

Embedding equivalence t-test results in Bland Altman Plots visualising rater reliability

**Jim Groeneveld,
OCS Consulting,
's Hertogenbosch, Netherlands.**

PhUSE 2011

AGENDA / CONTENTS

- A. Rater reliability (inter- / intra-)**
- B. Methods, variable type dependent**
- C. Equivalence t-test (quantitative)**
- D. Bland Altman Plots (qualitative)**
- E. Integration of both, visualising equivalence t-test results in Bland Altman Plots, showing quantitative (in)significant equivalence in the plots**
- F. Advantages of integration**

A. Rater reliability

1. Determine reliability of measuring instrument (device and/or human)
2. Repeated measurements (judgments by raters) on same objects
 - a. by **same** instrument: **intra-rater** or **within-rater** reliability (2 or more repetitions)
 - b. by similar, but **other** instrument: **inter-rater** or **between-rater** reliability (2 or more)
3. Application (before and after study):
 - a. **Certification** on representative data (before)
 - b. **QC** (on sample) of existing study data (after)

B. Methods, variable type dependent

1. Categorical data (nominal or ordered)

- a. Cohen's Kappa analysis (>2 cats: Fleiss)
- b. McNemar's test (>2 cats: McNemar-Bowker)

Application: non-missing vs missing (binary)

2. Continuous data (interval or ratio)

- a. Mean Absolute Difference (MAD) of pairs
- b. Intraclass Correlation Coefficient (ICC), pairs
- c. Equivalence t-test (quantitative interpretation)
- d. Bland Altman Plots (qualitative interpretation)

Application: ordered multi-level categorical data

C. Equivalence t-test (range limits)

- 1. on differences between paired measurements**
- 2. two one-sided non-inferiority t-tests**
- 3. user specification of equivalence range limits ((a)symmetrical)**

Result for each combination of pairs of matching, repeated measurements:

- 1. significant equivalence or not**
- 2. depending on range limits**

D. Bland Altman Plots

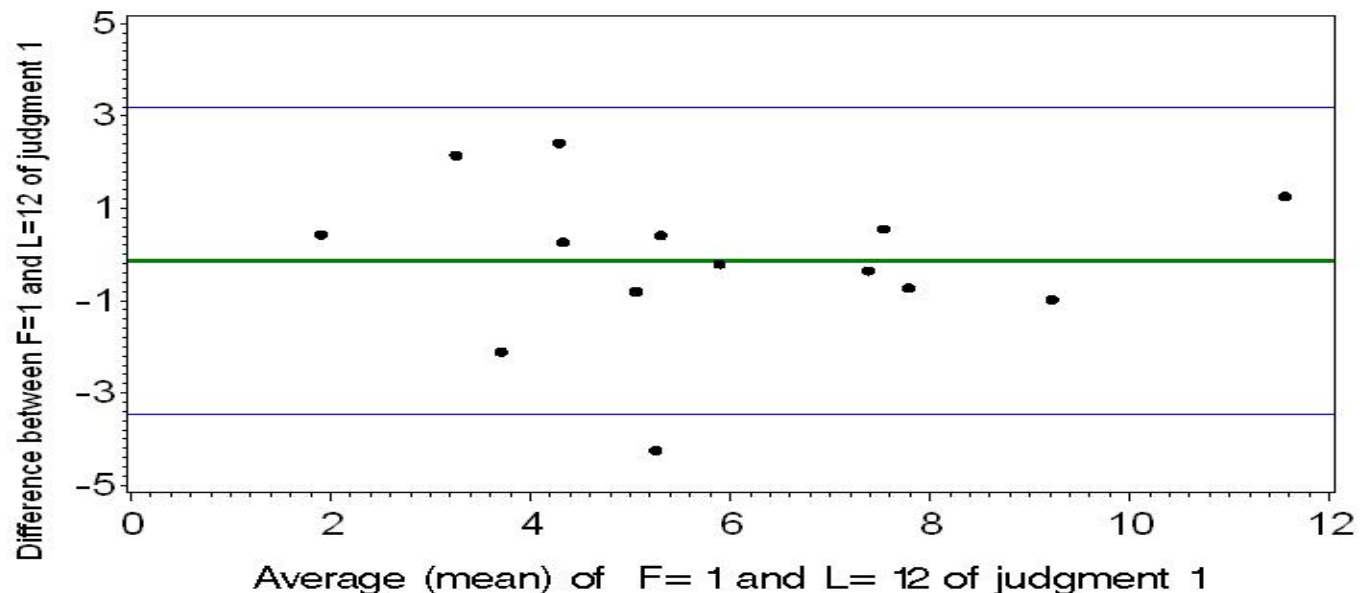
1. Scattergram of pairwise points of:
2. Mean of pairs: $X = (v_1 + v_2) / 2$ versus
3. Difference of pairs: $Y = v_1 - v_2$ including
4. Horizontal line of mean difference and
5. Confidence Interval (CI) of points, upper and lower horizontal lines
6. Qualitative interpretation of reliability

D. Bland Altman Plots (example)

PhUSE 2011 Conference / Jim Groeneveld, OCS Consulting

Example usual Bland Altman Plot from fake data (inter-rater reliability)

Between raters: Judgment=Measure Rater1=Alfred Rater2=Peter

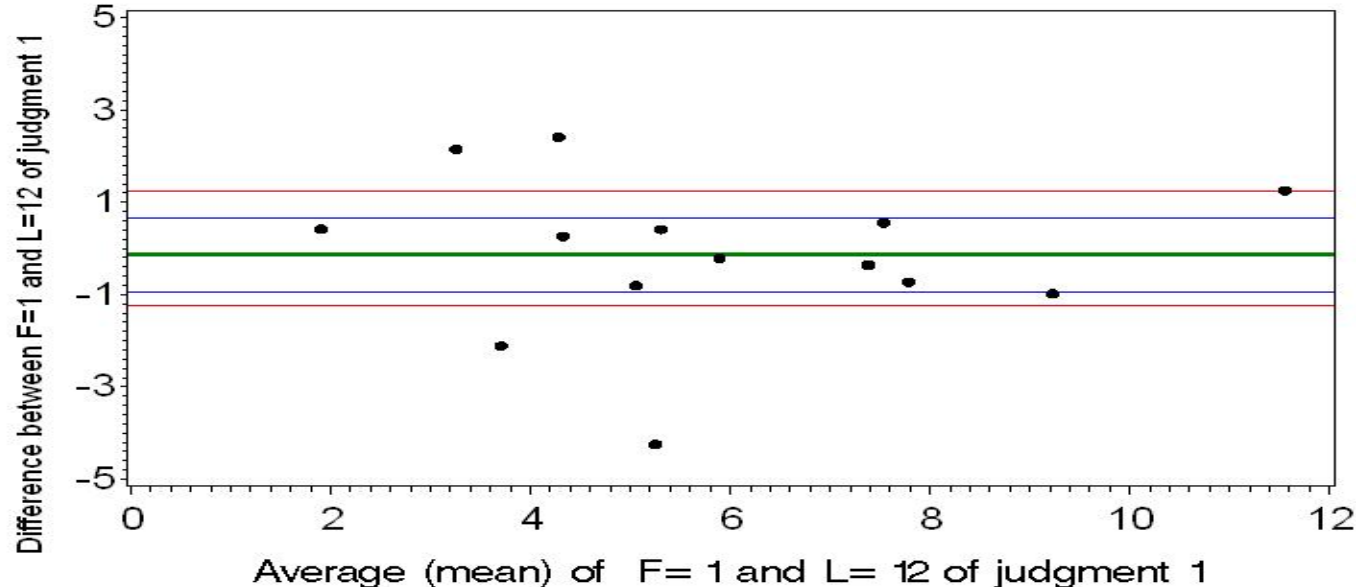


E. Integration of equivalence t-test and Bland Altman Plots

1. Scattergram of pairwise points of:
2. Mean of pairs: $X = (v_1 + v_2) / 2$ versus
3. Difference of pairs: $Y = v_1 - v_2$ including
4. Horizontal line of mean difference and
- 5. Confidence Interval (CI) of the mean, upper and lower horizontal lines**
- 6. T-test range limits, horizontal lines**
- 7. Quantitative interpretation of reliability**

E. Integration of equivalence t-test and Bland Altman Plots (example with significant equivalence)

PhUSE 2011 Conference / Jim Groeneveld, OCS Consulting
Example Bland Altman Plot with Equivalence T-test results from fake data
Equivalence range limits: -1.25 to $+1.25$ (influences p-value & significance)
Between raters: Judgment=Measure Rater1=Alfred Rater2=Peter



E. Integration of equivalence t-test and Bland Altman Plots

- 1. visualising equivalence t-test results in Bland Altman Plots**
- 2. showing quantitative significant equivalence in the plots**
- 3. if the Confidence Interval of the mean lies fully within the T-test range limits there is significant equivalence**

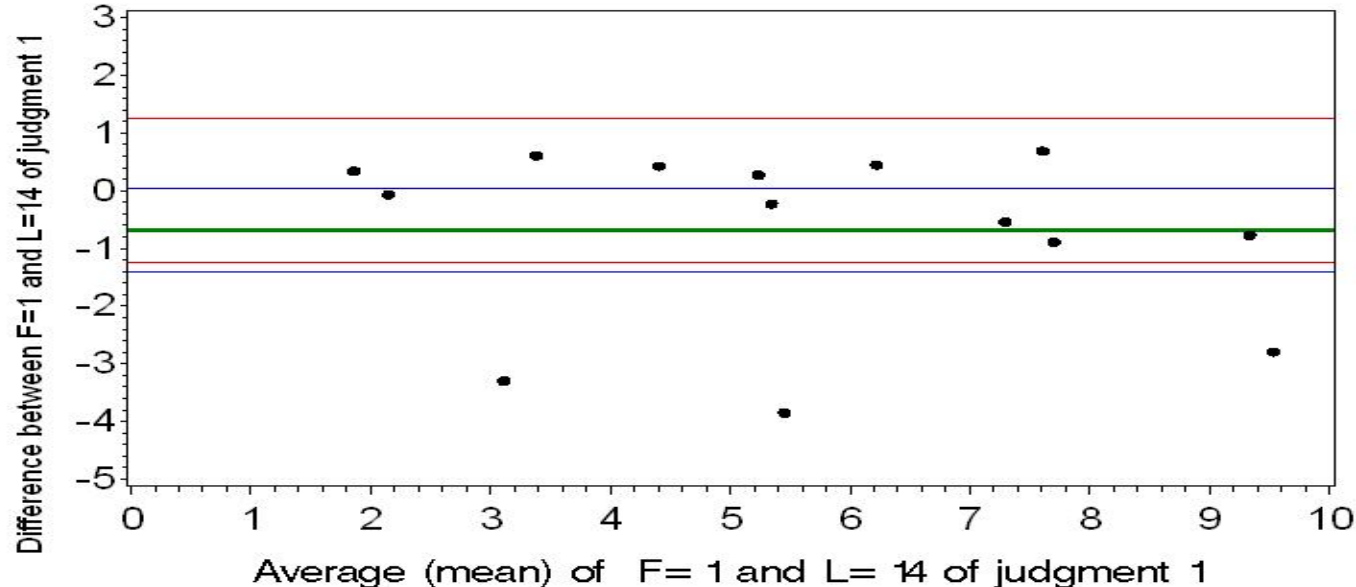
E. Integration of equivalence t-test and Bland Altman Plots (example with non-significant equivalence)

PhUSE 2011 Conference / Jim Groeneveld, OCS Consulting

Example Bland Altman Plot with Equivalence T-test results from fake data

Equivalence range limits: -1.25 to $+1.25$ (influences p-value & significance)

Between raters: Judgment=Measure Rater1=Alfred Rater2=Tommy



F. Advantages of integration

- 1. Extension of (value of) Bland Altman Plots with quantitative interpretation on equivalence (in)significance**
- 2. Equivalence (in)significance clearly visualised, depending on range limits**
- 3. Results of two reliability analysis methods in one plot**
- 4. showing a quantitative result and a qualitatively interpretable scatterplot**

QUESTIONS & ANSWERS

SASquestions@ocs-consulting.com

Jim.Groeneveld@ocs-consulting.com

<http://jim.groeneveld.eu.tf>

More than 2 matching measurements

**1. Pairwise analysis of repetitions
(may yield many pairs of more than 3)**

**2. If more than 3 reduce number of
analyses to “pairs” consisting of:**
 a. each individual measurement versus
 b. the mean of all other matching measurements

**This reduces the amount of “pairs” and
analyses and facilitates an overall
interpretation of the results.**

A SAS macro (Concord) is currently under development in which these techniques already are supported and applied.

Additional features: relative differences

1. difference between both values: $Y = v_1 - v_2$
2. **proportional difference with mean of both:** $Y = (v_1 - v_2) / \text{mean}[v_1, v_2] = 2 * (v_1 - v_2) / (v_1 + v_2)$
3. **(relative) proportion of both values, minus 1:** $Y = (v_1 / v_2) - 1 = (v_1 - v_2) / v_2$
4. **proportion of 1 value of mean of both, minus 1:** $Y = (v_1 / \text{mean}[v_1, v_2]) - 1 = (v_1 - v_2) / (v_1 + v_2)$

SAS Macro TickMark (*version 0.0.1*)

Neat automatic ticmarks for graphs based on minimum and maximum of an existing value range (tickmarks 1 to 2 significant digits).

Optional specification: desired minimum and maximum number of tick marks and minimum percentage of coverage of existing data range by generated value range (default values: minimum=7, maximum=12, pct coverage=80).

Return of From, To and By values via macro variables or as a single return value.