Faculty of Health Sciences

Variance component models

Analysis of repeated measurements, 2015

Julie Lyng Forman & Lene Theil Skovgaard

Department of Biostatistics, University of Copenhagen

Topics for today

Linear mixed models for clustered data and repeated measurements in general, i.e. not just for longitudinal data.

New concepts:

- random effects
- variance components
- multi-level models

Suggested reading:

- Fitzmaurice et al. (2011): chapters 8, 21, 22.
- Bland and Altman: *Statistical methods for assessing agreement between two methods of clinical measurement* The Lancet (1986).

Outline

Analysis of repeated measurements

Many applications:

- Longitudinal data
- Treatments applied to multiple limbs, teeth, etc within the same subject.
- Cross-over trials.
- Cluster randomized trials/multi-center studies.
- Reproducibility/reliability of measurement methods.

ATT: Measurements belonging to the same subject/cluster are correlated. If we fail to take this correlation into account we will experience:

- p-values that are too small or too large.
- confidence intervals that are too wide or too narrow.

Sources of variation / correlation

Measurements belonging to the same subject/cluster tend to be correlated (look alike) due to e.g.

- Environmental variation.
  - Between regions, hospitals or countries.
- Biological variation.
  - Between individuals, families or animals.

Today: Use random effects (variance components) to model various sources of variation in a linear mixed model framework.

Outline

One-way analysis of variance – with **random** variation

The siimplest possible model for cllustered data.

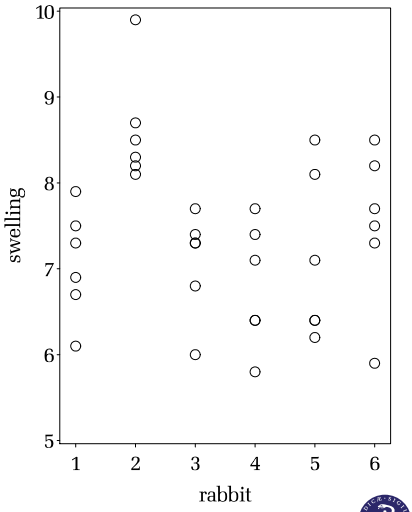
- Comparison of k groups or clusters, satisfying:
- The groups are of no individual interest and it is of no relevance to test whether they have identical means.
- The groups may be thought of as representatives from a population, that we want to describe.

Example: Rabbit data

- $R = 6$ rabbits vaccinated.
- In $S = 6$ spots on the back.

Response: swelling in cm^2

Research question:
How much swelling can be expected in reaction to the vaccine?



Random effects anova (the two-level model)

We let each rabbit have its own level of swelling described as

$$Y_{rs} = A_r + \varepsilon_{rs}$$

,

- We **assume** that these individual levels are randomly sampled from a normally distributed population,

$$A_r \sim \mathcal{N}(\mu, \omega_B^2)$$

- The error terms are considered to be independent normal,

$$\varepsilon_{rs} \sim \mathcal{N}(0, \sigma_W^2)$$

The rabbit levels are so-called random effects and the variances ω_B^2 and σ_W^2 are so-called variance components describing the variance **between rabbits** and **within rabbits**, respectively.

Implications of random effects anova

All observations are considered as randomly sampled measurements from the **same population**. Thus, the model implies that all measurements follow the same normal distribution:

$$Y_{rs} \sim N(\mu, \omega_B^2 + \sigma_W^2)$$

- Population mean μ, the grand mean.
- Population variance $\omega_B^2 + \sigma_W^2$, the total variation.

But: Measurements made on the same rabbit are correlated with the so-called intra-class correlation

$$\text{Corr}(y_{r1}, y_{r2}) = \rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2}$$

Compound symmetry

The implied covariance of the repeated measurements has a compound symmetry-structure:

$$\begin{pmatrix} \omega_B^2 + \sigma_W^2 & \omega_B^2 & \dots & \omega_B^2 \\ \omega_B^2 & \omega_B^2 + \sigma_W^2 & \dots & \omega_B^2 \\ \vdots & \vdots & & \vdots \\ \omega_B^2 & \omega_B^2 & \dots & \omega_B^2 + \sigma_W^2 \end{pmatrix}$$

In particular all pairs of spots on the same rabbit are assumed to be equally correlated (with the intra-class correlation).

Exchangeability

If any two pairs of measurements are equally correlated we say that the measurements are exchangeable.

- Are the spots randomly selected?

If this is not the case, an unstructured covariance is more apropriate

- Some spots are expected to respond more similarly than others.

In other situations with clustered data exchangeability is more obvious

- E.g. patients sampled from several GPs

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Random effects ANOVA in PROC MIXED

```
PROC MIXED DATA=rabbit;
  CLASS rabbit;
  MODEL swelling = / SOLUTION;
  RANDOM rabbit;
RUN;

Covariance Parameter Estimates

Cov Parm      Estimate

rabbit          0.3304
Residual        0.5842

                  Solution for Fixed Effects

                          Standard
Effect       Estimate       Error      DF    t Value    Pr > |t|
Intercept      7.3667      0.2670       5      27.59      <.0001
```

Estimation of variance components

Level	Variation	Variance component	Estimate	%of variation
1	Between	ω_B^2	0.3304	36%
2	Within	ω_W^2	0.5842	64%
	Total	$\omega_B^2 + \sigma_W^2$	0.9146	100%

We can use the `covtest`-option in

```
PROC MIXED COVTEST DATA=rabbit; ...
```

to get standard errors for the variance components:

- 95%CI for **Intra**-rabbit variation σ_W^2 : (0.37 – 1.04).
- 95%CI for **Inter**-rabbit variation ω_B^2 : (0.06 – 2.48).

Beware not to overinterpret the estimates in a small dataset!

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Computing variance components (technical)

In balanced data (same number of observations per cluster):

Explicit solution:

$$\tilde{\sigma}_W^2 = \mathrm{MS}_W \quad \text{and} \quad \tilde{\omega}_B^2 = \mathrm{MS}_B - \frac{\mathrm{MS}_W}{n}$$

- n is the number of observations per cluster.
- MS_W and MS_B are Mean Squares within and between clusters, defined as in one-way ANOVA.

This is deduced from $E(\mathrm{MS}_B) = n\omega_B^2 + \sigma_W^2$ and $E(\mathrm{MS}_W) = \sigma_W^2$.

Typical differences

Difference between spots on the **same** rabbit:

$$\begin{aligned} y_{rs_1} - y_{rs_2} &= \varepsilon_{rs_1} - \varepsilon_{rs_2} \\ &\sim N(0, 2\omega_W^2) \end{aligned}$$

- Normal region: $\pm 2\sqrt{2\omega_W^2} = \pm 2.16\ cm^2$

Difference between spots on **different** rabbits:

$$\begin{aligned} y_{r_1 s_1} - y_{r_2 s_2} &= \alpha_{r_1} - \alpha_{r_2} + \varepsilon_{r_1 s_1} - \varepsilon_{r_2 s_2} \\ &\sim N(0, 2\sigma_B^2 + 2\omega_W^2) \end{aligned}$$

- Normal region: $\pm 2\sqrt{2\sigma_B^2 + 2\omega_W^2} = \pm 2.70\ cm^2$

Why not use traditional one-way anova?

Focus on rabbit means: and test $H_0: \mu_1 = \ldots = \mu_6$.

One-way anova table:

	SS	df	MS=SS/df	F
Between rabbits	12.8333	$R-1 = 5$	2.5667	4.39
Within rabbit	17.5266	$R(S-1) = 30$	0.5842	
Total	30.3599	$RS-1 = 35$	0.8674	

Test for identical rabbits means: $F = 4.39 \sim F(5, 30)$, $P = 0.004$.

But: We are not interested in these particular 6 rabbits, only in rabbits in general, as a **species**! Presumably these 6 rabbits have been **randomly sampled** from the species.

One-way anova with and without random variation

Classical one-way anova

- The rabbit means μ_r are fixed parameters, - supposedly of an interest of their own.
- We say that the rabbit factor is a fixed effect.

Random effects one-way anova

- The rabbit levels A_r are considered random and their population mean μ and variance $\omega_B^2 + \sigma_W^2$ is the major interest.
- We say that the rabbit factor is a random effect.
- (If data is from a pilot study used in the planning of some trial, the intra-class correlation will also be of interest).

Comparison of modeling strategies

Quantifying overall swelling

Four strategies for estimating the grand mean (i.e. of the rabbit population).

method	estimate (s.e.)
1: forget rabbit	7.367 (0.155)
2: fixed rabbit	7.367 (0.127)
3: rabbit averages	7.367 (0.267)
4: random rabbit	7.367 (0.267)

1. We assume independence between all 36 measurements
2. We estimate the mean swelling of *exactly these* 6 rabbits by classical one-way anova
3. We analyse the sample of averages for the six rabbits (summary statistics).
4. We estimate the mean swelling of rabbits *as a species* in the random effects anova model (the correct approach)

Comments on the strategies:

1. Ignoring the clustering is wrong!
   - leads to systematic underestimation of the standard error.
2. In the fixed effect one-way anova the grand mean has a different interpretation!
   - leads to systematic underestimation of the standard error.
3. Looking at the sample of averages may be OK.
   - At least in balanced designs (otherwise the individual averages have unequal variances and the standard error may be affected)
   - But we loose all information on within subject variation. (E.g. not possible to test for systematic spot-differences.)

Comparison of modeling strategies

When the 3 smallest measurements from rabbit 2 (largest level) are omitted, the results become:

method	estimate (s.e.)
1: forget rabbit	7.291 (0.163)
2: fixed rabbit	7.291 (0.136)
3a: rabbit averages (weighted)	7.291 (0.265)
3b: rabbit averages (unweighted)	7.436 (0.333)
4: random rabbit	7.390 (0.298)
Full sample	7.367 (0.267)

1 we have omitted some of the largest observations

2+3a rabbit 2 has a lower weight in the average (only 3 observations)

3b average for rabbit 2 has increased

4 rabbit 2 has a lower weight in the average due to a larger standard error

Estimation of individual rabbit means

Sometimes estimates of individual random effects are used for e.g. prediction of future disease status.

How do we estimate them?

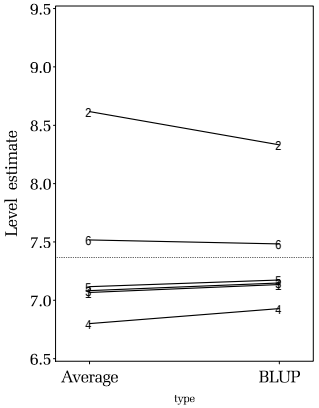
- Simple averages $\bar{y}_{r\cdot}$ of the individual measurements.
- Best unbiased linear predictors (BLUPs) are **weighted averages** of the individual and the population mean:

$$\frac{\tilde{\omega}_B^2}{\tilde{\omega}_B^2 + \frac{\tilde{\sigma}_W^2}{S}} \bar{y}_{r\cdot} + \frac{\frac{\tilde{\sigma}_W^2}{S}}{\tilde{\omega}_B^2 + \frac{\tilde{\sigma}_W^2}{S}} \bar{y}_{\cdot\cdot}$$

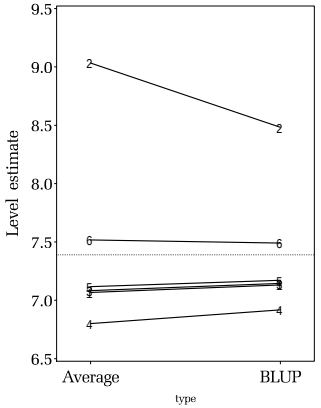
They have been **shrinked** towards the grand mean, $\bar{y}_{\cdot\cdot}$.

BLUPs vs averages

Full data



Reduced data



Note: We see larger shrinkage for rabbit no. 2 when the 3 smallest measurements from this rabbit have been removed (i.e. we are *borrowing strenght from the neighbours*).

Outline

General variance component models

Generalisations of ANOVA and GLM models involving several sources of random variation, so-called variance components.

Examples of sources of random variation:

- Environmental variation.
  - Between regions, hospitals or countries.
- Biological variation.
  - Between individuals, families or animals.
- Within-individual variation.
  - Between arms, teeth, days.
- Variation due to uncontrollable circumstances.
  - E.g. time of day, temperature, observer.
- Measurement error.

Multilevel models

Variance component models are also called multilevel models.

- Levels are most often hierarchical.
- We have variation, i.e. a variance component, on each level.
- And possibly systematic effects (covariates) on each level.

individual	$\rightarrow$	*context/cluster*	$\rightarrow$	*context/cluster*
level 1	$\rightarrow$	**level 2**	$\rightarrow$	**level 3**
students	$\rightarrow$	classes	$\rightarrow$	schools
patient	$\rightarrow$	clinic	$\rightarrow$	regions
visit	$\rightarrow$	girl	$\rightarrow$	
spot	$\rightarrow$	rabbit	$\rightarrow$	

Example: A three-level model

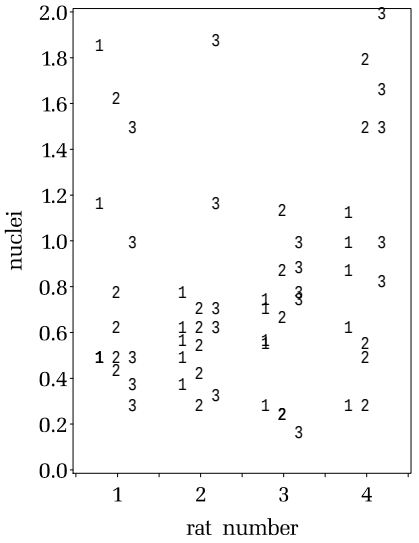
Outcome: Number of nuclei per cell in the rat pancreas (used for the evaluation of cytostatica)

- $R = 4$ rats.
- $S = 3$ sections for each rat.
- $F = 5$ randomly chosen fields from each section.

level 1	$\rightarrow$	**level 2**	$\rightarrow$	**level 3**
fields	$\rightarrow$	sections	$\rightarrow$	rats
σ^2		τ^2		ω^2

Reference: Henrik Winther Nielsen, Inst. Med. Anat.

Three-level variations



Variation	Estimate
Rats (ω^2)	0.0179 (8.2%)
Sections (τ^2)	0.0029 (1.3%)
Fields (σ^2)	0.1968 (90.4%)
Total	0.2176 (100%)

Typical differences (normal regions)

- For sections on **different rats**:

$$\pm 2 \times \sqrt{2 \times (0.0179 + 0.0029 + 0.1968)} = \pm 1.319$$

- For **different sections** on the **same rat**:

$$\pm 2 \times \sqrt{2 \times (0.0029 + 0.1968)} = \pm 1.264$$

- For **different fields** on the **same section**:

$$\pm 2 \times \sqrt{2 \times 0.1968} = \pm 1.255$$

Correlation

Estimated correlations between two measurements on the same rat:

- If they are measured on **the same section**:

$$\text{Corr}(y_{rs1}, y_{rs2}) = \frac{\omega^2 + \tau^2}{\omega^2 + \tau^2 + \sigma^2} = 0.096.$$

- If they are measured on **different sections**:

$$\text{Corr}(y_{r11}, y_{r22}) = \frac{\omega^2}{\omega^2 + \tau^2 + \sigma^2} = 0.082.$$

Merits of multilevel models

We get a better understanding of the various sources of variation.

Effects *within* may be estimated more precisely (higher power), since some sources of variation are eliminated, e.g. by making comparisons within a family. This is analogous to the **paired comparison** situation.

When planning investigations, estimates of the variance components are needed in order to compare the power of various designs, and help us decide

- How many replicates do we need at each level?
- Should we randomize entire clusters or randomize *within* the clusters?

Design considerations

(**Note** in analogy with cluster-randomized trials.)

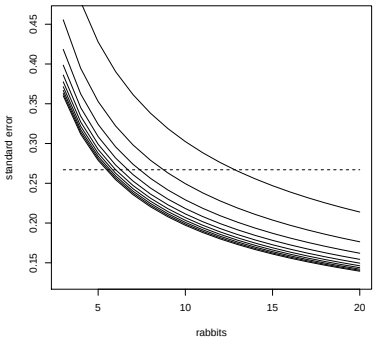
Plan an experiment with:

- R rabbits.
- S spots for each rabbit.
- $R \times S$ measurements.

Std. error of grand mean,

$$\text{var}(\bar{y}) = \frac{\omega_B^2}{R} + \frac{\sigma_W^2}{RS},$$

decreases with R and S.



The different curves correspond to S varying from 1 to 10.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Effective sample size

How many rabbits would we need to obtain the same precision in estimating the grand mean if we had **only one measurement** on each of R_1 rabbits?

Solve an equation to get:

$$R_1 = \frac{R \times S}{1 + \rho(S-1)}$$

where ρ is the within rabbit correlation.

- Estimate: $\rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2} = \frac{0.3304}{0.3304+0.5842} = 0.361 \Rightarrow R_1 = 12.8$

I.e. one measurement on each of thirteen rabbits gives the **same precision** as six measurements on each of six rabbits.

Drawbacks of multilevel models

Their statistical analysis is more difficult.

- When making inference (estimation and testing), it is important to take all sources of variation into account, and effects have to be evaluated against the relevant variation.

If we fail to take the correlation into account, we will experience:

- Possible bias in the mean value estimates.
- Too small standard errors (type 1 error) for estimates of level 2 covariates (between-cluster effects).
- Too large standard errors (type 2 error) for estimates of level 1 covariates (within-cluster effects)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Outline

Fixed or random effect?

Fixed effects such as treatment, gender, and time.

- Typically a limited number of carefully selected groups.
- Group names are specific and cannot be shuffled.
- Each group must have a decent size in order to reach interesting conclusions (statistical power).

Random effect such as subject, rat or familly.

- Possibly a large number of different groups.
- Group names are non-informative (number of subject, rat or family) and could be shuffled without consequence.
- Allows inference to be extended beyond the subjects in the experiment and to the population they were sampled from.
- The number of groups matters not the size of the groups.

Testing fixed effects

Imagine that rabbits are grouped in two (e.g. treatments):

level	variation	covariates
1	within rabbit	spot
2	between rabbits	group

- Part of the variation *between rabbits* could be explained by systematic differences between groups.
- Part of the variation *within rabbits* could be explained by systematic differences between spots.

Testing fixed effects with PROC MIXED

```
PROC MIXED DATA=rabbit;
  CLASS group rabbit spot;
  MODEL swelling = group spot / SOLUTION CL DDFM=KR;
  RANDOM rabbit;
RUN;
```

Output:

```
Covariance Parameter Estimates

Cov Parm          Estimate

rabbit              0.3694  <-------------- smaller than before
Residual            0.5477  <-------------- smaller than before
```

Testing fixed effects with PROC MIXED

```
Type 3 Tests of Fixed Effects

              Num     Den
Effect         DF      DF    F Value    Pr > F

group           1       4       0.64    0.4675
spot            5      25       1.40    0.2584
```

```
Solution for Fixed Effects

Effect     spot  group  Estimate  StdError    DF  t Value  Pr > |t|   Alpha    Lower    Upper

Intercept              6.9111    0.4792     4    14.42    0.0001    0.05   5.5807   8.2416
group          1       0.4444    0.5542     4     0.80    0.4675    0.05  -1.0942   1.9831
group          2            0         .     .        .         .       .        .        .
spot       a           0.6500    0.4273    25     1.52    0.1408    0.05  -0.2300   1.5300
spot       b          0.05000    0.4273    25     0.12    0.9078    0.05  -0.8300   0.9300
...
```

Disregarding repeated measurements

When the **random rabbit variation** is ignored:

```
PROC GLM DATA=rabbit;
  CLASS group spot;
  MODEL swelling=group spot / SOLUTION CLPARM;
RUN;
```

```
Source                     DF      Type III SS      Mean Square    F Value    Pr > F

group                       1      1.77777778      1.77777778       2.08    0.1596
spot                        5      3.83333333      0.76666667       0.90    0.4954
```

```
                                                    Standard
Parameter            Estimate                          Error    t Value    Pr > |t|      95% Confidence Limits

Intercept       6.911111111 B                     0.40735835      16.97      <.0001     6.077969737  7.744252485
group      1    0.444444444 B                     0.30793397       1.44      0.1596    -0.185351236  1.074240125
group      2    0.000000000 B                              .          .           .               .            .
spot       a    0.650000000 B                     0.53335728       1.22      0.2328    -0.440838117  1.740838117
spot       b    0.050000000 B                     0.53335728       0.09      0.9260    -1.040838117  1.140838117
...
```

Too small standard errors for estimates of difference between groups and too large standard errors for estimates of differences between spots!

Outline

Example: Cortisol and stress-response

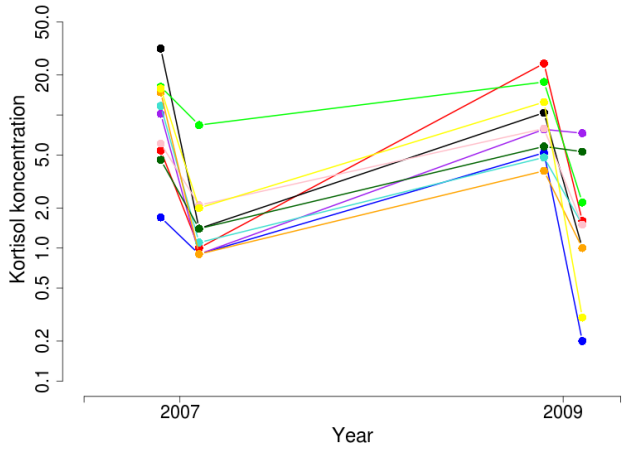
Outcome: Concentration of cortisol in blood samples taken morning and evening in workers in Aarhus amt and kommune in 2007 (3536 participants) with similar follow-up in 2009 (2408 participants)

Interest: effect of stressors: lifeevents, Effort Reward Index

level	variation	covariates
3	between persons	`gender, age`
2	within person: between days	`bmi, stressors`
1	within person: within days	`time (of day)`

Sample data

From 8 randomly selected men:



NOTE: concentrations on logarithmic scale.

Multi-level analysis

```
PROC MIXED DATA=prism COVTEST; WHERE sex EQ 'male';
  CLASS id year time;
  MODEL logcortisol = time  / SOLUTION CL DDFM=SATTERTH;
  RANDOM id id*year;
RUN;
```

Covariance Parameter Estimates

Cov Parm	Estimate	Std.Error	Z Value	Pr > Z
id	0.05993	0.01266	4.73	<.0001
id*year	0	.	.	.
Residual	0.5385	0.01794	30.01	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
time	1	1305	4916.89	<.0001

One of the variance component estimates is a **zero**!

Negative variance components

In case on of the variance component estimates becomes negative, SAS repports a zero.

What does it mean?

- The zero-estimate may be a chance finding due to statistical uncertainty.
- Or it might be the result of truly negative correlation within clusters - e.g. from competition (plants grown in same pot).

What can we do about it?

- Re-fit the model without the problematic random effect.
- Use a covariance pattern model which allows for negative correlation
- Include more covariates at the lower levels.

Estimated variance components

Level	Variation	Estimate
3	between persons (ω^2)	0.0599 (10.0%)
2	between days (τ^2)	0.0000 (0.0%)
1	within days (σ^2)	0.5385 (90.0%)
	Total	0.5984 (100%)

Level 2 covariates (stressors) can only have **very little impact on individual cortisol koncentrations**!

Systematic effects

Solution for Fixed Effects

Effect	time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept		0.4106	0.02209	448	18.59	<.0001	0.05	0.3672	0.4540
time	morn	2.0137	0.02872	1305	70.12	<.0001	0.05	1.9573	2.0700
time	even	0	.	.	.	.	.	.	.

Cortisol is measured on **log-scale**. Backtransformation $\exp(2.0137) \simeq 7.49$ yields that median levels of kortisol is an estimated 7.5 times higher in the morning than in the evening.

Exact time of measurement should be taken into account!!!

Outline

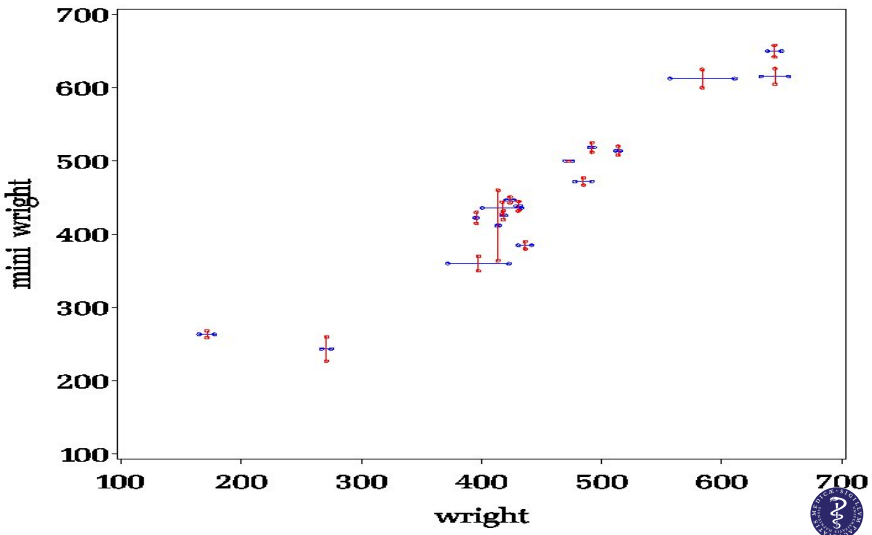
Comparing measurement devices

Example: Peak expiratory flow rate, l/min:

- 17 subjects, 2 measurement devices,
- two replicates with **each method**.

subject	*Wright*		*mini Wright*	
id	Y_{1p1}	Y_{1p2}	Y_{2p1}	Y_{2p2}
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443
Average	450.35	445.41	452.47	455.35
SD	116.31	119.61	113.12	111.32

Reference: Bland and Altman, Lancet (1986).



Aim of investigation

Quantify the **precision** of each measuring device

- Repeatability (variability=measurement error)

Quantify the **agreement** between the two devices.

- Bias of one method compared to the other.
- Variance of one method compared to the other.

Can the devices be used interchangably?

Simple approaches

For reliability of **each method separately** we could:

- make Bland Altman plots of differences vs averages.
- compute limits of agreement, i.e. the 95% normal range of the differences.

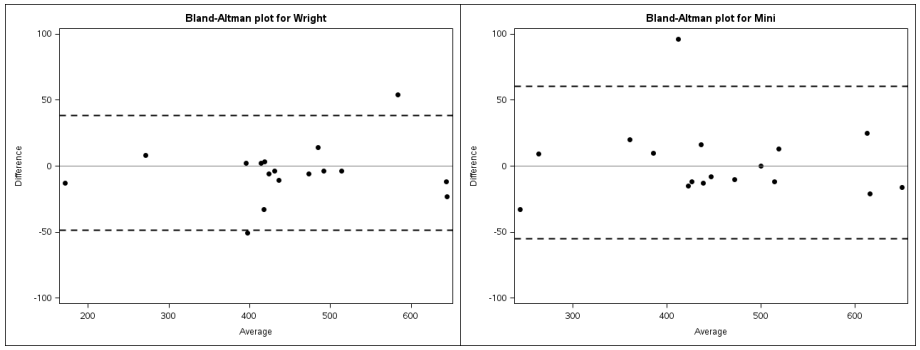
For reproducibility (method comparison) we might:

- compare the **averages** in a Bland-Altman plot?
- Not good - unless you also do averages in clinic!

For both at the same time:

- Mixed model for variance between and within methods.

Repeatability



No evidence of bias in either case

- **paired t-test:** P=0.36 for wright and P=0.68 for mini.

Two-level models

For each method $(i = 1, 2)$ we have a two-level model

$$Y_{ijk} = \mu_i + a_{ij} + \varepsilon_{ijk}$$

- μ_i population mean as anticipated by method i.
- a_{ij} deviation of subject j from population mean, assumed normally distributed $N(0, \sigma_i^2)$.
- ε_{ijk} deviation for replicate k (measurement error), assumed normally distributed $N(0, \omega_i^2)$.

PROC MIXED: Stratified analyses

```
PROC MIXED DATA=wright;  BY method;
CLASS id;
MODEL flow = / SOLUTION CL;
RANDOM id;
RUN;
```

```
method=mini

Cov Parm      Subject    Estimate
Intercept     id            12188
Residual                   396.44

Effect       Estimate       Error      DF    t Value    Pr > |t|     Alpha       Lower       Upper
Intercept      453.91     26.9921      16      16.82      <.0001      0.05      396.69      511.13


method=wright

Cov Parm      Subject    Estimate
Intercept     id            13683
Residual                   234.29

Effect       Estimate       Error      DF    t Value    Pr > |t|     Alpha       Lower       U
Intercept      447.88     28.4914      16      15.72      <.0001      0.05      387.48      50
```


Joint model for both methods

For methods $(i = 1, 2)$:

$$Y_{ijk} = \mu_i + a_{ij} + \varepsilon_{ijk}$$

- ε_{ijk} assumed normally distributed $N(0, \omega_i^2)$ and **independent across methods**.
- a_{ij} assumed normally distributed $N(0, \sigma_i^2)$ and correlated with $\rho = \mathrm{Cor}(a_{i1}, a_{i2})$.

Anticipated means for the same subject ought to look a lot like each other, so the a_{ij}'s are likely to be correlated across methods.

- Note that SAS models the covariance parameter $\sigma_{12} = \mathrm{Cov}(a_{1j}, a_{2j}) = \sigma_1 \cdot \sigma_2 \cdot \rho$.

PROC MIXED: Joint analysis

```
PROC MIXED DATA=wright;
CLASS method id;
MODEL flow=method / SOLUTION CL;
RANDOM method / TYPE=UN SUBJECT=id;
REPEATED / TYPE=simple GROUP=method SUBJECT=id*method;
RUN;
```

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate
UN(1,1)	id		12188
UN(2,1)	id		12542
UN(2,2)	id		13683
Residual	method*id	method mini	396.44
Residual	method*id	method wright	234.29

Solution for Fixed Effects

Effect	method	Estimate	StdError	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept		447.88	28.4914	32	15.72	<.0001	0.05	389.85	505.92
method	mini	6.0294	8.0532	32	0.75	0.4595	0.05	-10.3744	22.4332
method	wright	0	.	.	.	.	.	.	.

Repeatability

Typical differences (approximate 95% normal range) between two measurement with the **same method**:

$$\text{Wright: } \hat{\omega}_1^2 = 234.29 \;\rightarrow\; \pm 2\sqrt{2\omega_1^2} \simeq \pm 43.3$$

$$\text{Mini: } \hat{\omega}_2^2 = 396.44 \;\rightarrow\; \pm 2\sqrt{2\omega_2^2} \simeq \pm 56.3$$

Seemingly Wright is more precise, but is the difference significant?

$$F = \frac{396.44}{234.29} = 1.69 \sim F(17,17) \rightarrow P = 0.14$$

Don't form too firm a conclusion with too small data.

Reproducibility

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

- Estimated bias +6.0 (-10.4;22.4) for mini vs wright. P=0.46.

Typical differnces between the two methods:

$$\begin{aligned}\text{var}(Y_{1jk} - Y_{2jk}) &= \text{var}(a_{1j} - a_{2j} + \varepsilon_{1jk} - \varepsilon_{2jk}) \\ &= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} + \omega_1^2 + \omega_2^2 \\ &= 12188 + 13683 - 2 \cdot 12542 + 396.44 + 234.29 \\ &= 1417.73\end{aligned}$$

Limits-of-agreement: $6.03 \pm 2\sqrt{1417.7} = (-69.3, 81.3)$.

Multi-level model?

level	variation	covariates
3	between subjects (ω^2)	
2	between methods (τ^2)	method
1	within methods (σ^2)	

Specified as:

$$Y_{ijk} = \mu_j + a_i + b_{ij} + \varepsilon_{ijk}$$

- $A_i \sim \mathcal{N}(0, \omega^2)$ for subjects $i = 1, \ldots, 17$,
- $B_{ij} \sim \mathcal{N}(0, \tau^2)$ for methods $j = 1, 2$,
- $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ for replicate $k = 1, 2$.

Assuming the same variance for both methods!

Estimated variance components

```
PROC MIXED DATA=wright;
  CLASS method id;
  MODEL flow=method / SOLUTION CL;
  RANDOM intercept method / SUBJECT=id;
RUN;
```

```
 Covariance Parameter Estimates
Cov Parm      Subject    Estimate
Intercept     id            12542
method        id           393.57
Residual                   315.37

          Fit Statistics
-2 Res Log Likelihood          676.0
AIC (smaller is better)        681.6
```

What does this tell us about the precision of the measurements?

Typical differences

Between replicate measurements using the same method:

$$\begin{aligned} Y_{ijk_1} - Y_{ijk_2} &= \varepsilon_{ijk_1} - \varepsilon_{ijk_2} \\ &\sim \mathcal{N}(0, 2\sigma^2) \end{aligned}$$

Limits-of-agreement: $\pm 2\sqrt{2\sigma^2} \simeq \pm 50.23.$

Between measurements using the different methods:

$$\begin{aligned} Y_{ij_1k_1} - Y_{ij_2k_1} &= \mu_{j_1} - \mu_{j_2} + b_{ij_1} - b_{ij_2} + \varepsilon_{ij_1k_1} - \varepsilon_{ij_2k_1} \\ &\sim \mathcal{N}(\mu_{j_1} - \mu_{j_2}, 2\tau^2 + 2\sigma^2) \end{aligned}$$

Limits-of-agreement: $\mu_1 - \mu_2 \pm 2\sqrt{2\tau^2 + 2\sigma^2} \simeq 6.03 \pm 75.31.$

(where we include the non-significant systematic difference).

Systematic difference?

```
                    Solution for Fixed Effects

                                 Standard
Effect      method    Estimate      Error      DF    t Value    Pr > |t|

Intercept             447.88      27.7519      16      16.14      <.0001
method      mini      6.0294       8.0532      16       0.75      0.4649
method      wright         0            .       .          .           .
```

Conclusion: No evidence of **systematic** differences between the measurement methods.

BUT: Do we really want to assume that variances are equal when the power for testing if they are is low?