Introduction to Robust Data Analysis

Outline

- What is robust data analysis, why is it useful
- Methods
  - median
  - median absolute deviation
  - trimmed estimators
  - Windsor estimators
  - (M-estimators)
- Using R
- Using Python
- Using KNIME basic features
- (Using KNIME reporting features)
- Using HiTS for special needs
- Off topic: data transformations

What is robust data analysis?

http://en.wikipedia.org/wiki/Robust_statistics
A **robust statistic** is resistant to errors in the results, produced by deviations from assumptions
If the assumptions are only approximately met, the robust estimator will still have a reasonable efficiency, and reasonably small bias, as well as being asymptotically unbiased

Why robust statistics?

The outliers have smaller or no effect on the statistics
They eliminate/hide some of the errors

Methods - median

This is an alternative to mean/average.
When the values are ordered this is the middle element (or the average of the two middle element).

Methods - median absolute deviation

This is an alternative to the standard deviation.
$\text{median}(|x_i - \text{median}(x_i)|)$

Methods - trimmed estimators

Removing some of the outliers and then computing the mean/standard deviation/range or other statistics.
Special case:

- Interquartile range - 25% trimmed range.
  - for a normal-consistent estimate of the standard deviation, use IQR / 1.349

Methods - Windsor estimators

These are similar to the trimmed estimators, but the extreme values are not omitted, but replaced by less extreme values.

Using R

Really powerful.
Functions:

- median
- mad
- mean
- sd.trim from Gregor Gorjanc (similarly Windsor estimators)
- IQR

Using Python

Allows multiple column computation, although it does not support (without additional libraries) the computation of robust statistics.
Not covered in this presentation.

Using KNIME basic features

Conditional Box Plot computes the following robust statistics:

- median
- IQR (computable: `Q3-Q1`)

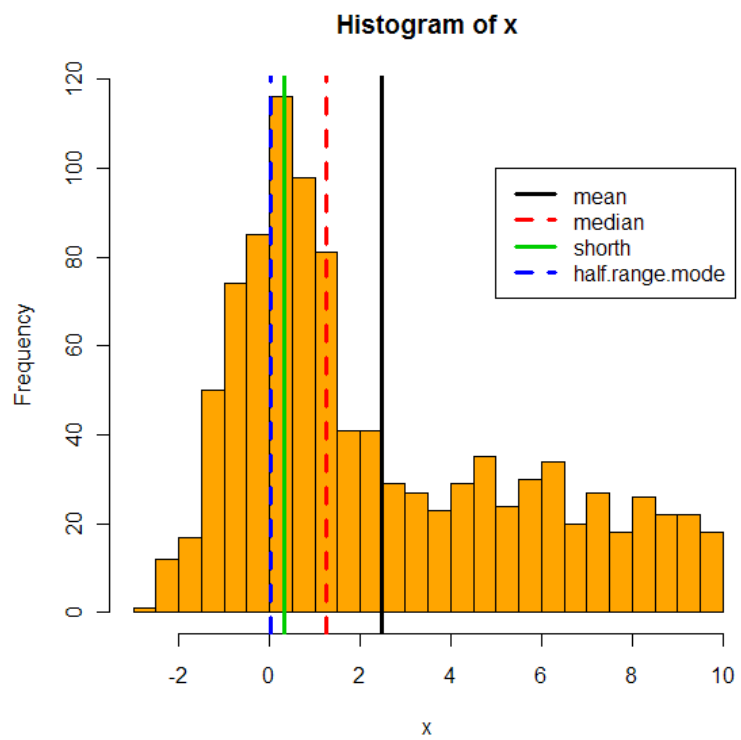
Sorter, ***Row Filter*** and ***Statatistics*** or ***GroupBy***:

- trimmed mean
- trimmed standard deviation

Using HiTS

HiTS was designed for a really special area of data analysis, High Content/Throughput Screens.
The data is expected to have an experimental structure, called plate. CellHTS2 is used to perform analysis.
Allows normalisation to controls, which is a bit tricky with basic KNIME nodes.
Supported methods:

- median,
- mean,
- POC,
- NPI,
- shorth,
- Bscore.



Data transformations

Subsets - generate every possible subset of a set of values
Direct product - direct product of two columns
Merge - kind of anti-sort
Pivot - from rows to columns
Unpivot - from columns to rows

Hierarchical Clustering

Some useful additions:

- Reverse Order
- Leaf Ordering
- Sort by Cluster

Acknowledgements

Supported by a Marie Curie Fellowship in Trinity College Dublin.
Thanks for

- Aideen Long,
- Dara J. Dunican,
- Emliy Bennet,
- Antje Hoff,
- Michael Freeley, and all other users.

evopro Kft - workshop

Questions?

http://code.google.com/p/hits

http://drop.io/HiTS_sample_workflows

Thank you for your attention.