

Übereinstimmung von Messmethoden stetiger Variablen

Kaspar Rufibach

Abteilung Biostatistik
Institut für Sozial- und Präventivmedizin
Universität Zürich

11. September 2009
Stadtspital Waid, Zürich

- 1 Übereinstimmung von Messmethoden
- 2 Beispiele
- 3 Bland-Altman Plot
- 4 Fallzahl
- 5 Erweiterungen & Software

Medizin: Zuverlässigkeit von Bewertungen bedeutsam.

- **Interrater agreement:** Beurteilung eines Merkmals durch mehrere Bewerter.
Klassifizierung einer Krankheit anhand eines diagnostischen Verfahrens, Einteilung des psychischen Zustands eines Patienten.
- **Intrarater agreement:** Wiederholte Beurteilung eines Merkmals durch einen Bewerter.
Befundung desselben Röntgenbildes durch denselben Radiologen, Einteilung von Gewebeproben durch einen Pathologen jeweils zu verschiedenen Zeitpunkten.

Beurteilung der **Übereinstimmung:** Auskunft über **Reliabilität** oder **Reproduzierbarkeit** einer Messung.

Kalibrierung: Die **korrekten** Werte sind bekannt, wir "eichen" eine neue Methode daran.

Wahl der Methode hängt vom **Datentyp** ab.

- Nominale Daten: Cohen's κ - Koeffizient.
Normal vs. nicht-normal, vier (ungeordnete) Kategorien einer Krankheit.
- Ordinale Daten: Gewichteter κ - Koeffizient.
Grad einer Krankheit.
- Stetige Daten, Beurteilung der Übereinstimmung zweier Methoden:
Bland-Altman Analyse.
Messungen, Beurteilung mittels eines Scores auf einer Skala von 0-100.

Wahl der Methode hängt vom **Datentyp** ab.

- Nominale Daten: Cohen's κ - Koeffizient.
Normal vs. nicht-normal, vier (ungeordnete) Kategorien einer Krankheit.
- Ordinale Daten: Gewichteter κ - Koeffizient.
Grad einer Krankheit.
- Stetige Daten, Beurteilung der Übereinstimmung zweier Methoden:
Bland-Altman Analyse.
Messungen, Beurteilung mittels eines Scores auf einer Skala von 0-100.

13'651× zitiert Stand 9. September 2009

STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT

J. Martin Bland, Douglas G. Altman

Department of Clinical Epidemiology and Social Medicine, St. George's Hospital Medical School, London SW17 ORE; and Division of Medical Statistics, MRC Clinical Research Centre, Northwick Park Hospital, Harrow, Middlesex

SUMMARY

In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analysed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, is described, together with the relation between this analysis and the assessment of repeatability.

(*Lancet*, 1986; **i**: 307-310)

- 1 Übereinstimmung von Messmethoden
- 2 Beispiele**
- 3 Bland-Altman Plot
- 4 Fallzahl
- 5 Erweiterungen & Software

Glukosetest:

- Wechsel von Hitachi912 auf Cobas6000.

Variable	n	Min	\tilde{x}	\bar{x}	Max	s	IQR
Hitachi 912	39	1.9	6.5	10.0	36.8	8.0	8.7
Cobas 6000	39	1.9	6.7	9.7	35.4	7.8	8.7

Bestimmung der Folsäure:

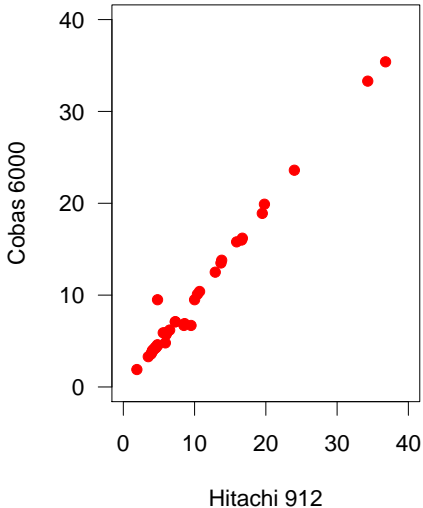
- Wechsel von alter auf neue Methode.

Variable	n	Min	\tilde{x}	\bar{x}	Max	s	IQR
Folsäure, alte Methode	40	11.5	20.6	23.5	45.4	10.6	12.4
Folsäure, neue Methode	40	8.2	16.6	18.0	45.4	8.8	9.0

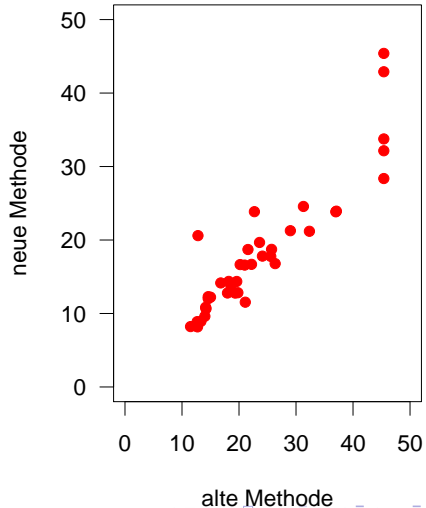
Fragestellungen:

- Hängt der Messfehler vom “wahren” Wert ab?
Gültigkeit Limits of agreement, Konfidenzintervalle.
- Gibt es auffällige Beobachtungen (“Ausreisser”)?
- Wie stark unterscheiden sich die Messungen beider Methoden?
Quantifizierung des **Bias**.
- Ist ein allfälliger Unterschied **klinisch relevant**?

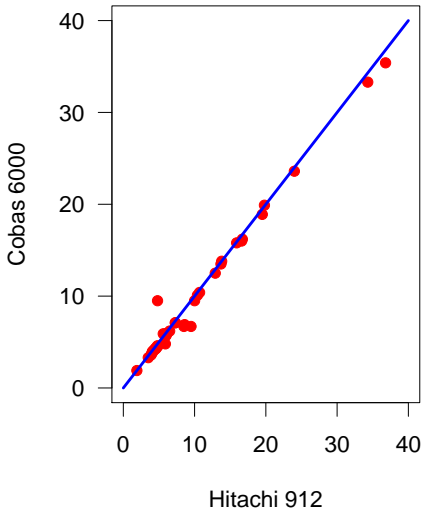
Glukose



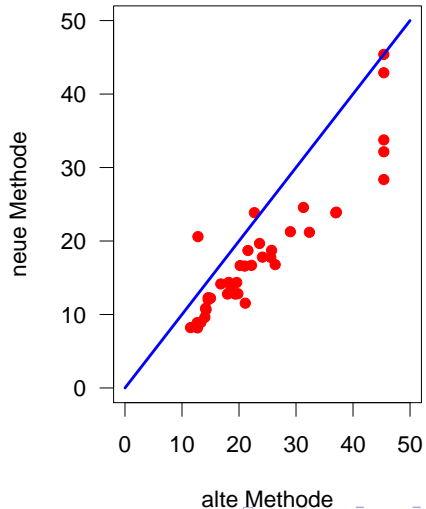
Folsäure



Glukose



Folsäure



Glukose

- Methoden scheinen sehr ähnliche Werte zu liefern.
- (Mindestens) ein “Ausreisser”.

Folsäure

- Neue Methode liefert systematisch kleinere Werte.
- Alte Methode liefert $5\times$ den Wert 45.4?

Blaue Linie: Winkelhalbierende, **NICHT** Regressionsgerade.

Korrelationskoeffizient r : Sollte **NICHT** angegeben werden.

- r misst Korrelation, nicht Übereinstimmung. Übereinstimmung: Punkte entlang Winkelhalbierende \Leftrightarrow Korrelation von 1: Punkte entlang irgendeiner Gerade.
- Änderung der Skala: Belässt Korrelation, ändert selbstverständlich Übereinstimmung.
- Korrelation hängt vom Wertebereich ab \Rightarrow Methoden werden über ganzen gemessenen Bereich verglichen, Korrelation sowieso hoch.
- Schlecht übereinstimmende Methoden können hohe Korrelation aufweisen.

Regression: Macht ebenfalls wenig Sinn.

Zitat aus Bland & Altman (1986):

*Why has a totally inappropriate method, the correlation coefficient, become almost universally used for this purpose? Two processes may be at work here - namely, **pattern recognition** and **imitation**. Once the correlation approach has been published, others will read of a statistical problem similar to their own being solved in this way and will use the same technique with their own data. Medical statisticians who ask "why did you use this statistical method?" will often be told "**because this published paper used it**".*

- Hängt der Messfehler vom “wahren” Wert ab?
Gültigkeit Limits of agreement, Konfidenzintervalle.
- Gibt es auffällige Beobachtungen (“Ausreisser”)?
- Wie stark unterscheiden sich die Messungen beider Methoden?
Quantifizierung des **Bias**.
- Ist ein allfälliger Unterschied **klinisch relevant**?

- 1 Übereinstimmung von Messmethoden
- 2 Beispiele
- 3 Bland-Altman Plot**
- 4 Fallzahl
- 5 Erweiterungen & Software

Beurteile Abhängigkeit des Messfehlers vom “wahren” Wert.

x-Achse: Mittelwert der beiden Methoden.
Schätzung des “wahren” Werts.

y-Achse: Differenz der beiden Methoden.
Differenzenbildung eliminiert Variabilität zwischen Beobachtungen \Rightarrow übrig
bleibt **Messfehler**.

Größenordnung und Muster der individuellen Abweichungen besser sichtbar als in
Punktediagramm.

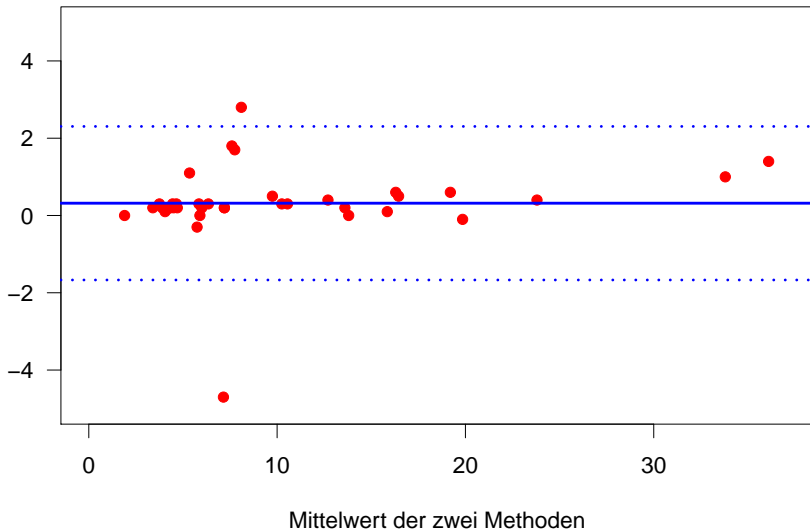
Plot der Differenzen gegen eine der beiden Methoden ungeeignet: Differenz
korreliert mit jeweils einer Messmethode.

Aus dem **Bland-Altman** Plot können wir ablesen:

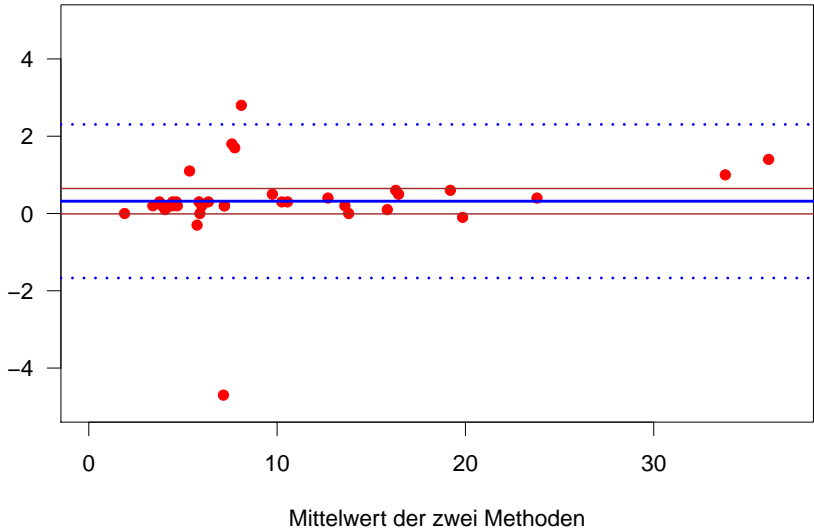
- **Bias \bar{d}** (systematische Abweichung): Mittelwert der Differenzen, inklusive **Konfidenzintervall** (NV-Annahme).
- **Standardabweichung s der Differenzen**, d.h. Messfehler: \approx NV, auch falls Messungen \neq NV.
- **Limits of agreement**: 95% der Punkte in $[\bar{d} - 1.96 \cdot s, \bar{d} + 1.96 \cdot s]$ wobei s die Standardabweichung der Differenzen ist.
Wenn Differenzen innerhalb $\bar{d} \pm 1.96 \cdot s$ **klinisch nicht relevant** sind \Rightarrow Methoden sind austauschbar.
- Liegt eine Abhängigkeit des Messfehlers vom “wahren” Wert vor? Punkte bilden nach rechts offenen Trichter \Rightarrow Transformation.
log wenn Differenzen proportional zum “wahren” Wert.
- Liegen “Ausreisser” vor?

Limits of agreement sind **Schätzungen**, eine andere Stichprobe würde andere LOA liefern \Rightarrow Konfidenzintervalle für LOA können ebenfalls angegeben werden (NV-Annahme).

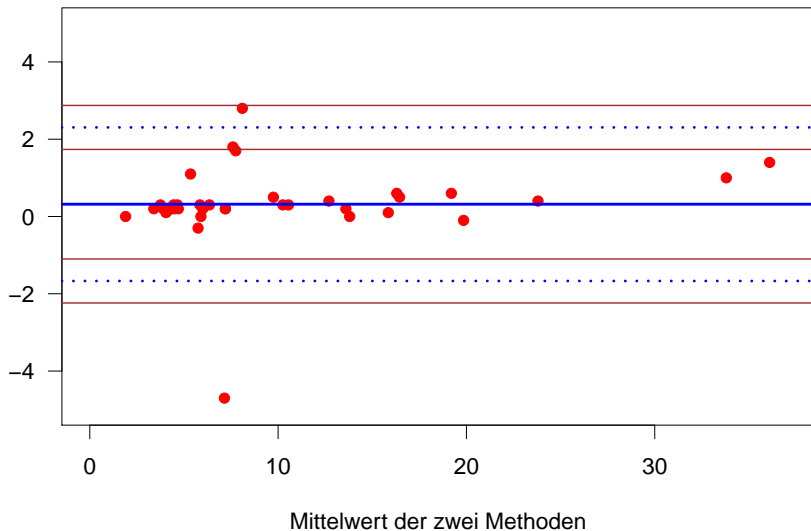
Differenz der zwei Methoden

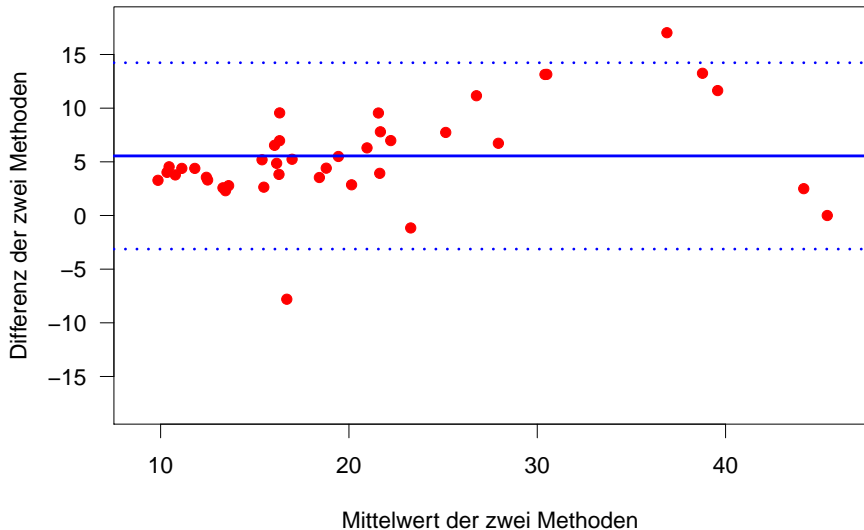


Differenz der zwei Methoden

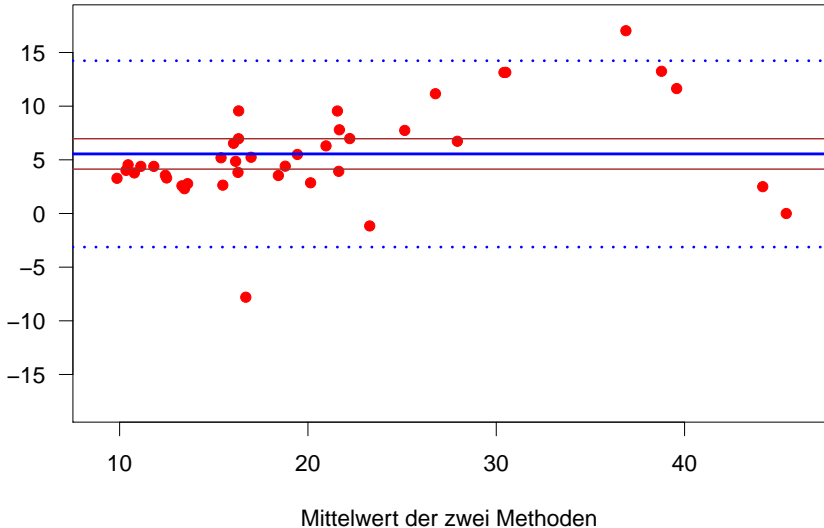


Differenz der zwei Methoden

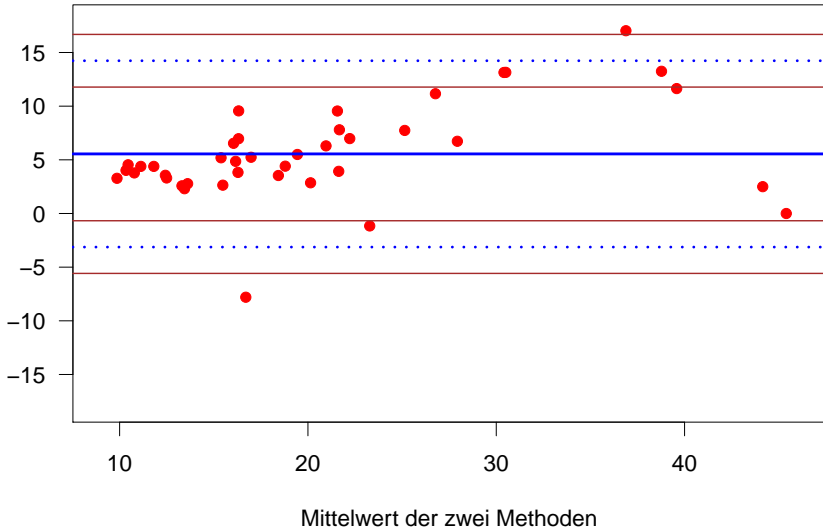




Differenz der zwei Methoden



Differenz der zwei Methoden



Wenn Differenzen innerhalb $\bar{d} \pm 1.96 \cdot s$ **nicht klinisch relevant** sind \Rightarrow Methoden sind austauschbar.

Klinische, keine **statistische** Beurteilung!

Welche Abweichung noch als klinisch irrelevant betrachtet wird, sollte **vor** den Messungen festgelegt werden!

Betonung auf Quantifizierung des Unterschieds, nicht auf statistischem Test für "Gleichheit" der Methoden.

Beurteilung der Übereinstimmung von Messmethoden gelingt nicht mit **einer einzigen** Masszahl.

Zu einer vollständigen Bland-Altman Analyse gehören:

- Bland-Altman Plot,
- Diskussion Ausreisser, Abhängigkeit Messfehler vom “wahren” Wert, Transformationen,
- Bias mit Konfidenzintervall,
- Limits of agreement mit Konfidenzintervallen,
- Diskussion der klinischen Relevanz der Differenzen und der Austauschbarkeit.

Wir kennen “wahren Messwert” (\Rightarrow Kalibrierung) nicht:

- Nur **Vergleich** der zwei Methoden möglich,
- wertende Aussage: “Welche Methode ist besser?” oder
- eine Aussage ob überhaupt eine Methode zur Messung der zugrundeliegenden Grösse geeignet ist, ist **nicht möglich**.

- 1 Übereinstimmung von Messmethoden
- 2 Beispiele
- 3 Bland-Altman Plot
- 4 Fallzahl**
- 5 Erweiterungen & Software

Mehr Beobachtungen \Rightarrow Schätzungen

- des Bias und
- der Limits of agreement

werden präziser, d.h. die Konfidenzintervalle enger.

Limits of agreement werden **nicht** enger!

Lege **maximale Breite** der Konfidenzintervalle für die Limits of agreement fest \Rightarrow damit kann Fallzahl berechnet werden.

- 1 Übereinstimmung von Messmethoden
- 2 Beispiele
- 3 Bland-Altman Plot
- 4 Fallzahl
- 5 Erweiterungen & Software

Stetige Daten, Übereinstimmung mehrere Methoden und mehrerer Bewerter:
Intraklassen-Korrelationskoeffizienten.

Mehrere Radiologen zählen Anzahl entzündete Läsionen mit verschiedenen Methoden.

Wiederholte Messungen: identisches Vorgehen wie bei zwei verschiedenen Methoden.

Beurteilung Übereinstimmung zweier Methoden macht wenig Sinn wenn Resultate **einer** Methode nicht präzis genug reproduziert werden können.

Longitudinale Messungen: farbliche Unterscheidung der Punkte im Plot, kompliziertere Methoden.

Messungen eines Scores an **mehreren Zeitpunkten** mit zwei Methoden.

Mehrere Messungen an einem Subjekt: farbliche Unterscheidung der Punkte im Plot, kompliziertere Methoden.

Mehrere Messungen der Gefäßdicke der Herzarterie pro Person.

Brute force: Von Hand Plots erstellen und relevante Größen berechnen: mit beliebiger Software möglich.

SPSS: Nicht implementiert.

R: Funktionen zur Bland-Altman Analyse auf Anfrage. ICCs in diversen Zusatz-Paketen verfügbar.

Abteilung Biostatistik
Institut für Sozial- und Präventivmedizin
Universität Zürich
Hirschengraben 84
8001 Zürich
kaspar.rufibach@ifspm.uzh.ch

<http://www.biostat.uzh.ch>

Statistische Beratung

Skript "Einführung in Biostatistik" (175 Seiten, Sfr 22)

Vielen Dank für Ihre Aufmerksamkeit.