A casual exploration of outliers in small sample situations

Dhammika Amaratunga

Senior Director & Janssen Fellow, Nonclinical Statistics & Computing
Janssen R&D of Johnson & Johnson

Osmosis, Beerse, May 2013

Introduction

Literature

- ***Conflicting statements …***
  - **"The t-test is robust, so it is not drastically affected by departures from the underlying assumptions."**
  - **"Even small deviations from a normal distribution could cause classical parametric statistics to fail."**
  - **"Nonparametric tests are extremely robust to outliers because they use only the ranks of the values instead of the values themselves"**
  - **"Even non-parametric analyses may suffer from outlier cases."**

Outliers

- **Outliers are "unusual" observations; observations that are inconsistent with the general tendency or variability of the data.**
- **They are potentially problematic because they could significantly impact the analysis and interpretation of the data.**

Question: How should outliers be handled?

Example

Group 1: 19 21 20 16 27 27 32 24 24 18

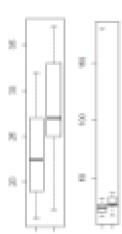
Group 2: 26 26 36 32 24 17 33 28 37 25

t test p-value = 0.037*

Group 1: 19 21 20 16 27 27 32 24 24 180

Group 2: 26 26 36 32 24 17 33 28 37 25

t test p-value = 0.512



Example

Group 1: 19 21 20 16 27 27 32 24 24 18

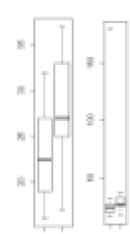
Group 2: 26 26 36 32 24 17 33 28 37 25

Wilcoxon test p-value = 0.049*

Group 1: 19 21 20 16 27 27 32 24 24 180

Group 2: 26 26 36 32 24 17 33 28 37 25

Wilcoxon test p-value = 0.198



Example

Group 1: 19 21 20 16 27 27 32 24 24 18

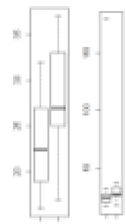
Group 2: 26 26 36 32 24 17 33 28 37 25

Robust t test p-value = 0.020*

Group 1: 19 21 20 16 27 27 32 24 24 180

Group 2: 26 26 36 32 24 17 33 28 37 25

Robust t test p-value = 0.034*



Observations

♦ **Outliers are indeed a problem! The t test and the Wilcoxon test were both affected by the outlier.**

♦ **The robust t-test (which used M-estimates with biweight weights) was not much affected by the outlier.**

Outlier detection

Examples

♦ *Example 1*	♦ *Example 2*	♦ *Example 3*
Data:	Data:	Data:
17	17	17
21	21	21
25	25	25
26	26	26
32	32	32
40	40	400
48	480	480
Mean = 29.9	Mean = 91.6	Mean = 143.0
SD = 10.9	SD = 171.4	SD = 204.3
Median = 26.0	Median = 26.0	Median = 26.0
MAD = 8.9	MAD = 8.9	MAD = 8.9

Outlier detection methods (1)

♦ **A general approach:**

Check how far the values are from the center of the distribution of values observed:

$$Distance = \frac{|Value - Center|}{Variability}$$

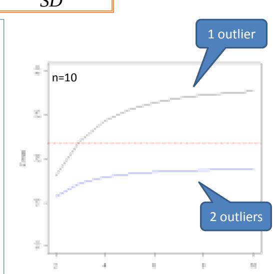
- **Values close to the center will have relatively small distances while values at the extremes will have relatively large distances.**
- **Outliers are identified as observations that have unusually large distances.**

Outlier detection methods (2)

♦ **A classical approach – Grubb's test:**

$$Distance = \frac{|Value - Mean|}{SD}$$

♦ *Ex 1*	♦ *Ex 2*	♦ *Ex 3*
Data:	Data:	Data:
17	17	17
21	21	21
25	25	25
26	26	26
32	32	32
40	40	400
48	480*	480
p=0.22	*p*<0.001	*p*=0.22

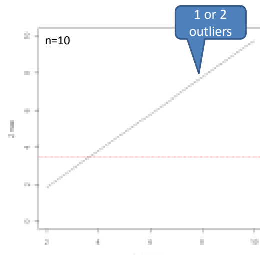


Outlier detection methods (3)

♦ **An improved approach – Median-MAD test:**

$$Distance = \frac{|Value - Median|}{MAD}$$

♦ *Ex 1*	♦ *Ex 2*	♦ *Ex 3*
Data:	Data:	Data:
17	17	17
21	21	21
25	25	25
26	26	26
32	32	32
40	40	400*
48	480*	480*



Construction of outlier detection tests

♦ **Outlier detection tests should really be based on outlier-resistant statistics (i.e., statistics that have high breakdown).**

♦ **Breakdown refers to the minimum percentage of observations that can be moved to infinity without the statistic also moving off to infinity.**

♦ Example: Mean and SD have BP=0%.

♦ Example: Median and MAD have BP≅50%.

➢ Median = middle observation when X is sorted in order.

➢ MAD (median absolute deviation) = median of |X-Median|.

Outlier detection using resistant statistics

♦ Median-MAD method: **An observation is considered an outlier if it lies outside the interval:**

$$\text{MEDIAN} \pm \lambda\text{MAD}.$$

➢ **BP of fence is 50%.**
➢ **Fence is symmetric.**

♦ Boxplot method: **An observation is considered an outlier if it lies outside the interval:**

$$(Q_L - \lambda IQR,\ Q_U + \lambda IQR).$$

➢ **BP of fence is 25%.**
➢ **Fence is asymmetric.**

Performance of some outlier detection tests

♦ **The table below gives the probability of Grubb's Mean-SD test and the Median-MAD test (with cutoff of 3.5) finding at least one outlier in a sample when the "good" data is N(0,1) and the outliers are at 5.**

	-- N=5 --			-- N=10 --			-- N=20 --		
#Outliers	0	1	2	0	1	2	0	1	2
Grubb's	10	59	0	10	95	0	10	99	32
Med-MAD	20	63	21	12	80	57	10	91	83
Hybrid	8	39	43	10	85	64	12	98	94

♦ **Grubb's test is slightly better for the cases where there is at most one outlier but overwhelmingly weaker for the cases where there are multiple outliers.**

♦ **Personal preference: Median-MAD test (or hybrid!).**

Robust statistics

Robustness

♦ **Instead of trying to identify and remove outliers, robust statistics attempts to accommodate outliers, usually by downweighting observations which lie far from the center.**

♦ **Robust statistics are constructed to have high efficiency at both the nominal as well as contaminated situations.**

M estimates of location

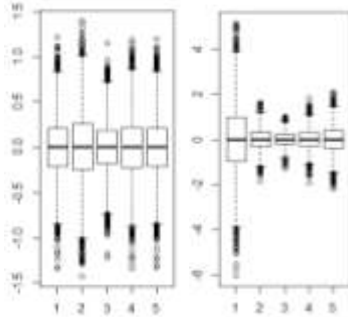
- Can think of M estimates as solutions to

  $\mu = (\sum w_i x_i)/(\sum w_i)$ where $w_i = W(x_i - \mu)$

- This has to be solved iteratively for μ starting from an outlier resistant (but possibly inefficient) estimate for μ (sometimes 1-step estimators are also effective).
- Weight functions:
  - Arithmetic mean: $W(u) = 1$
  - Huber: $W(u) = \min(1, k/|u|)$
  - Biweight: $W(u) = (1-(u/k)^2)^2 \, I(|u|<k)$
- Robust weight functions reduce the influence of non-central observations, with the biweight especially strongly downweighting extreme values.

Performance of some estimates of location

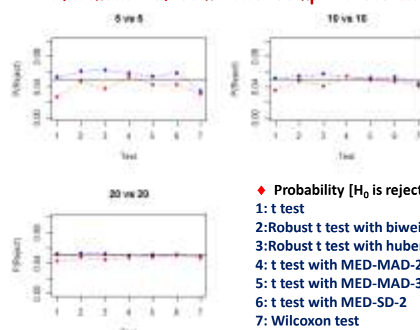
- Sampling distribution of estimates when N=5 for:
  1: Mean
  2: Median
  3: Biweight (1-step)
  4: Biweight
  5: Huber
- **Personal preference**: Biweight mean



Normal Contaminated Normal

Hypothesis testing in the presence of outliers

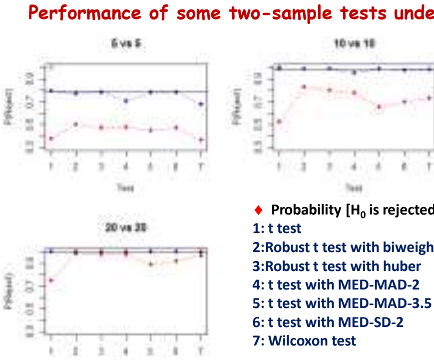
Performance of some two-sample tests under H_0



- Probability [H_0 is rejected at 5%|Δ=0]
  1: t test
  2:Robust t test with biweight
  3:Robust t test with huber
  4: t test with MED-MAD-2
  5: t test with MED-MAD-3.5
  6: t test with MED-SD-2
  7: Wilcoxon test

(- Normal / - Contaminated Normal)

Performance of some two-sample tests under H_1



- Probability [H_0 is rejected at 5%|Δ=2]
  1: t test
  2:Robust t test with biweight
  3:Robust t test with huber
  4: t test with MED-MAD-2
  5: t test with MED-MAD-3.5
  6: t test with MED-SD-2
  7: Wilcoxon test

(- Normal / - Contaminated Normal)

Points to note – parametric tests

- When there are outliers in the data

  Outliers $\Rightarrow \uparrow SE$ (compared to signal) $\Rightarrow \downarrow t \Rightarrow \downarrow$ reject H_0
  $\Rightarrow \downarrow Size \;\&\; \downarrow Power$

- Loss of Size makes the t test more conservative (some may feel this is ok in confirmatory settings)
- Loss of Power makes it more difficult to detect signal (can be damaging in exploratory settings)
- Robust alternatives (including simple outlier removal) don't increase Size much, but do increase Power

Points to note – nonparametric tests

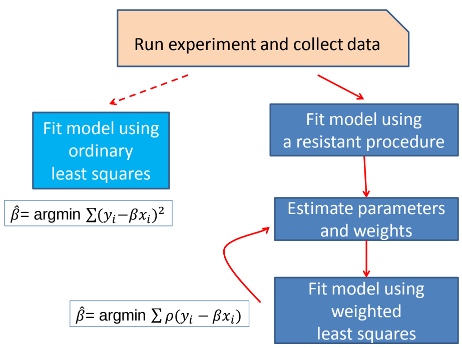
- When there are outliers in the data, nonparametric tests such as the Wilcoxon test are robust of validity (they maintain their Size at α), but, in small sample situations, they are not necessarily robust of efficiency (they tend to lose power).

- The same applies to permutation tests which are not based on a robust test statistic.

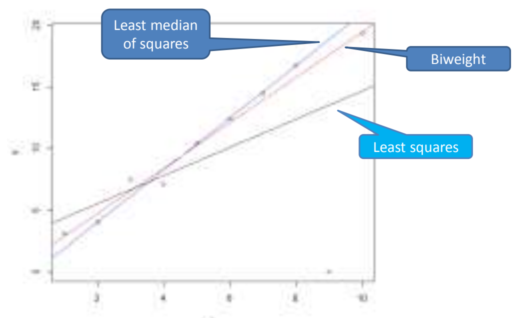
- <u>Personal preference</u>: Robust test.

Model fitting in the presence of outliers

Analysis procedure – schematic

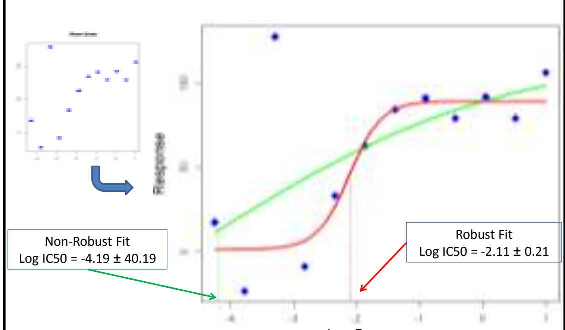


Example – linear model



Example: nonlinear model

- Inhibition Effect on IgG1



Wrap Up

Software

♦ R packages and functions

♦ Some outlier-handling methods will be implemented in EZ-R4Excel, which provides an Excel interface to R:
- Median-MAD test
- Robust t test and F test
- Robust 4 parameter logistic model fit
- Other methods (based on need)

(contact: Jyotsna Kasturi)

♦ TransData is a SAS based system that performs a robust transformation of the data.
(contact: Al Barron)

Wrap up

♦ **Conclusion:** There is little lost and much gained by using methods that work well in a range of situations rather than methods that are optimal in a specific situation which may or may not be the situation for the data being analyzed.

♦ **Tukey:** "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

♦ Website:
www. amaratunga.com

♦ Email:
damaratu@its.jnj.com