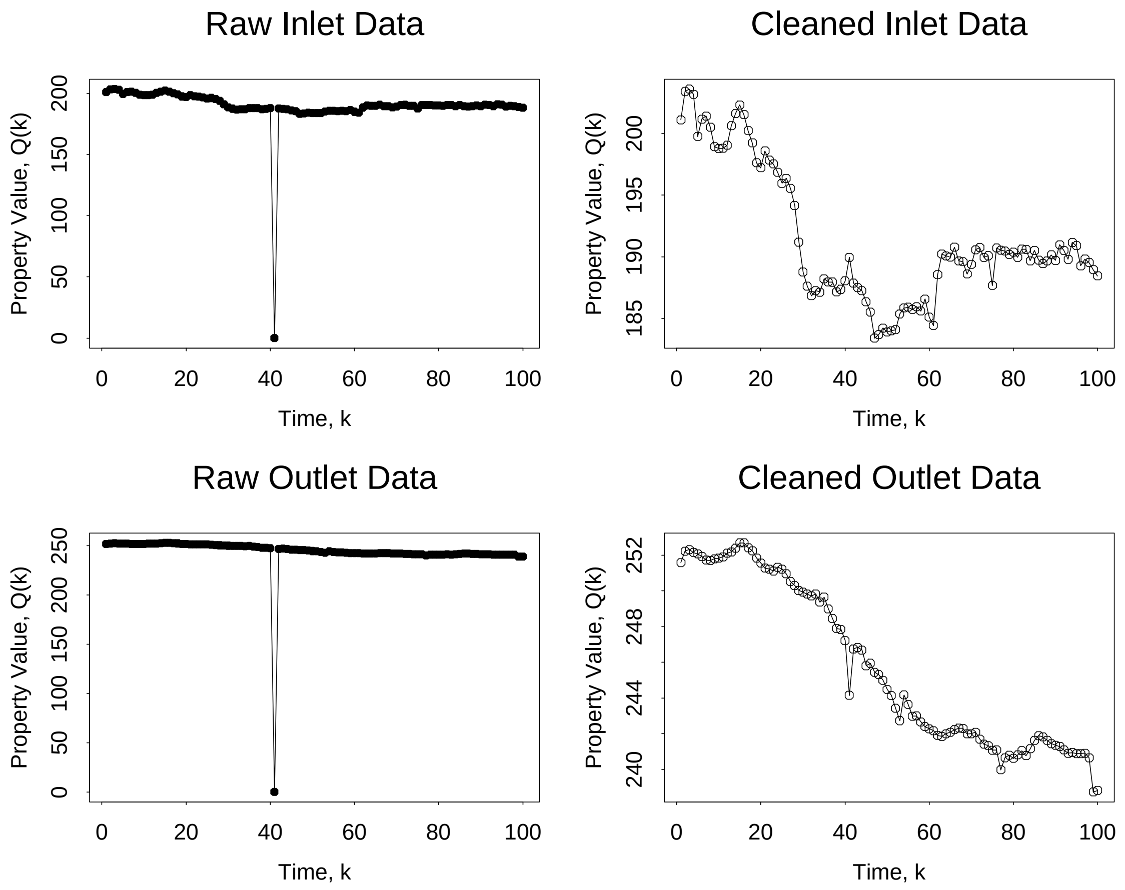
Here, outliers correspond to unusually good materials



Example 6:
Common mode outlier example

$\rightsquigarrow$ NOTE:

Univariate outliers can be highly correlated in different variables



The 3σ Edit Rule

- Procedure:

$$|x_k - \bar{x}| > 3\hat{\sigma} \;\Rightarrow\; x_k \text{ is an outlier}$$

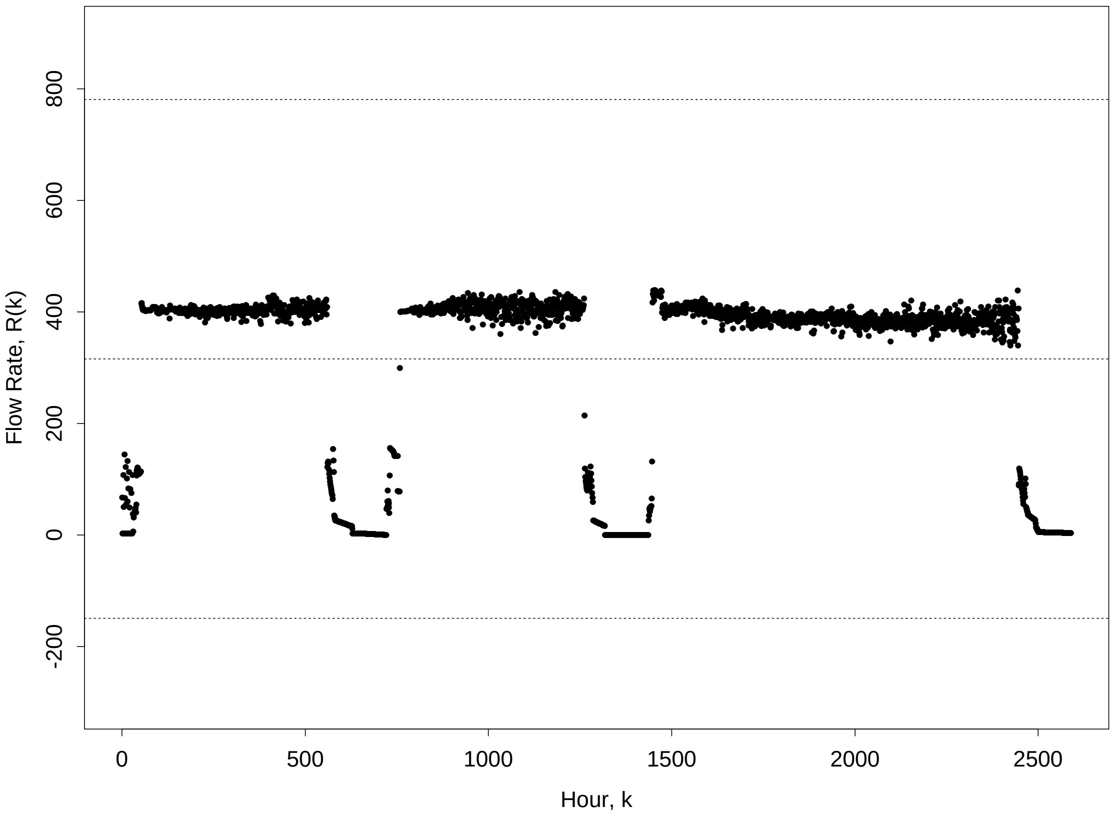
- Motivation:
  1. the Gaussian assumption $x_k \sim N(\mu, \sigma^2)$ is *very* popular
  2. under this assumption:
     Prob $\{|x_k - \mu| > 3\sigma\} \sim 0.3\%$

- History:
  - dates back at least a century:
    T. Wright, *A Treatise on the Adjustment of Observations by the Method of Least Squares*, Van Nostrand, 1884
  - still advocated today:
    S. Draghici, *Data Analysis Tools for DNA Microarrays*, Chapman and Hall/CRC, 2003

A Spectacular Failure:
The flow rate dataset

$\rightsquigarrow$ **NOTE:**

This dataset contains $\sim 20\%$ *visually obvious outliers: none are detected by the* 3σ *edit rule*



Why?

- Basic reason:
  - the mean μ and standard deviation σ are unknown and must be estimated from data

  $\rightsquigarrow$ standard estimators are extremely sensitive to the presence of outliers

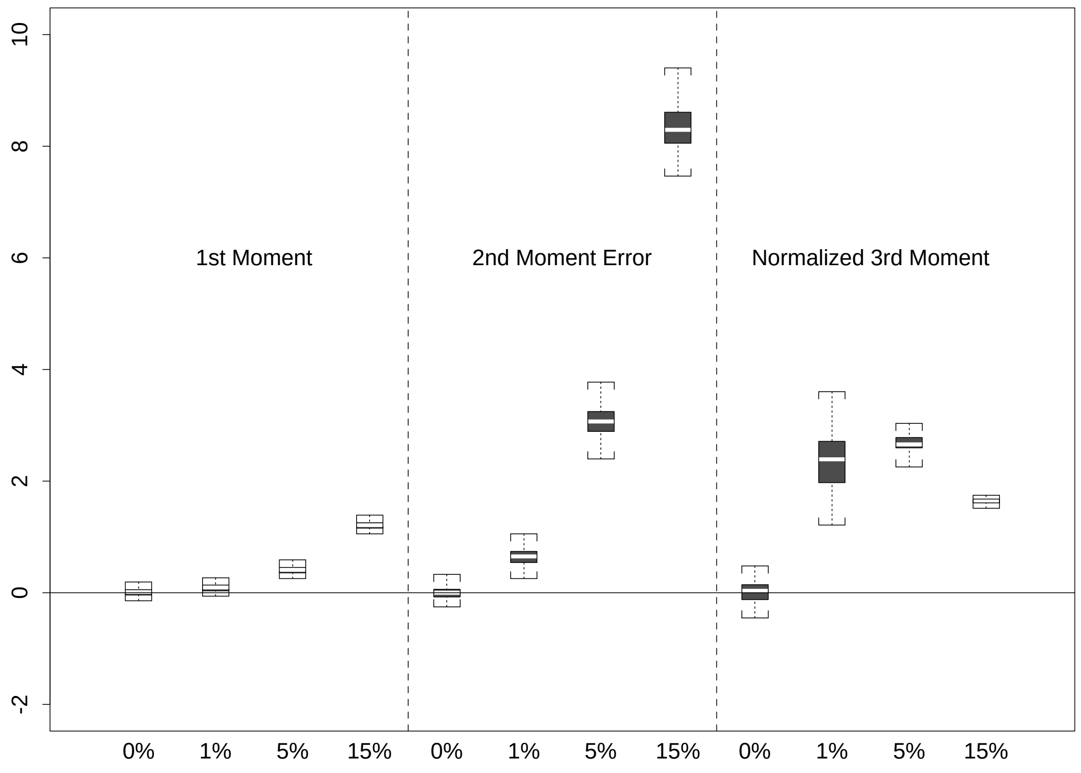
- Specific observation:

  At point contamination levels greater than 10%, the 3σ edit rule will fail completely: *no* outliers will be detected

- To overcome this problem:
  1. replace the mean with an outlier-resistant alternative (e.g., median)
  2. replace the standard deviation with an outlier-resistant alternative (e.g., MAD)

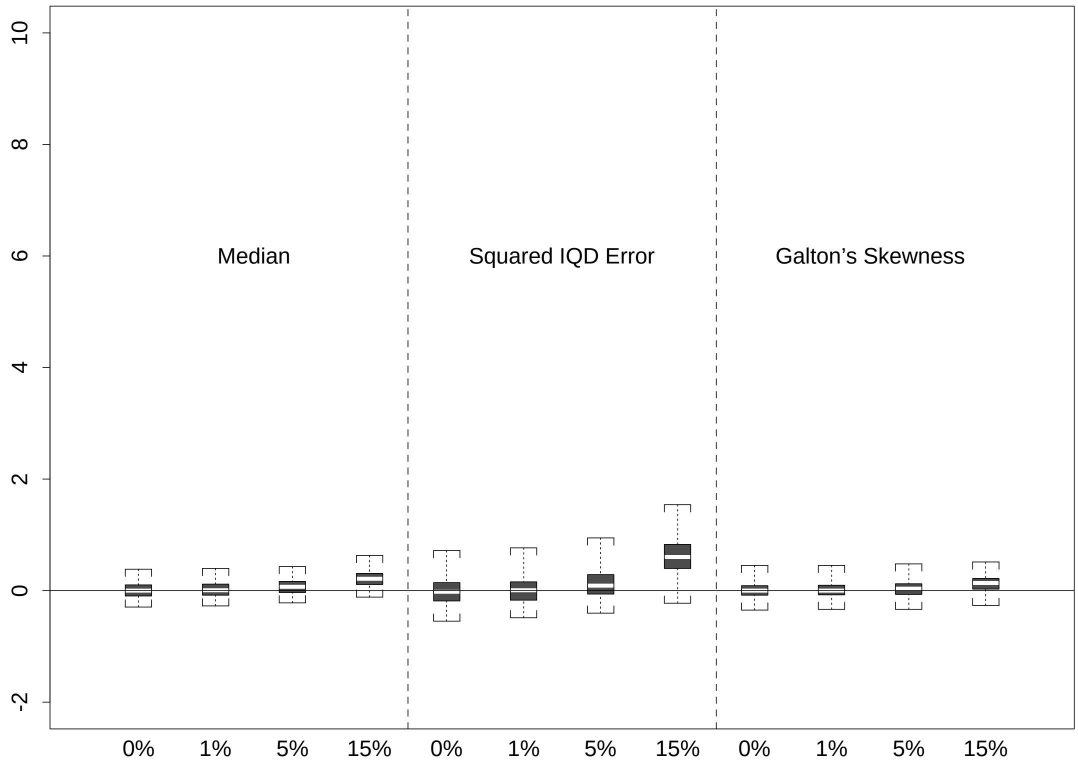
Outlier Sensitivity of Standard Moment Estimators:

Mean, variance, and skewness



Order-based Alternatives:

Median, square of interquartile distance, Galton's skewness



The Hampel Identifier

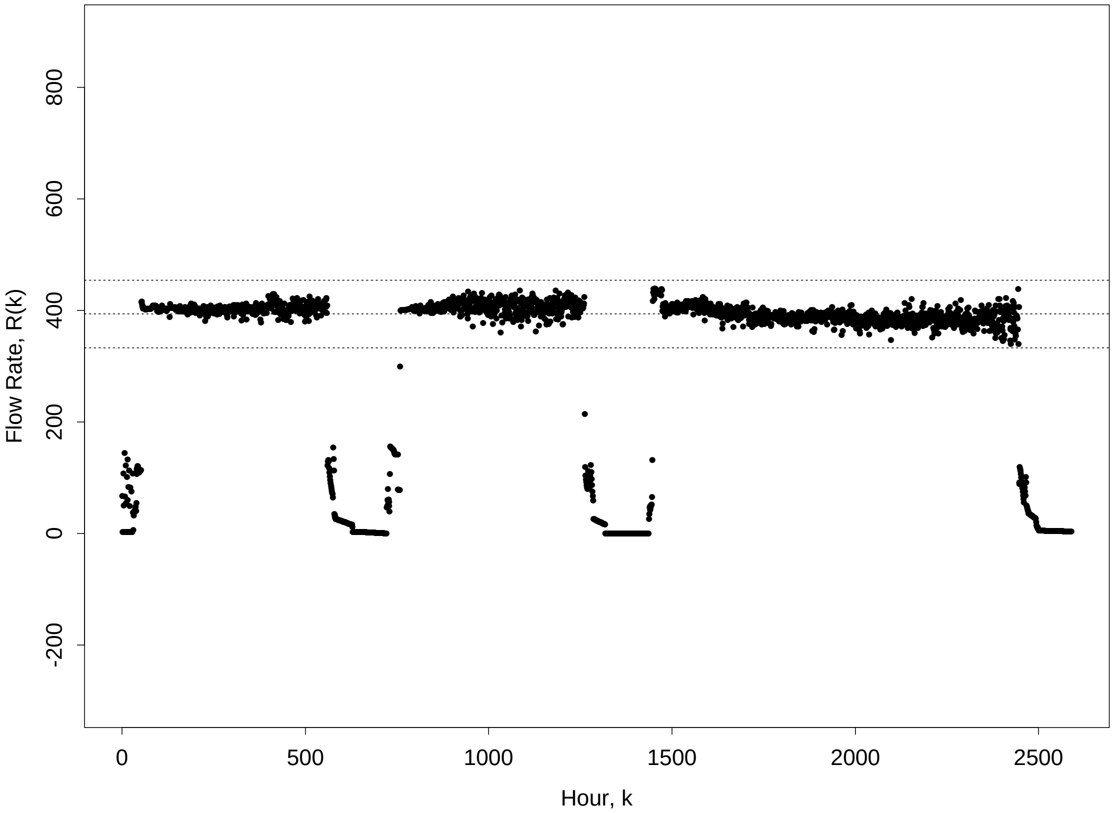
- Idea:
  - replace the mean $\bar{x}$ with the outlier-resistant median $x^\dagger$
  - replace the standard deviation $\hat{\sigma}$ with the outlier-resistant MAD scale estimate S
- The MAD scale estimate:

$$S = 1.4826 \text{ median } \{|x_k - x^\dagger|\}$$

- Interpretation:
  - $d_k = |x_k - x^\dagger|$ measures the distance of each point x_k from the reference value $x^\dagger$
  - the median d_k value tells how far a "typical" point lies from $x^\dagger$
  - the factor 1.4826 makes S an unbiased estimate of σ for Gaussian data

The Flow Rate Dataset Revisited

The Hampel identifier provides a clean separation between normal operation and shutdown episodes



The Boxplot Edit Rule

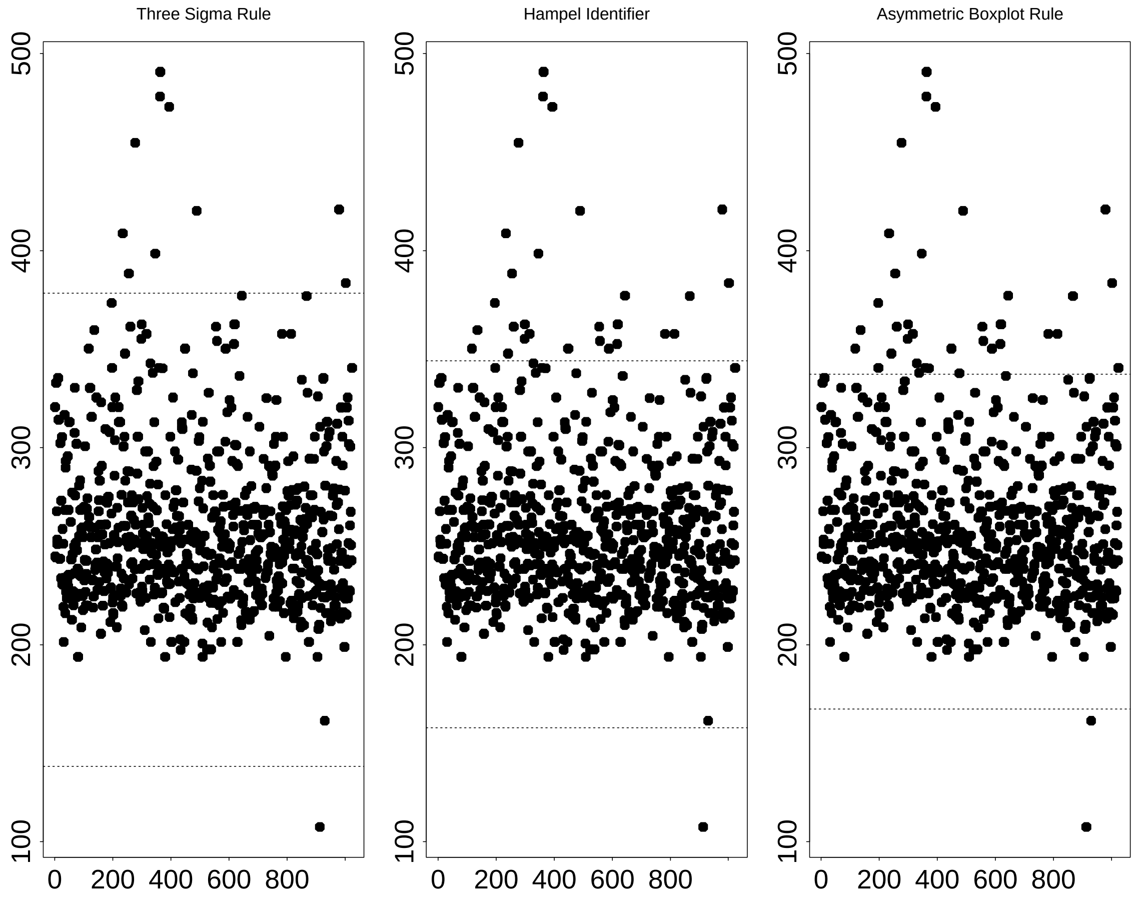
- Symmetric version:
  - like Hampel identifier, replace $\bar{x}$ with $x^\dagger$
  - replace $\hat{\sigma}$ with the outlier-resistant interquartile distance Q
- Quartiles:
  - x_U = upper quartile $\Rightarrow$ 75% of data values lie below this observation
  - x_L = lower quartile $\Rightarrow$ 25% of data values lie below this observation
  - $Q = x_U - x_L$

$\rightsquigarrow$ Asymmetric version:

  - $x_k < x_L - tQ \Rightarrow$ lower outlier
  - $x_k > x_U + tQ \Rightarrow$ upper outlier

Asymmetric Example:
The industrial pressure dataset

Comparison of three outlier detection rules



Mis . . . ing Data

- Problem: some x_k values are unavailable
  - ignorable case: increases variability
  - $\rightsquigarrow$ nonignorable case: introduces bias
  - $\rightarrow$ 1936 *Literary Digest* election poll
- Autocorrelation example:

$$\tilde{R}_{xx}(k) = \frac{1}{|S|} \sum_{\ell \in S} x_\ell x_{\ell+k}$$

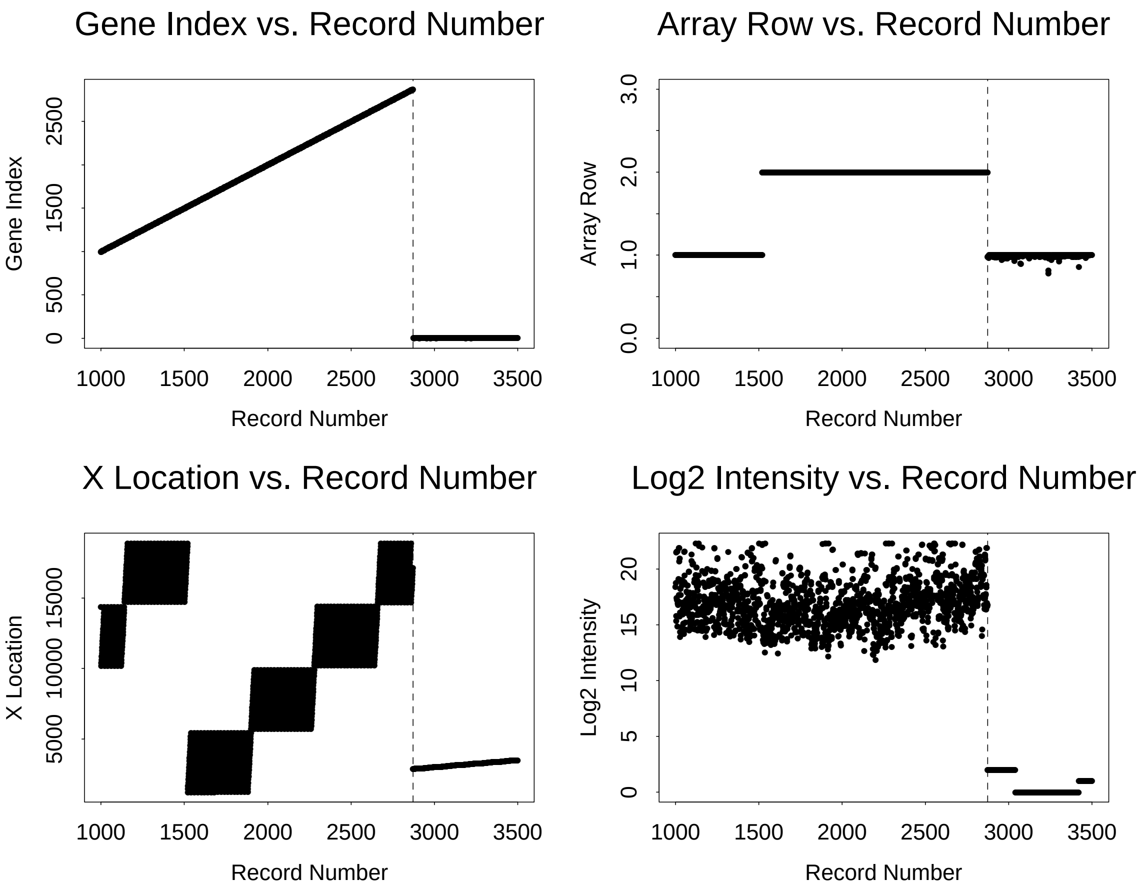
  - $S =$ random subset of $\{1, 2, \ldots, N\} \Rightarrow$ ignorable case: causes increased variability of $R_{xx}(k)$ estimates
  - $S =$ even k only $\Rightarrow$ non-ignorable case: cannot estimate $R_{xx}(k)$ for any odd k
- Additional consequences:
  - missing values can be converted into outliers (storage tank example)
  - missing values can cause misalignments

Misalignment:

Four corrupted data sequences caused by unexpected "blank" records

⇝ **NOTE:**

Difficulty of detection varies strongly from one variable to another



The CAMDA Challenge Dataset

- CAMDA: Critical Assessment of Microarray Data Analysis
  - annual data analysis competition
  - CAMDA 2002 challenge datsets:
    1. Latin square Affymetrix benchmark
    2. normal mouse cDNA microarray study
- Structure of the normal mouse dataset:
  - derived from 72 individual microarrays
  - 3 organ samples from each of 6 mice
  - 4 microarrays per sample
  - 2 channels per microarray: reference & experimental

  ⇝ reformated into three organ-specific summary datasets

The CAMDA Challenge Dataset

- Stivers *et al.* obtained anomalous results from a preliminary principal components analysis
  - expected clustering: common reference cluster, 3 organ clusters
  - observed: unreasonable splitting of the reference cluster
  - subsequently observed: disagreements of gene ID/slide position combinations between different organ datasets
- What happened?
  - 1932 of 5304 genes were mis-annoted
  - cause: error in procedure that combined the 72 individual microarray datasets into 3 organ-specific summary datasets

Softwear Errors

- Source of both misalignment examples:
  1. inconsistent handling of missing values between Excel and S-plus
  2. (Stivers et al.):

     > The data used here were assembled into packages, probably manually using *ad hoc* database, spreadsheet, or perl script. Under these conditions, it is remarkably easy for the row order to be changed accidentally . . .

- Some relevant observations:
  1. Wall, *et al.* (2000):

     > It is a standing joke in the *Perl* community that the next big stock market crash will probably be caused by a bug in someone's *Perl* script.

  2. Kanert *et al.* (1999):

     > About one in three attempts to fix a program doesn't work or causes a new problem.

  3. Beizer (1990):

     > estimates between 1 and 3 errors per 100 executable statements, *after the code has been debugged*

Noninformative Variables

- Externally noninformative variables:
  - variables x_k that are *a priori* irrelevant
  - ⇝ **Murphy's law:** irrelevant variables sometimes aren't
  - R.W. McClure's example
- Inherently noninformative variables:
  - completely missing variables
  - constant variables
  - exact duplicate variables
- Application-irrelevant variables:
  - e.g., variables that become inherently noninformative when analysis is restricted to a subset of interest
  - specific example: anomaly indicator variables in the analysis of nominal data
  - (sometimes:) noise variables

⇝ Why is this important?

A Clustering Example

- Eight datasets compared:
  - $k = 4$ well-separated clusters
  - three informative components in each attribute vector $\mathbf{x}_k$
  - 0 to 7 non-informative components in $\mathbf{x}_k$
- Clustering procedure:
  - Partitioning Around Medoids (PAM) - Kaufman and Rousseeuw (1987)
  - better-behaved alternative to k-means
- Performance assessment:
  - average silhouette coefficient (Kaufman and Rousseeuw, 1987)
  - assesses both intracluster cohesion and intercluster separation
  - bounded between -1 (horrible misclassification) and $+1$ (perfect classification)

Clustering Results:

Influence of noninformative variables

Average silhouette coefficients $\bar{s}$
$k = 2 \Rightarrow$ *spurious clustering*
$k = 4 \Rightarrow$ *correct clustering*

Noise Components	$\bar{s}$, $k = 2$	$\bar{s}$, $k = 4$
0	0.636	0.750
1	0.619	0.709
2	0.604	0.675
3	0.587	0.638
4	0.579	0.619
5	0.568	0.595
6	0.557	0.573
7	0.548	0.555

A Final Example

- Consider the effects of "small" deletions:
  - datasets: four different 17 point sequences
  - deletions: all possible 2 point deletions

  $\Rightarrow \begin{pmatrix} 17 \\ 2 \end{pmatrix} = 136$ possible 15 point subsets

- The data sequences:

  0: uniformly distributed on $[-1.1, 1.1]$

  1: 8 points uniformly distributed on $[-1.1, -0.9]$, one zero value, 8 points uniformly distributed on $[0.9, 1.1]$

  2: middle 5 points of Sequence 0 set to zero (one common missing data model)

  3: Sequence 0 with 2 outliers, rescaled into original $[-1.1, 1.1]$ range

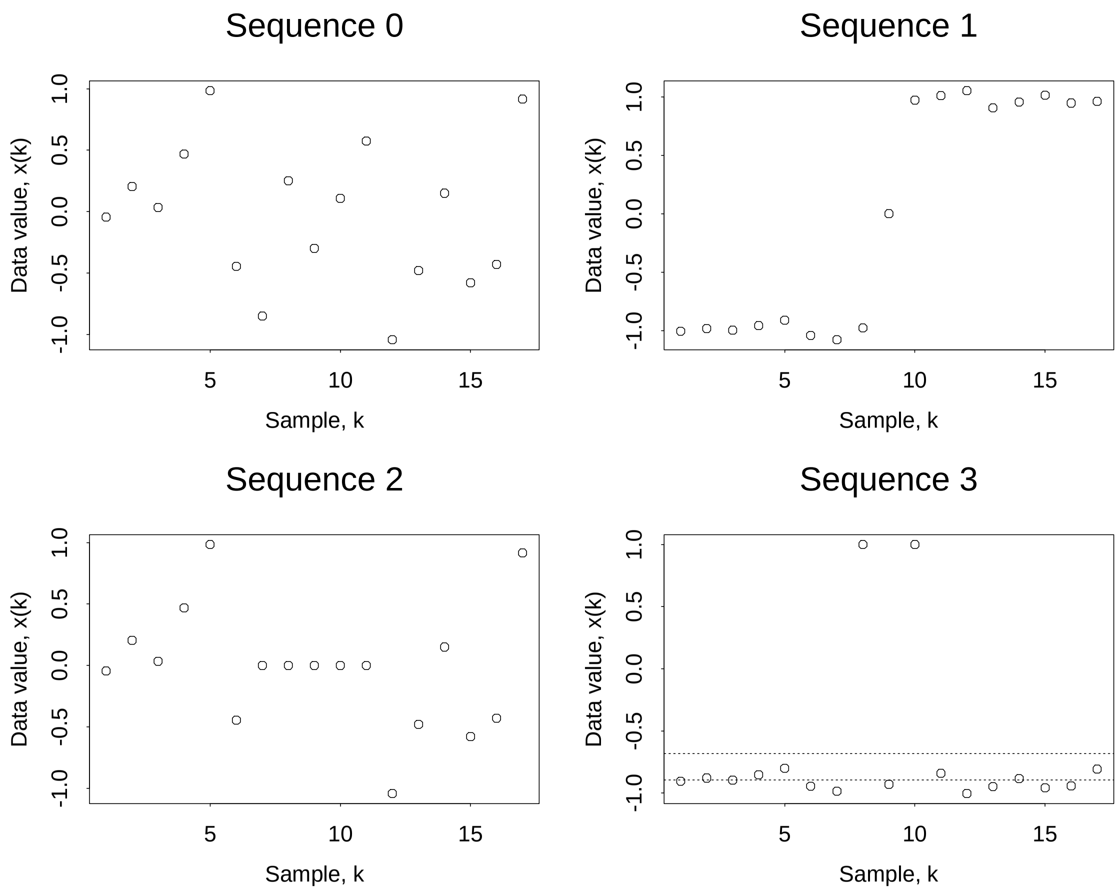
- The scale estimates:

  A. the standard deviation $\hat{\sigma}$

  B. the interquartile distance Q

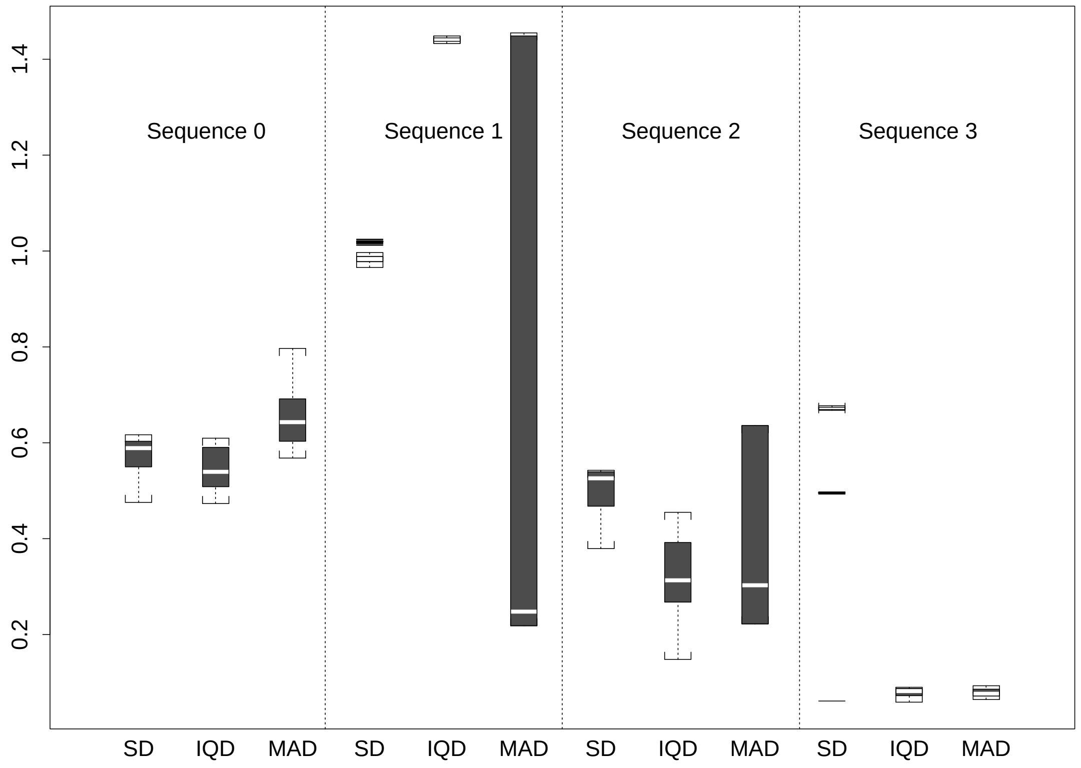
  C. the MAD scale estimate S

Four Simulated Data Sequences



Scale Estimates:

Consequences of all possible 2-point deletions



Summary:

Three Key Conclusions

1. Unimaginable anomalies infest real datasets

⇝ Yogi Bera:

> If something has a 50% chance of happening,
> then 9 times out of 10 it will.

⇝ Dasu and Johnson (2003, p. 186):

> Take NOTHING for granted. The data are never what they are supposed to be, even after they are "cleaned up." The schemas, layout, content, and nature of content are never completely known or documented and continue to change dynamically.

2. Different analysis methods exhibit different sensitivities to different data anomalies

3. Comparison of what *should* be "equivalent" analyses across different scenarios can be extremely useful in uncovering anomalies