Faculty of Health Sciences

Reliability of measurement methods

Statistical analysis of repeated measurements 2024

Julie Forman

Section of Biostatistics, University of Copenhagen

Course day 4 contents

Part I. Cross-over studies

- Consideration about study design and statistical analysis
- Single measures from an AB-BA design
- Repeated measurements from an AB-BA design

Part II. Reliability of measurement methods

- Considerations about study design and statistical analysis
- Reliability of a single measurement method
- Agreement between two measurement methods

Outline

Evaluating measurement methods

Reliability of a single measurement method

Agreement between two measurement methods

Case Study: Comparison of two devices

- Two devices for measuring peak expiratory flow rate (l/min).
- 17 test persons, two replicates with **each device**.

	Wright-meter		**mini Wright-meter**	
id	wright1	wright2	mini1	mini2
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
16	423	372	350	370
17	427	421	451	443

Reference: Bland and Altman, *Statistical methods for assessing agreement between two methods of clinical measurement*, Lancet (1986).

Plan for a typical investigation

Can the new device replace the old in clinic?

Quantify the precision of each device.

- How precise are the two devices?

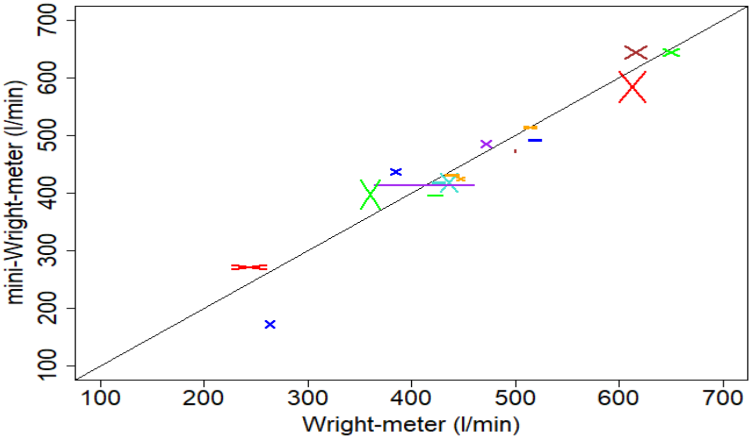
We look at the typical differences between two replicates; aka **limits of agreement** (normal range $\pm 2\sqrt{2}\cdot$ replication error SD).

Quantify the agreement between the two devices.

- Is the new device biased compared to the old?
- Are the two devices equivalent within a reasonable margin?
  - On average? - At the individual level?

We look at the mean difference and the typical differences between measurements with the new and old device; **limits of agreement**.

Case study: Starplot



Vertices connect pairs of measurements from the same test person (wright1,mini1)↔(wright2,mini2), (wright2,mini1)↔(wright1,mini1)

Considerations about study design and statistics

Do you have a gold standard/know the ground truth?

- Then you can evaluate *bias* and *accuracy*.
  Otherwise you can only evaluate *precision*.

Does the study include one or more devices (conditions)?

- Devices (conditions) may differ systematically (fixed effect), while technical replicates do not (variance component).

What is the total number of measurements per subject?

- > 2 in general requires a mixed model, but there are work-arounds for balanced data (e.g. the case study).

Report intraclass correlations (ICC) or limits of agreement?

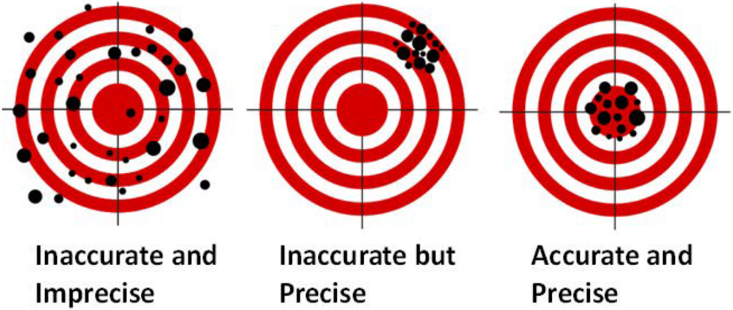
- **Recommended choice:** *limits of agreement*.

Did you make a power calculation?

- Lack of evidence is not the same as *equivalence*.

Basic concepts: Bias, accuracy and precision

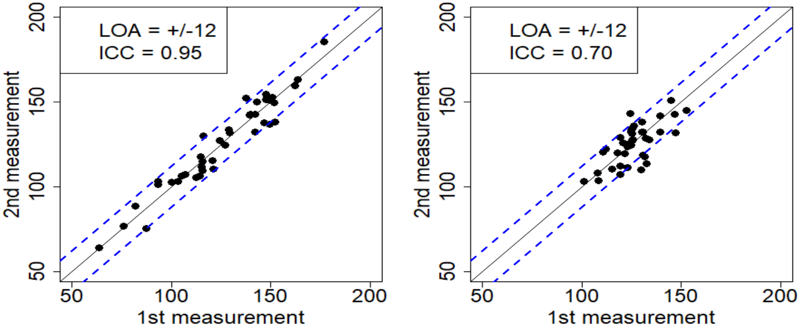
A biased device may be precise but not accurate.



- To evaluate accuracy you need to know what the truth is (ground truth in a planned experiment or gold standard).
- Precision can be assessed from technical replicates alone.

intraclass correlation (ICC) vs limits of agreement

Hypothetical example: Same device evaluated in two different populations, one very homogenous and one very heterogeneous.



- **ATT:** Same limits of agreement, but very different ICCs.

We disrecommend ICC as a measure of reliability. ICCs are not comparable between studies and not clinically operational.

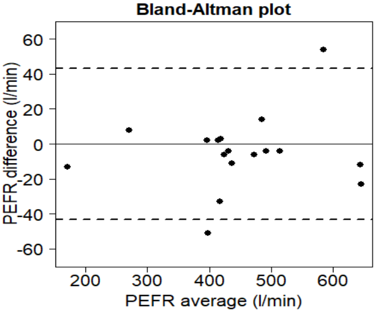
Outline

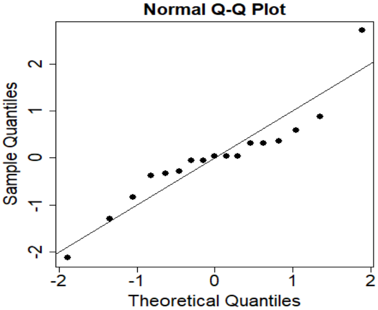
Evaluating measurement methods

Reliability of a single measurement method

Agreement between two measurement methods

Precision of a single measurement method (`wright`)





- Plot difference vs average of the two replicates.
- Model assumption: Differences are normally distributed.
- No bias since technical replicates are exchangeable.

Symmetric limits of agreement: $\pm 2 * \mathrm{SD(dif)}$

Same analysis based on a two-level model

Describe the k'th replicate from the jth subject as:

$$Y_{jk} = \mu + A_j + \varepsilon_{jk}$$

μ : Mean outcome in population (intercept).

A_j: Individual deviation (random effect of subject).

ε_{jk}: Replication error (residual).

We assume that A_js and ε_{jk}s are independent and normally distributed with zero mean. Normality is important for the ε_{jk}s.

level	**variance component**
2	τ^2 between subjects
1	ω^2 within subjects (replication error variance)

Note: The parameter of primary interest is ω.

Two-level model in R

Note: No covariates in the model formula, only an intercept.

```
library(lme4)

sym.lme <- lmer(pefr~1+(1|id), data=subset(long, method=="wright
```

Estimated variance components from summary:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 13682.8  116.97
 Residual               234.3   15.31
```

Compute limits of agreement

The difference between the two replicates is

$$\mathsf{dif}_j = Y_{jk_1} - Y_{jk_2} = \varepsilon_{jk_1} - \varepsilon_{jk_2} \sim \mathcal{N}(0, 2\omega^2)$$

Hence, the normal range for the differences is $\pm\, 2\sqrt{2}\omega$.

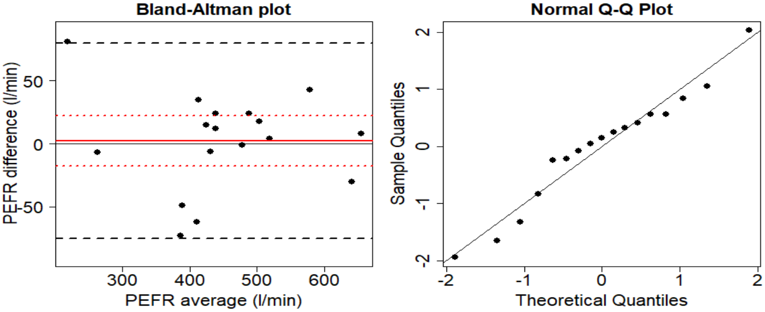
Case: Limits-of-agreement for the wright-meter:

$$\pm 2\sqrt{2} \cdot 15.31 \;\simeq\; \pm\, 43.3 \text{ l/min}$$

Note: Use `confint(sym.lme)` to get a confidence interval.

Outline

One measurement from each device (mini vs wright)



- Plot difference vs average of the two replicates.
- Model assumption: Differences are normally distributed.
- Possible bias: Compute mean difference $\overline{\text{dif}}$ with 95% CI.

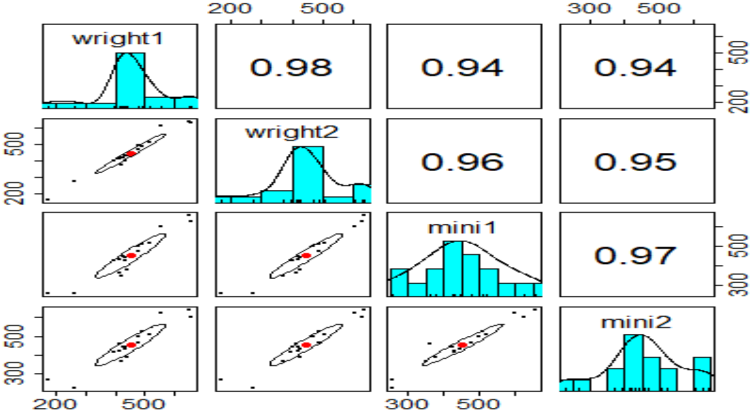
Asymmetric limits of agreement: $\overline{\text{dif}} \pm 2 * \text{SD}(\text{dif})$

What if we have replicates for each method?

Common approaches with replicates:

1. Make an analysis based on the first replicate only.
   - We throw away half of the data.
2. Average the replicates before comparing the mehods.
   - Averaging reduces the natural replication error.
3. Treat replicates as independent data from a new person.
   - Ignoring correlation may bias the results.
4. Model the replicates in a linear mixed model.
   - Makes optimal use of the data (but technical).

Case: Repeated measurements



- Data looks reasonably normal.
- Somewhat stronger correlation between measurements made with same device compared to with different devices.

Modeling considerations

Means: Fixed effect of `method`

- Two means for the methods, μ_1 and μ_2.
- The bias is $\mu_2 - \mu_1$.

Covariance pattern: Blocked compound symmetry.

- Two variances for the two methods, σ_1^2 and σ_2^2
- Two correlations within the two methods, ρ_1 and ρ_2, and one correlation between them, κ

Correlation matrix

$$\begin{pmatrix} 1 & \rho_1 & \kappa & \kappa \\ \rho_1 & 1 & \kappa & \kappa \\ \kappa & \kappa & 1 & \rho_2 \\ \kappa & \kappa & \rho_2 & 1 \end{pmatrix}$$

Covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_1^2\rho_1 & \sigma_1\sigma_2\kappa & \sigma_1\sigma_2\kappa \\ \sigma_1^2\rho_1 & \sigma_1^2 & \sigma_1\sigma_2\kappa & \sigma_1\sigma_2\kappa \\ \sigma_1\sigma_2\kappa & \sigma_1\sigma_2\kappa & \sigma_2^2 & \sigma_2^2\rho_2 \\ \sigma_1\sigma_2\kappa & \sigma_1\sigma_2\kappa & \sigma_2^2\rho_2 & \sigma_2^2 \end{pmatrix}$$

R-code

Step 1. Create a factor corresponding to the four replicates:

```
long$rep.method <- interaction(long$replicate, long$method)

table(long$rep.method)

1.wright 2.wright   1.mini   2.mini
      17       17       17       17
```

Step 2. Fit the linear mixed model with:

```
fit.ba <- lmm(pefr~method,
              repetition=~rep.method|id,
              structure=CS(~method),
              data=long)
```

Results: Bias between methods

```
Fixed effects: pefr ~ method

            estimate     se     df   lower   upper p.value
(Intercept)  447.882 28.491 16.002 387.484 508.281  <0.001
methodmini     6.029  8.053 15.996 -11.043  23.102   0.465

Degrees of freedom were computed using a Satterthwaite approximation.
```

No **evidence** of systematic differences between the two methods.

But note that:

- This doesn't neccessarily imply that devices are **equivalent**.
- We **cannot rule out a bias** within -11 to +23 l/min.
- Is this within a pre-specified **equivalence margin**?

Results: Covariance parameters

```
Residual variance-covariance: block unstructured

  - correlation structure: ~method
           1.wright 2.wright 1.mini 2.mini
    1.wright    1.000    0.983  0.948  0.948
    2.wright    0.983    1.000  0.948  0.948
    1.mini      0.948    0.948  1.000  0.968
    2.mini      0.948    0.948  0.968  1.000

  - variance structure: ~method
                   standard.deviation     ratio
    sigma.wright             117.9708 1.0000000
    sigma.mini               112.1782 0.9508982
```

Resullts: Limits of agreement

The standard deviation of the difference between two measurements with different devices is:

$$\text{SD(dif)} = \sqrt{\mathrm{SD}(M_1)^2 + \mathrm{SD}(M_2)^2 - 2 \cdot \mathrm{SD}(M_1) \cdot \mathrm{SD}(M_2) \cdot \mathrm{Cor}(M_1, M_2)}$$

E.g. for the difference between Mini and Wright:

$$\sqrt{112.17^2 + 117.97^2 - 2 \cdot 112.17 \cdot 117.97 \cdot 0.948} \simeq 75.3$$

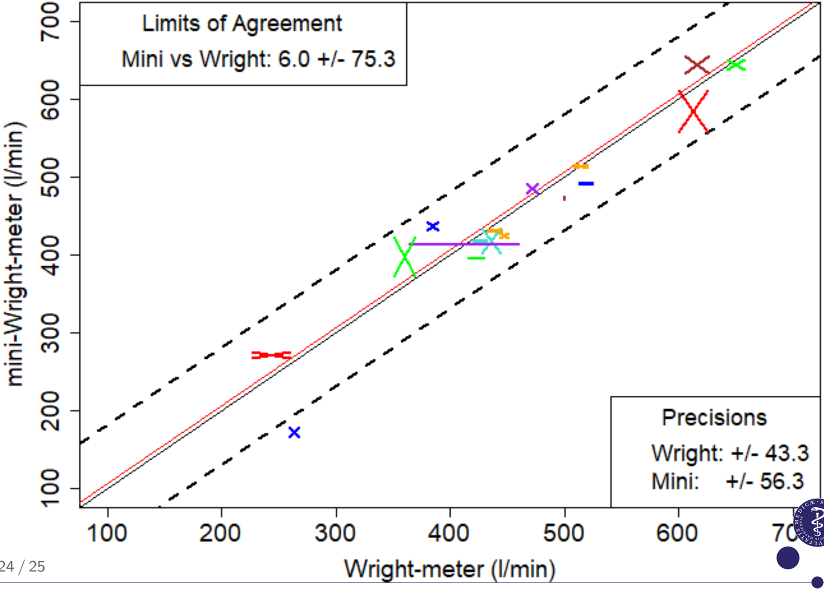
Thus we get the limits of agreement:

Mini vs Wright: 6.0 ± 75.3 l/min.

Wright vs Wright: ± 43.3 l/min.

Mini vs Mini: ± 56.3 l/min.

Case study: starplot with limits of agreement



Alternative: Two naive approaches revisited

If we only had one measurement from each device, we could make an ordinary Bland-Altman analysis. . .

2. Average the replicates ($n = 17$ averages per device):

- Average the replicates before comparing the two methods.
- Correct estimate and 95% CI for bias*.
- Too narrow limits of agreement.

3. Ignore the replicates ($n = 34$ measurements per device):

- Treat replicates as data from a new person.
- Too narrow 95% CI for the bias.
- Correct limits of agreement in all but tiny samples*.

$\star$ Assuming data is balanced and complete (i.e. same number of replicates for each test person with each method).