Method Comparison Studies

Kevin O'Brien (kevin.obrien@ul.ie, University of Limerick)

Medical statistics

- Applications to medicine and the health sciences, including epidemiology, public health, forensic medicine, and clinical research.
- "Biostatistics" more commonly connotes all applications of statistics to biology.
- Clinical Research is main focus for this talk - Method Comparison Studies

Medical Statistics

"It is the science of summarizing, collecting, presenting and interpreting data in medical practice, and using them to estimate the magnitude of associations and test hypotheses.
It has a central role in medical investigations. It not only provides a way of organizing information on a wider and more formal basis than relying on the exchange of anecdotes and personal experience, but also takes into account the intrinsic variation inherent in most biological processes."

(Kirkwood, Betty R. (2003). "Essential Medical Statistics")

Pharmaceutical statistics

Pharmaceutical statistics is the application of statistics to matters concerning the pharmaceutical industry.

This can be from issues of design of experiments, to analysis of drug trials, to issues of commercialization of a medicine.

There are many professional bodies concerned with this field including:

- European Federation of Statisticians in the Pharmaceutical Industry (EFSPI)
- Statisticians In The Pharmaceutical Industry (PSI)

Medical Measurement

Method Comparison Studies

- Commonly encountered issue in medical statistics
- “Do two methods of measurement agree statistically?".
- “Can the two methods be used interchangeably?"
- Sources of disagreement can arise from differing population means (i.e. inter-method bias), differing between-subject and with-in subject variances [1].

Gold Standards

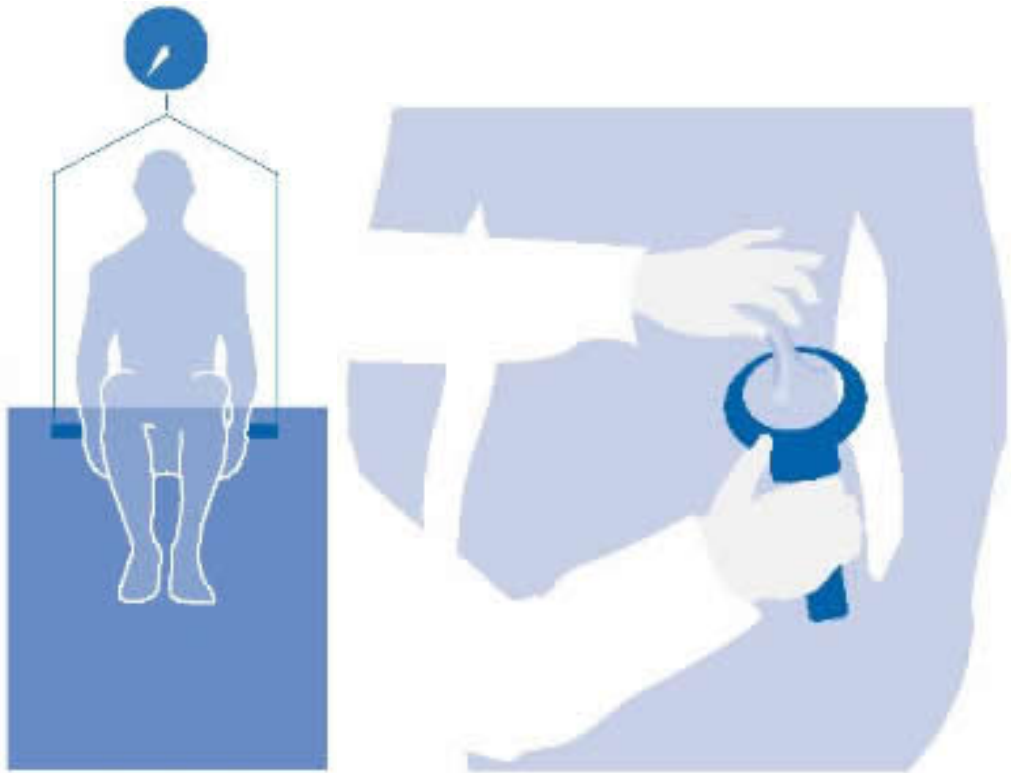
Gold Standard Methods of Measurement

- Gold standard test usually refers to a diagnostic test or benchmark that is the best available under reasonable conditions.
- Other times, gold standard is used to refer to the most accurate test possible without restrictions.

For instance, for the diagnosis of aortic dissection, the "gold standard" test used to be the aortogram, which had a sensitivity as low as 83% and a specificity as low as 87%.
Since the advancements of magnetic resonance imaging, the magnetic resonance angiogram (MRA) has become the new "gold standard" test for aortic dissection, with a sensitivity of 95% and a specificity of 92%.
Before widespread acceptance of any new test, the former test retains its status as the "gold standard."

Methods of Measurement

Comparing against a Gold Standard



December 18, 2012 | By Ioana Patringenaru

Small, Portable Sensors Allow Users to Monitor Exposure to Pollution on Their Smart Phones

Computer scientists at the University of California, San Diego have built a small fleet of portable pollution sensors that allow users to monitor air quality in real time on their smart phones. The sensors could be particularly useful to people suffering from chronic conditions, such as asthma, who need to avoid exposure to pollutants.

CitiSense is the only air-quality monitoring system capable of delivering real-time data to users' cell

The CitiSense sensors transmit their air quality readings to smart phones. More pictures of the sensor and its smart phone interface can be found here

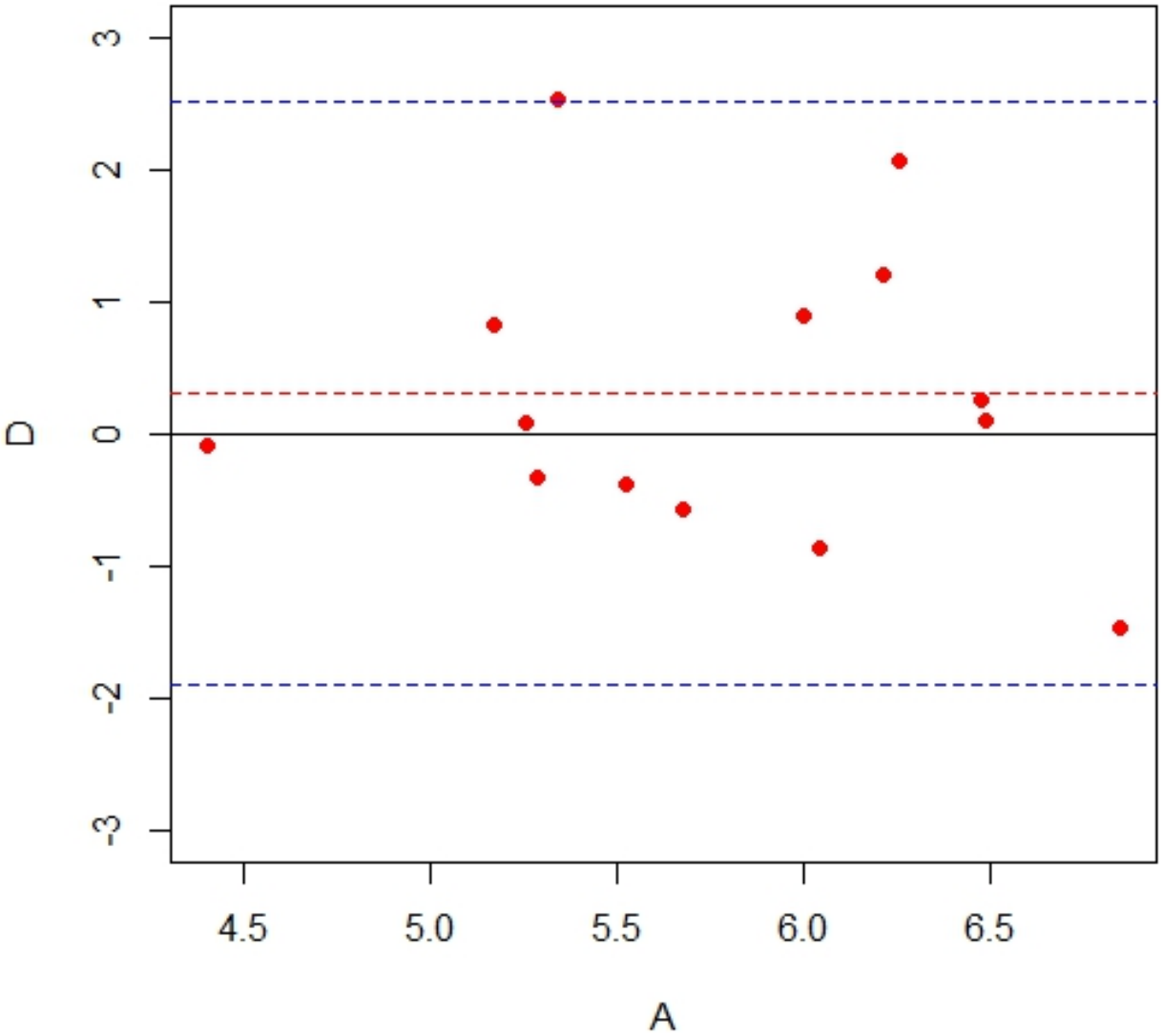
The Bland-Altman Plot

- The Bland-Altman plot [2, 3] is a very simple graphical method to compare two measurements techniques.
- In this approach the case-wise differences between the two methods are plotted against the corresponding case-wise averages of the two methods.
- A horizontal lines is drawn at the mean difference(the **inter-method bias**) , and at the **limits of agreement**, which are defined as the inter-method bias plus and minus 2 times the standard deviation of the differences.

Bland-Altman Plot

```
>X = rnorm(14,6,1);Y = rnorm(14,5.3,1.1)
>
>A=(X+Y)/2 #case-wise averages
>D=X-Y     #case-wise differences
>
>Dbar=mean(D) #inter-method bias
>SdD=sd(D) #standard deviation of the differences
>
>plot(A,D,pch=16,col="red", ylim=c(-3,3))
>
>abline(h=Dbar,lty=2)
>abline(h=(Dbar-2*SdD),lty=2)
>abline(h=(Dbar+2*SdD),lty=2)
```

Inter-method Bias : 0.27 | Limits of Agreement: [-1.98, 2.52]



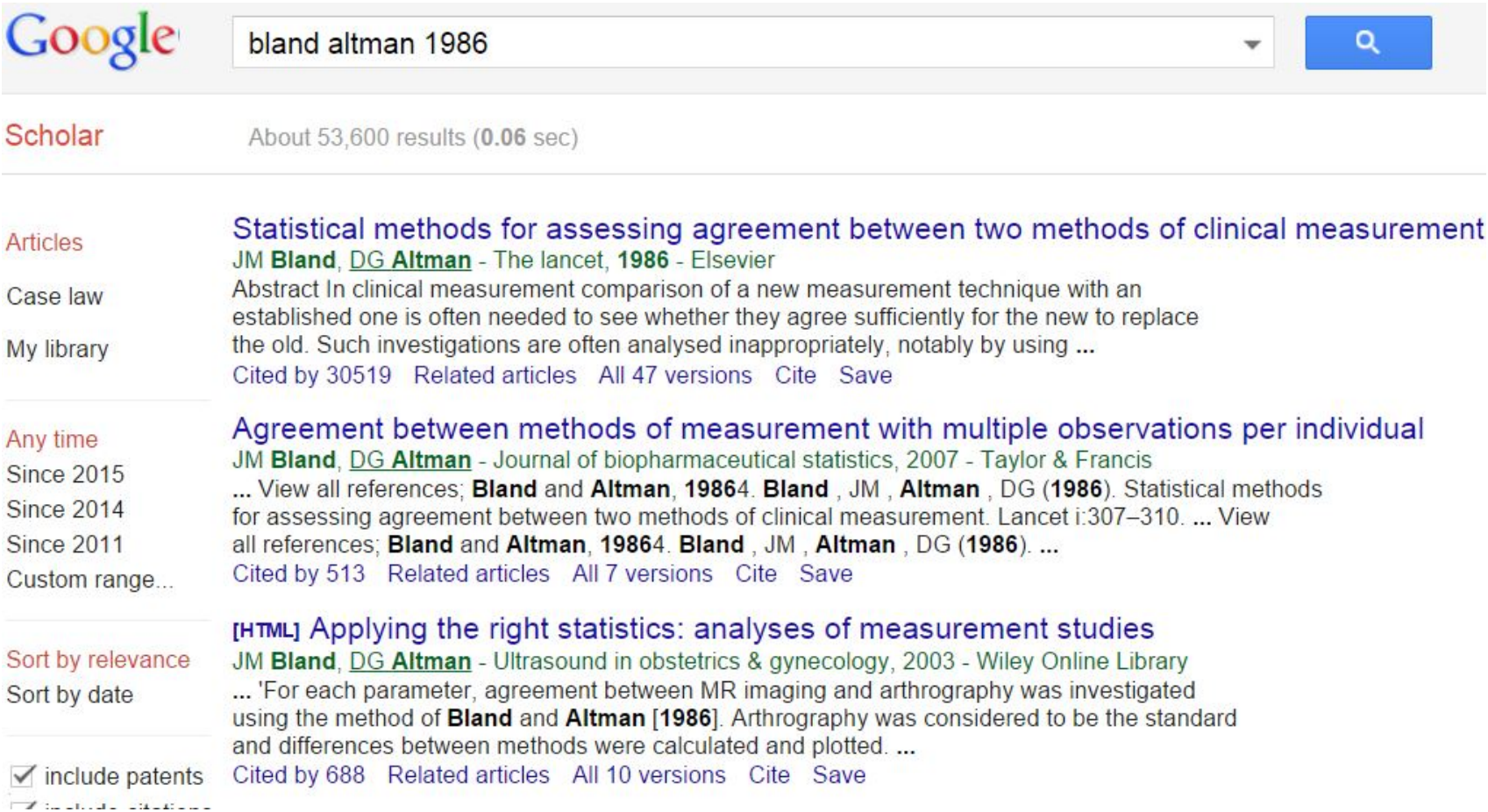
Bland-Altman Plot

Building Blocks

1. Simple Arithmetic Operations
2. Sample Mean - `mean()`
3. Sample Standard deviation - `sd()`
4. Scatter plot - `plot()`
5. Normal Distribution Theory
6. Enhancing plots - basic `R` knowledge

Nothing here that is beyond a Stats 101 course in college.

In excess of 30000 citations



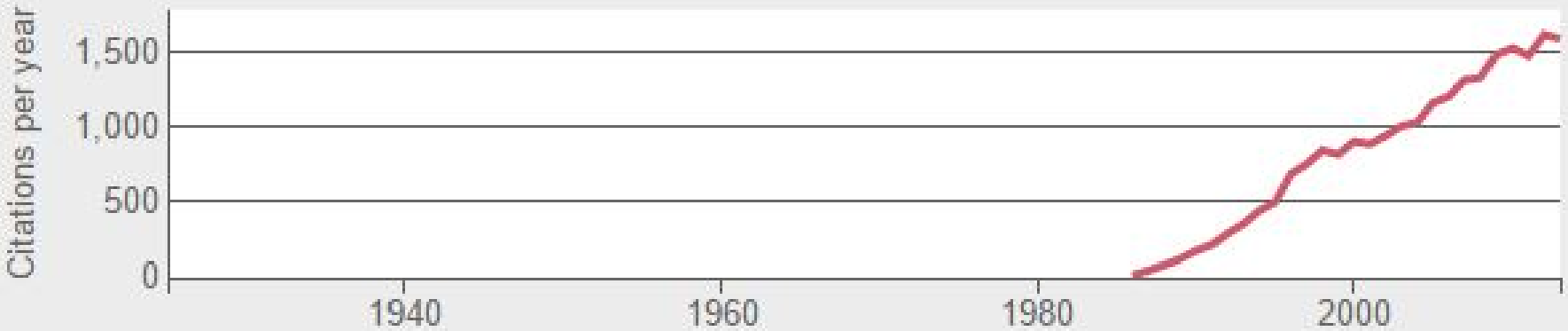

nature
International weekly journal of science
Home
News & Comment
Research
Careers & Jobs
Current Issue
Archive
Audio & Video
For Authors
Archive
Volume 514
Issue 7524
News Feature
Article
NATURE | NEWS FEATURE
عربي
The top 100 papers
Nature explores the most-cited research of all time.
Richard Van Noorden, Brendan Maher & Regina Nuzzo
29 October 2014

Rank: **29** Citations: **23,826**

Statistical methods for assessing agreement between two methods of clinical measurement.

Bland, J. M. & Altman, D. G.

Lancet **327**, 307–310 (1986).



> The Kaplan–Meier paper was a sleeper hit, receiving almost no citations until computing power boomed in the 1970s, making the methods accessible to non-specialists. Simplicity and ease of use also boosted the popularity of papers in this field. British statisticians Martin Bland and Douglas Altman made the list (number 29) with a technique[17] — now known as the Bland–Altman plot — for visualizing how well two measurement methods agree. The same idea had been introduced by another statistician 14 years earlier, but Bland and Altman presented it in an accessible way that has won citations ever since.

R Packages for Bland-Altman Analysis

PairedData has a function `plotBA` based on ggplot2 and no stats as return value

ResearchMethods has a function `BlandAltman` which focuses on a GUI and has no return values.

epade has a function `bland.altman.ade` which appears to have no return values.

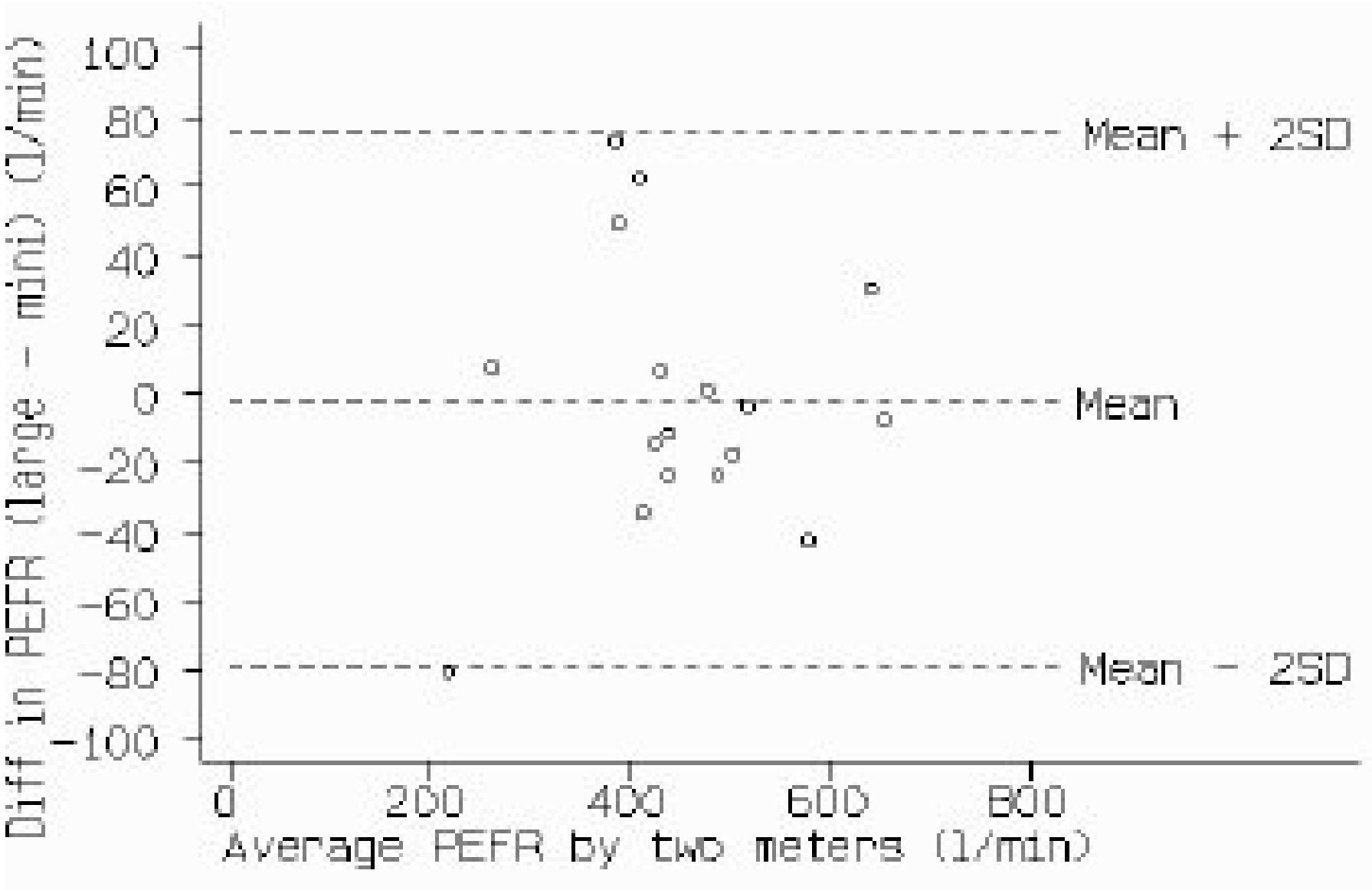
MethComp has a functino BlandAltman that is deprecated and a function `ba.plot` which does a lot, mainly regression methods

MethComp: Functions for Analysis of Agreement in Method Comparison Studies

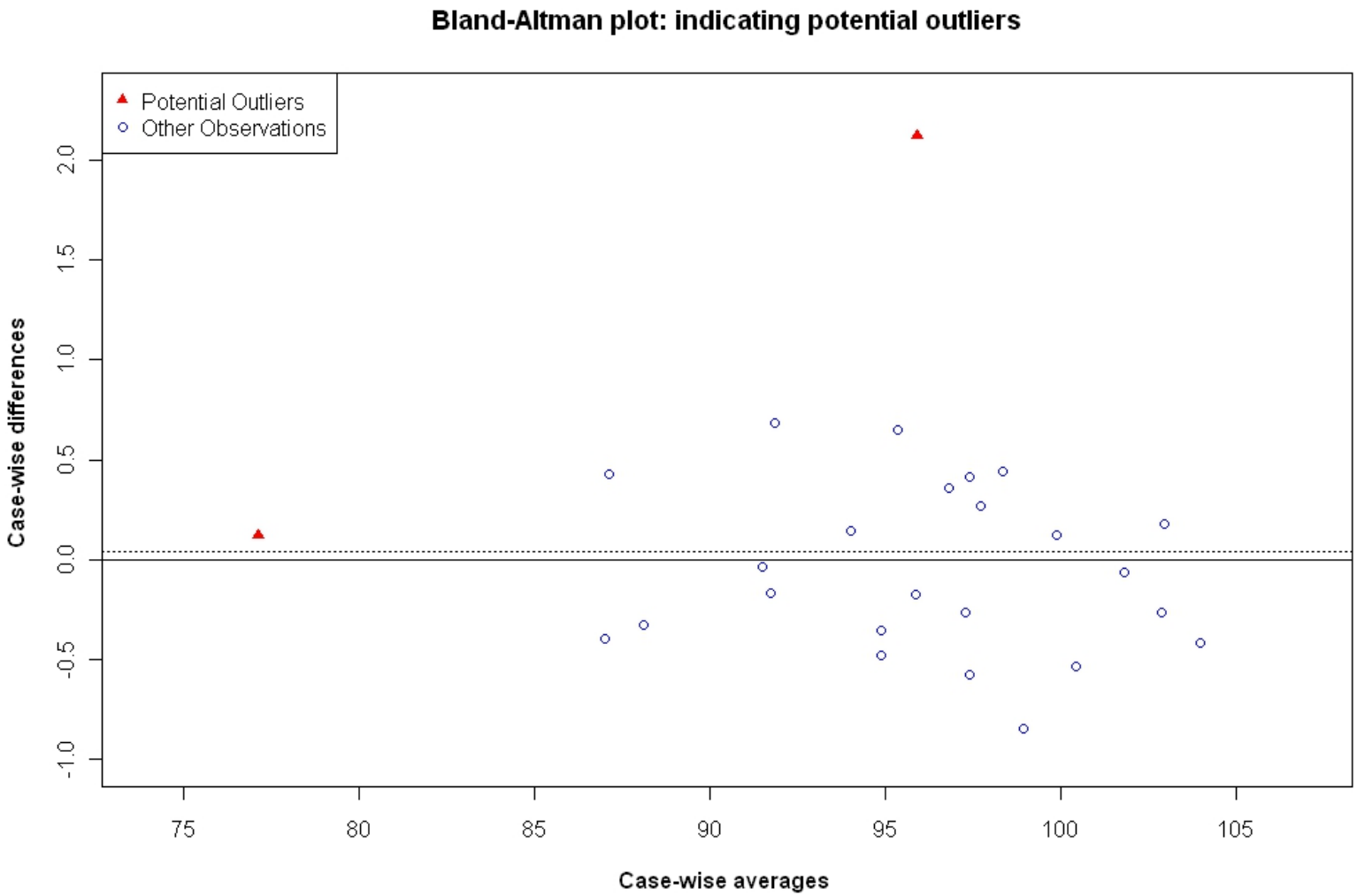
Methods (standard and advanced) for analysis of agreement between measurement methods.

Version:	1.22.2
Depends:	R (≥ 3.0.0), nlme
Suggests:	R2WinBUGS, BRugs, rjags, coda, lattice, lme4
Published:	2015-03-31
Author:	Bendix Carstensen, Lyle Gurrin, Claus Ekstrom, Michal Figurski
Maintainer:	Bendix Carstensen <bxc at steno.dk>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	http://BendixCarstensen.com/MethComp/
NeedsCompilation:	no
CRAN checks:	MethComp results

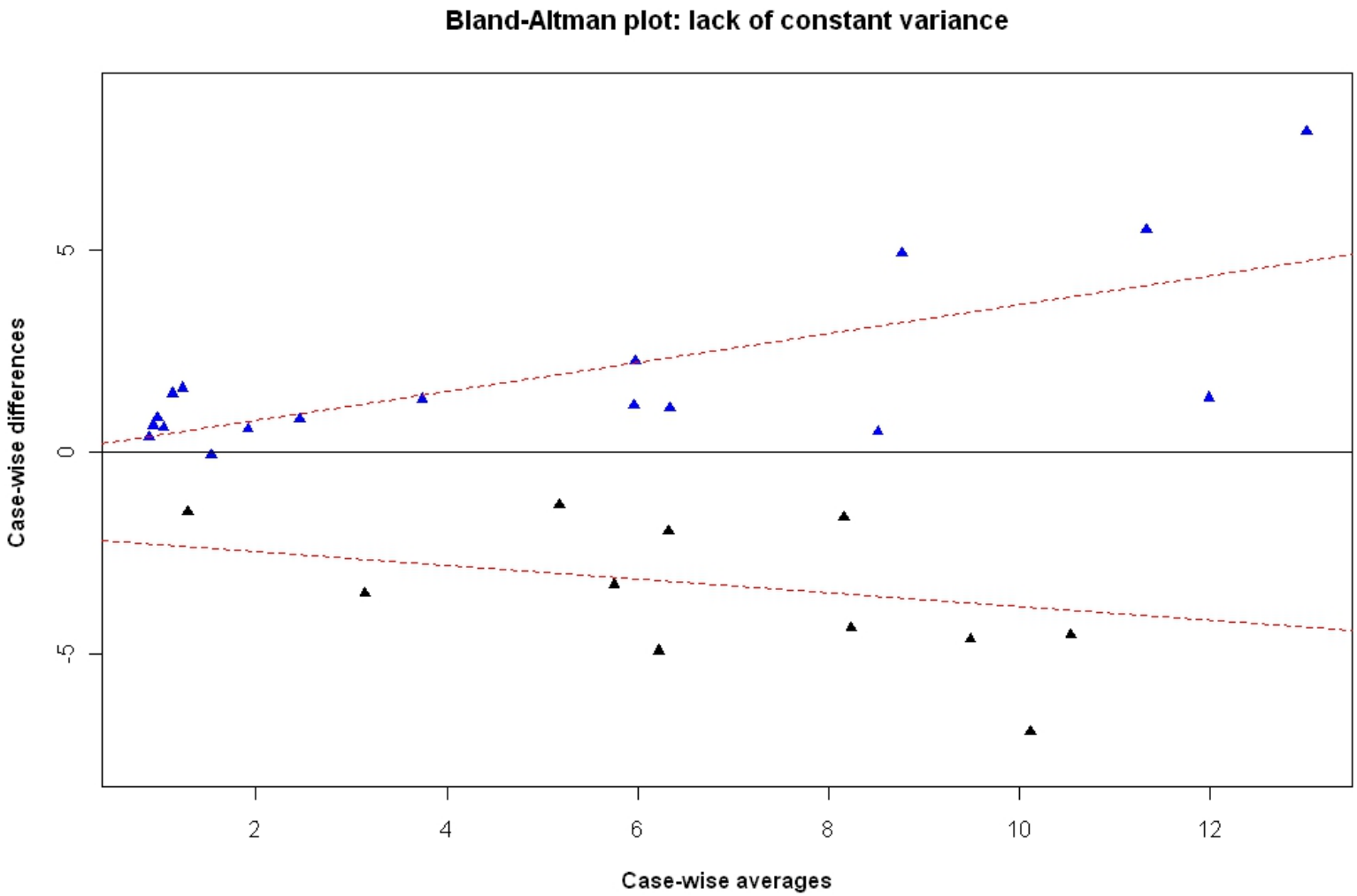
Interpreting the Bland-Altman Plot



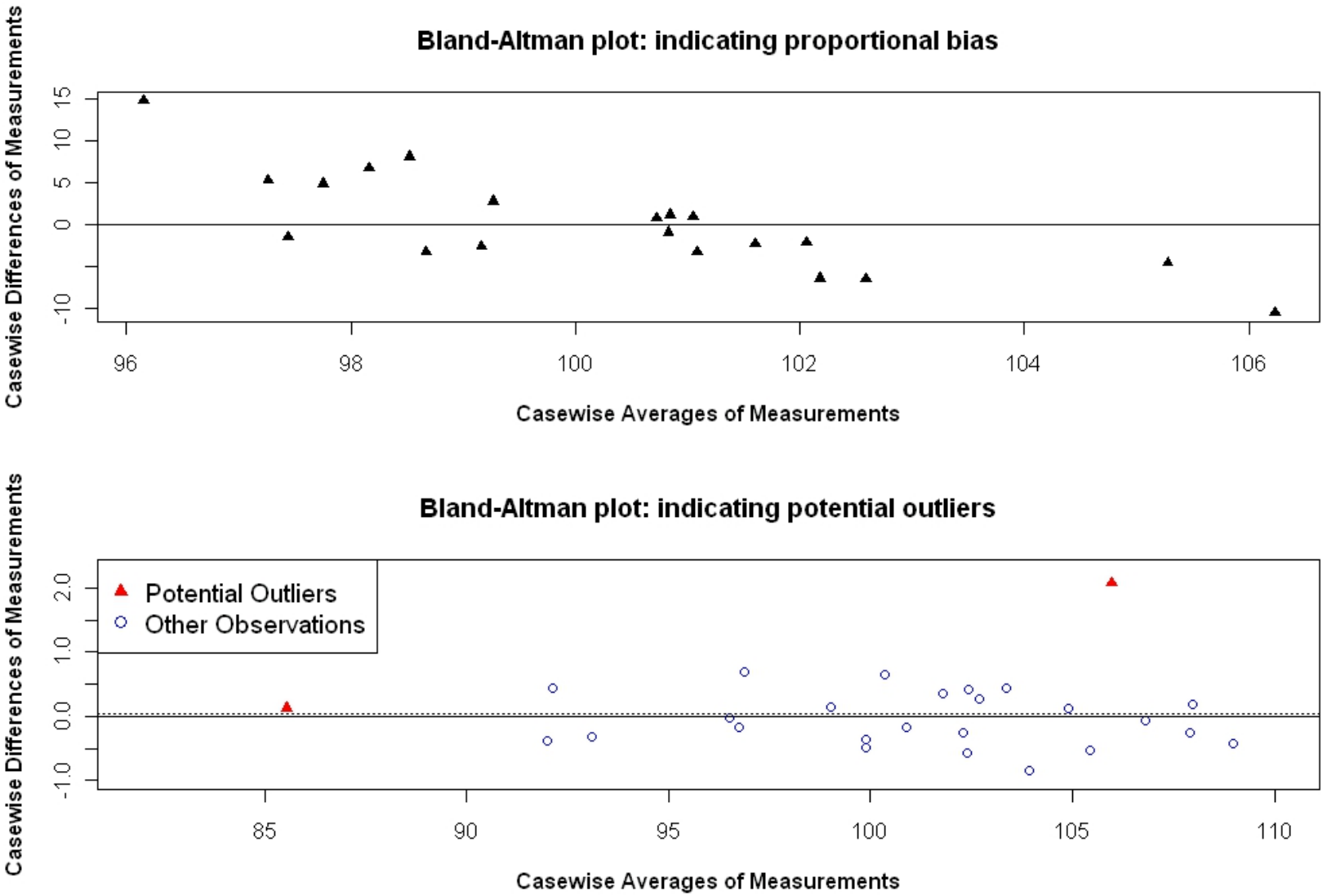
Interpreting the Bland-Altman Plot



Interpreting the Bland-Altman Plot



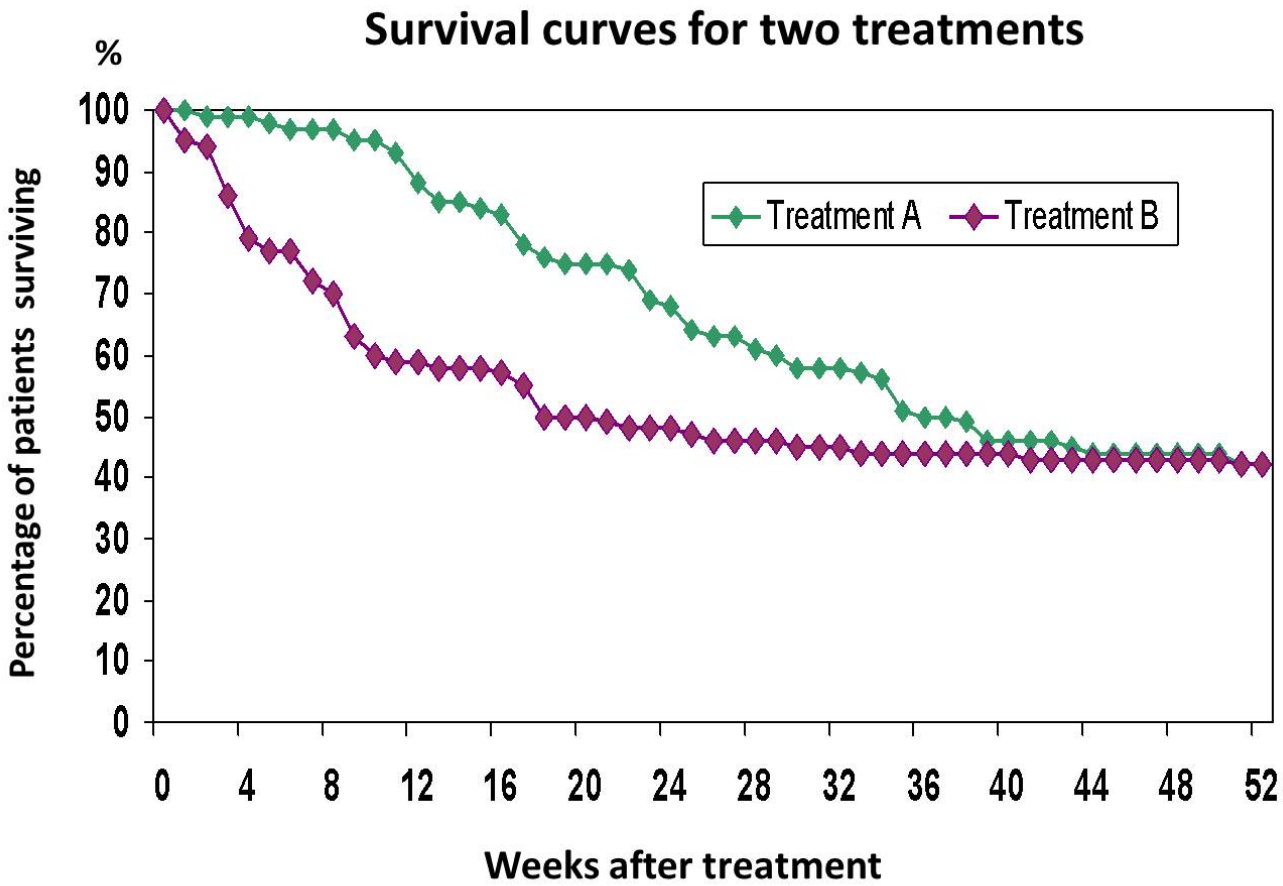
Interpreting the Bland-Altman Plot



The Bland-Altman Plot: Prevalence

- Limits of Agreement are used extensively in medical literature for assessing agreement between two methods.
- Building Blocks are featured in almost every undergraduate statistics course (i.e. Mean, Standard Deviation, Scatterplot, Normal Distribution)
- Other graphical techniques, such as *Survival-Agreement Plot* (based on Kaplan-Meier Curve) and *Mountain Plot* have been developed, but are not prevalent at all.

Kaplan Meier Survival Curve



Technology Acceptance Model

Davis (1989) proposes the TAM model, which suggests an hypothesis as to why users may adopt particular technologies, and not others.

When users are presented with a new technology, two important factors will influence their decision about how and when they will adopt it.

Perceived usefulness (PU) - This was defined by Fred Davis as "the degree to which a person believes that using a particular system would enhance his or her job performance".

Perceived ease-of-use (PEOU) - Davis defined this as "the degree to which a person believes that using a particular system would be free from effort"

Technology Acceptance Model

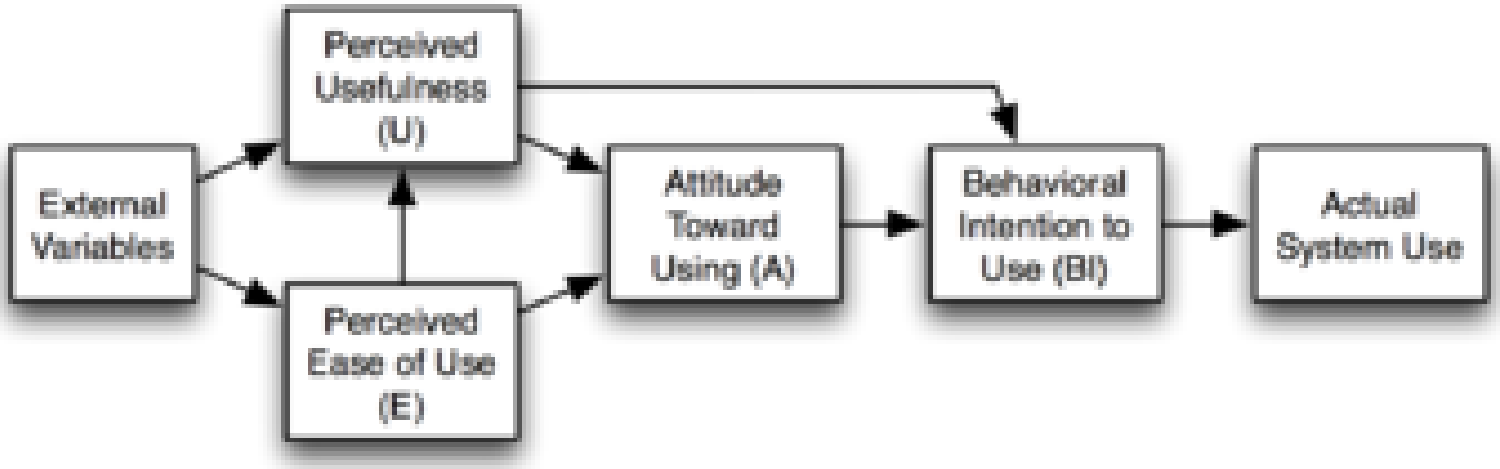


Figure: Technology Acceptance Model Flowchart (Davis,1989)

- Bland-Altman method not very good on it's own.
- Does not account for Replicate Measurements.
- Useful as a diagnostic method subsequent to other methods.
- Develop a proper methodology for MCS and Get people to use it!


Shiny
by RStudio
A web application framework for R

Shiny Web Applications with R

Useful Shiny Resources

- shiny.rstudio.com
- showmeshiny.com
- shiny.snap.uaf.edu/

Shiny-phyloseq

Shiny-phyloseq is an interactive web application that provides a graphical user interface to the microbiome analysis package for R, called phyloseq. For details about using the phyloseq package directly, see The phyloseq Homepage.

Citation

Shiny-phyloseq is provided under a free-of-charge, open-source license (A-GPL3). All we require is that you cite/attribute the following in any work that benefits from this code or application.

Citing the Web Application

McMurdie and Holmes (2014) Shiny-phyloseq: Web Application for Interactive Microbiome Analysis with Provenance Tracking. **Bioinformatics** *in press*.

Replicate Measurements

- Bland and Altman's approach originally devised for a single measurement on each item by each of the methods.
- Their 1999 paper [3] extended their approach to replicate measurements:
  By replicates we mean two or more measurements on the same individual taken in identical conditions.
  In general this requirement means that the measurements are taken in quick succession.
- Emphasis put on "repeatability".

Three Conditions

For two methods of measurement to be considered interchangeable, the following conditions must apply [1]:

- No significant inter-method bias
- No difference in the between-subject variabilities of the two methods
- No difference in the within-subject variabilities of the two methods (repeatability)

LME models

- In a linear mixed-effects model, responses from a subject are due to both fixed and random effects. A random effect is an effect associated with a sampling procedure.
- Replicate measurements would require use of random effect terms in model.
- Can have differing number of replicate measurements for different subjects.

Roy's Approach

- Roy proposes an LME model with Kronecker product covariance structure in a doubly multivariate setup.
- Response for *i*th subject can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_{1i} z_{i1} + b_{2i} z_{i2} + \epsilon_i$$

- β_1 and β_2 are fixed effects corresponding to both methods. (β_0 is the intercept.)
- b_{1i} and b_{2i} are random effects corresponding to both methods.

Roy's LME model

- Let $\boldsymbol{y}_i$ be the set of responses for subject i (in matrix form).
- $\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \epsilon_i$
- $\boldsymbol{b}_i \sim N_m(0, \boldsymbol{D})$ (m: number of methods)
- $\epsilon_i \sim N_{n_i}(0, \boldsymbol{R})$ (n_i: number of measurements on subject i)

Variance-covariance matrix

- Overall variance covariance matrix for response vector $\boldsymbol{y}_i$

$$\text{Cov}(\boldsymbol{y}_i) = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{R}_i$$

- can be re-expressed as follows:

$$\boldsymbol{Z}_i \begin{bmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{bmatrix} \boldsymbol{Z}_i' + \left(V \otimes \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

- Overall variability between the two methods is sum of between-subject and within-subject variability,

$$\text{Block } \boldsymbol{\Omega}_i = \begin{bmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Variance-Covariance Structures

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

- Symmetric structure specifies that σ_1^2 may differ from σ_2^2.
- Compound symmetric structure specifies that $\sigma_1^2 = \sigma_2^2$.
- In both cases, σ_{12} may take value other than 0.

The `nlme` Package

- LME models can be implemented in `R` using the `nlme` package, one of the core packages.
- Authors: Jose Pinheiro, Douglas Bates (up to 2007), Saikat DebRoy (up to 2002), Deepayan Sarkar (up to 2005), the `R` Core team
  (source: `nlme` package manual)
- "Mixed-Effects Models in S and S-PLUS" by JC Pinheiro and DM Bates (Springer,2000)

The Reference Model

```
REF = lme(y ~ meth,
    data = dat,
    random = list(item=pdSymm(~ meth-1)),
    weights=varIdent(form=~1|meth),
    correlation = corSymm(form=~1 | item/repl),
    method="ML")
```

- LME model that specifies a symmetric matrix structure for both between-subject and within-subject variances.

The Nested Model 1

```
NMB = lme(y ~ meth,
    data = dat,
    random = list(item=pdCompSymm(~ meth-1)),
    weights=varIdent(form=~1|meth),
    correlation = corSymm(form=~1 | item/repl),
    method="ML")
```

- LME model that specifies a compound symmetric matrix structure for between-subject and symmetric structure within-subject variances.

The Nested Model 2

```
NMW = lme(y ~ meth,
    data = dat,
    random = list(item=pdSymm(~ meth-1)),
    #weights=varIdent(form=~1|meth),
    correlation = corCompSymm(form=~1 | item/repl),
    method="ML")
```

- LME model that specifies a symmetric matrix structure for between-subject and compound symmetric structure within-subject variances.

The Nested Model 3

```
NMO = lme(y ~ meth,
    data = dat,
    random = list(item=pdCompSymm(~ meth-1)),
    #weights=varIdent(form=~1|meth),
    correlation = corCompSymm(form=~1 | /repl),
    method="ML")
```

- LME model that specifies a compound symmetric matrix structure for both between-subject and within-subject variances.

Example: Blood Data

- Used in Bland and Altman's 1999 paper [3]. Data was supplied by Dr E O'Brien.
- Simultaneous measurements of systolic blood pressure each made by two experienced observers, J and R, using a sphygmometer.
- Measurements also made by a semi-automatic blood pressure monitor, denoted S.
- On 85 patients, 3 measurement made in quick succession by each of the three observers (765 measurements in total)

Example: Blood Data

Inter-method Bias between J and S: 15.62 mmHg

```
>summary(REF)
.....
Fixed effects: y ~ meth
             Value Std.Error  DF t-value p-value
(Intercept) 127.41    3.3257 424  38.310       0
methS        15.62    2.0456 424   7.636       0
.....
```

Between-subject variance covariance matrix

```
..
Random effects:
 Formula: ~method - 1 | subject
 Structure: General positive-definite
            StdDev    Corr
methodJ   30.396975 methdJ
methodS   31.165565 0.829
Residual   6.116251
..
```

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 923.97 & 785.34 \\ 785.34 & 971.29 \end{pmatrix}$$

Within-subject variance covariance matrix

```
Correlation Structure: General
 Formula: ~1 | subject/obs
 Parameter estimate(s):
 Correlation:
  1
2 0.288
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | method
 Parameter estimates:
       J        S
1.000000 1.490806
```

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix}$$

Overall variance covariance matrix

- Overall variance

$$\text{Block } \hat{\Omega} = \hat{\boldsymbol{D}} + \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix}$$

- Standard deviation of the differences can be computed accordingly : 20.32 mmHg.

- Furthermore, limits of agreement can be computed: $[15.62 \pm (2 \times 20.32)]$ (mmHg).

Some useful R commands

- `intervals` :

  This command obtains the estimate and confidence intervals on the parameters associated with the model.
  This is particularly useful in writing some code to extract estimates for inter-method bias and variances, and hence estimates for the limits of agreement.

- `anova` :

  When a reference model and nested model are specified as arguments, this command performs a likelihood ratio test.

Formal Tests: Between-subject Variances

- Test the hypothesis that both methods have equal between-subject variances.
- Constructed an alternative model "Nested Model B" using ***compound symmetric*** form for between-subject variance (hence specifying equality of between-subject variances).
- Use a likelihood ratio test to compare models.

```
...
> anova(REF,NMB)
    Model df ...      logLik   Test   L.Ratio p-value
REF     1  8 ...  -2030.736
NMB     2  7 ...  -2030.812 1 vs 2 0.1529142  0.6958
...
```

- Fail to reject hypothesis of equality.

Formal Tests: Within-subject Variances

- Test the hypothesis that both methods have equal within-subject variances.
- Constructed an alternative model “Nested Model W" using compound symmetric form for within-subject variance (hence specifying equality of within-subject variances).
- Again, use a likelihood ratio test to compare models.

```
...
> anova(REF,NMW)
    Model df ...    logLik   Test  L.Ratio p-value
REF     1  8 ... -2030.736
NMW     2  7 ... -2045.044 1 vs 2 28.61679  <.0001
```

- Reject hypothesis of equality.

Formal Tests : Outcomes

- Inter-method bias: Significant difference in mean values detected.
- Between-subject variance: No significant difference in between-subject variances between the two methods detected.
- Within-subject variance: A significant difference in within-subject variances is detected.
- Can not recommend switching between the two methods.

Remarks

- Can perform a test for equality of overall variances.
- This can be done by specifying a compound symmetry structure for both between-subject and within-subject variances when constructing a nested model.
- Roy controls the family-wise error rate in paper, using Bonferroni correction procedure.

References

- A Roy (2009): *An application of linear mixed effects model to assess the agreement between two methods with replicated observations* Journal of Biopharmaceutical Statistics
- Bland JM, Altman DG (1986) *Statistical method for assessing agreement between two methods of clinical measurement*.
- Bland JM, Altman DG (1999) *Measuring agreement in method comparison studies.* Statistical Methods in Medical Research
- Pinheiro JC, Bates DM (2000): *Mixed-effects models in S and S-PLUS*, Springer.

Thanks

- Dr Kevin Hayes, University of Limerick
- Dr Kevin Burke, University of Limerick
- Peter Fennell