Faculty of Health Sciences

Analysis of cross-over studies

Statistical analysis of repeated measurements 2024

Julie Forman

Section of Biostatistics, University of Copenhagen

Course day 4 contents

Part I. Cross-over studies

- Considerations 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

Case study: An experiment with potassium

Study with $n = 25$ healthy volunteers randomized to either

- AB Four months of potassium supplement (period 1) followed by two month wash-out and four months of placebo (period 2) .
- BA Four months of placebo (period 1) followed by two month wash-out and four months of potassium supplement (period 2) .

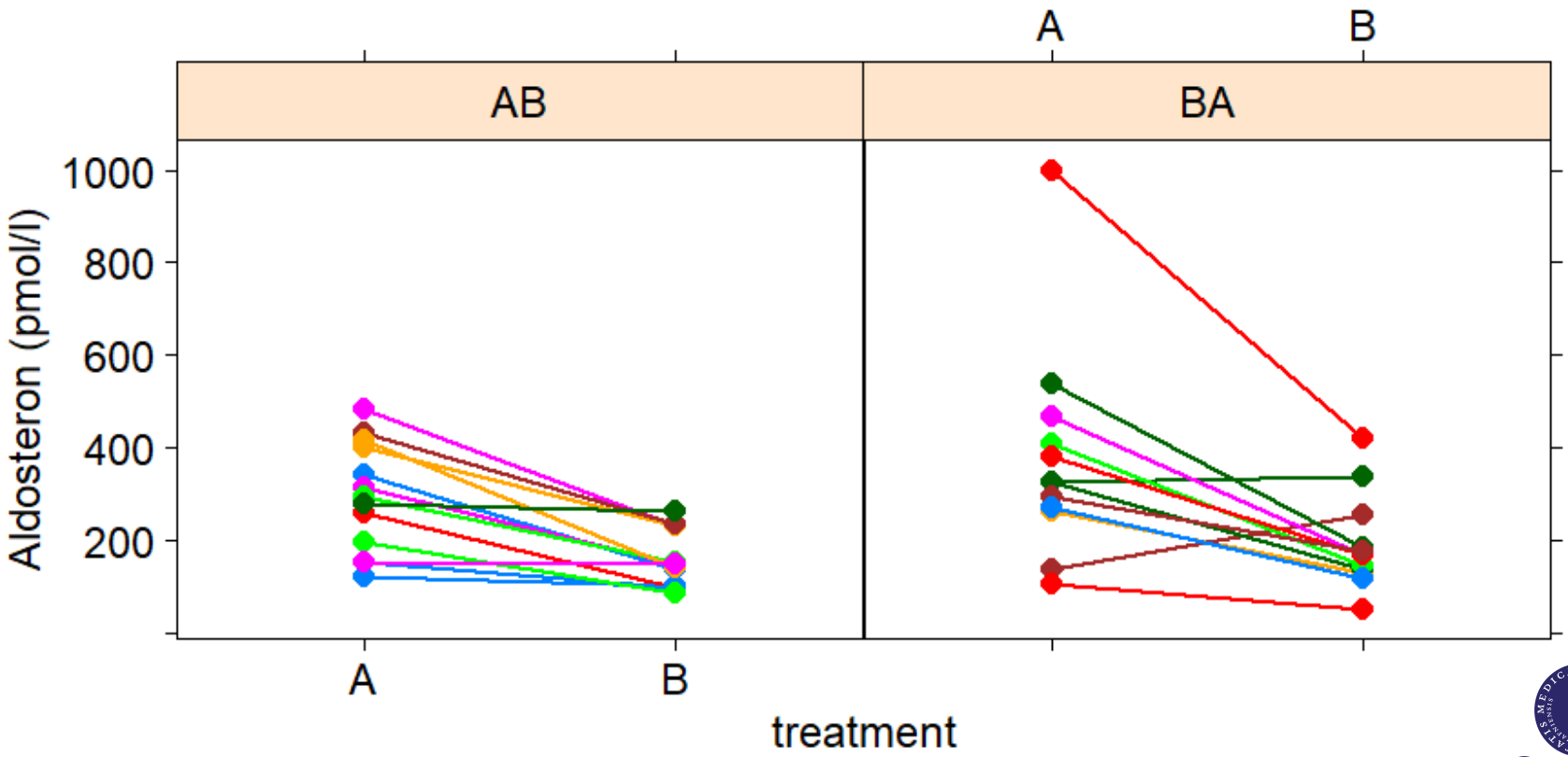
Several outcomes were collected at the end of each treatment period, but today we consider only serum aldosteron (pmol/l) measured before and during angiotensin stimulation.

Reference: Dreier et al. (2021): *Effect of increased potassium intake on the renin-agiotensin-adeosterone system and subcutaneous resistance arteries: a randomized crossover study*, Nephrology, Dialysis, Transplantation 36.

Case study: Single measure

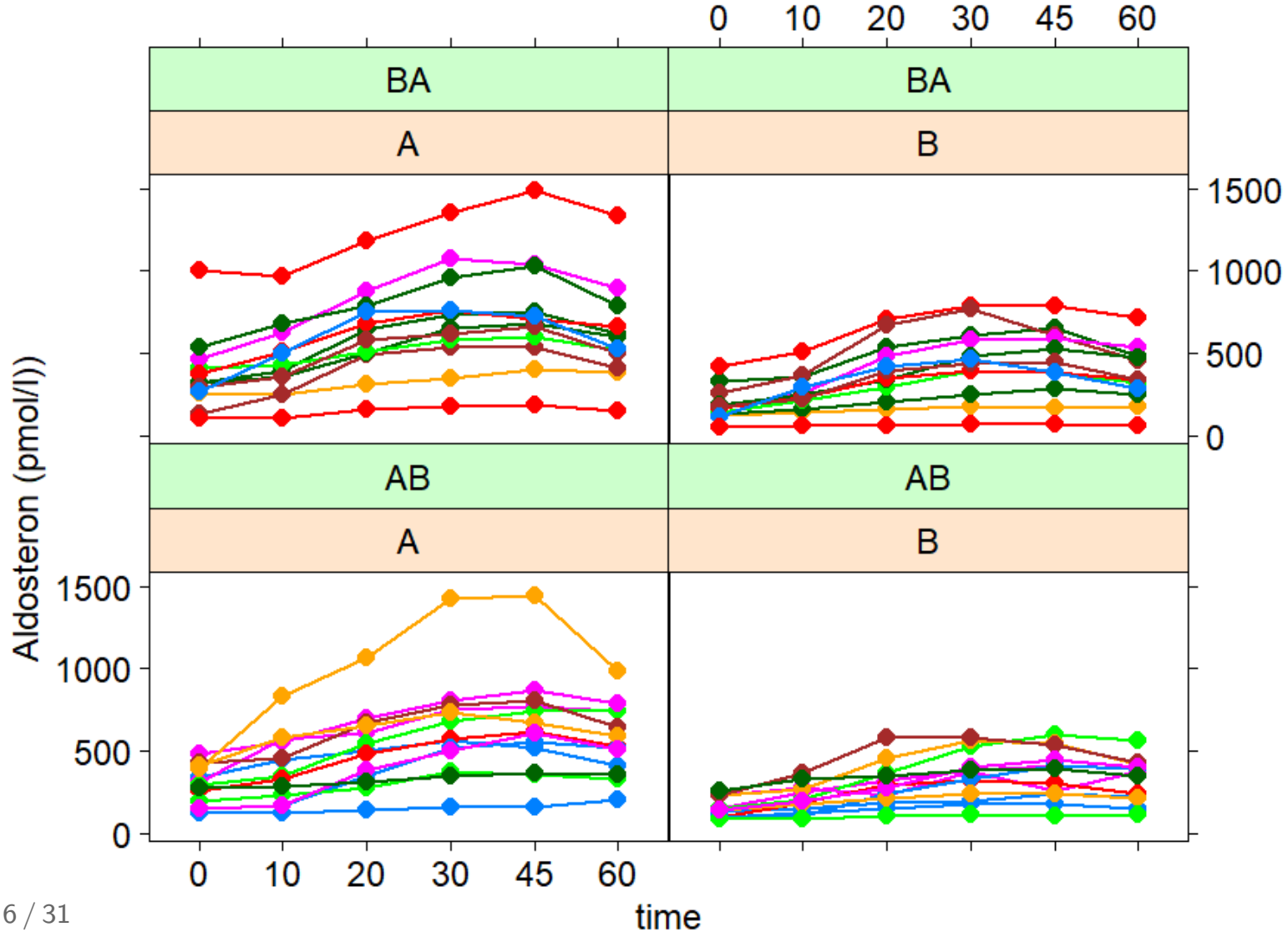
Serum concentration of aldosteron (pmol/l) after four months of:

- **A:** potassium supplement
- **B:** placebo



Case study: Repeated measurements

. . . and in response to 30 minutes angiotensin stimulation.



Why use a cross-over design?

When using a paired study design, we get a **higher statistical power** from **fewer testpersons** compared to a two-sample design.

We can make use of this by applying **different treatments** to e.g. the left and right eye of the same patient **at the same time**.

However, this only works with treatments that have a **local effect**. We *cannot* apply two treatments that have a **global effect** simultaneously, since then their effects cannot be **separated**.

But we *can* apply two treatments that have a **temporary effect** to the same person **at two different time points**. That way, we can separate the effects as long as the **time between treatments** is sufficiently long that the previous effect has **washed-out**.

Considerations about design and statistical analysis

Fixed vs randomized sequence?

- **Recommended choice:** Randomized.

Single measures or repeated measurements from each period?

- We recommend reducing the repeated measurements into **summary statistics** to simplify results and computations.

Should we adjust for period effect?

- **Recommended choice:** Yes.

Should we test for carry-over effect?

- **Recommended choice:** No.

Should we use a t-test or a linear mixed model?

- **It depends** on sample size, on whether there are missing data, and for what reason they are missing (see lecture 6).

Fixed vs randomized sequence

Fixed sequence (sometimes reasonable):

- All subjects receive treatments in the **same order**, A→B→...
- **Limitation:** Cannot separate the treatment and period effects.
- Analyze as single group follow-up study (course day 1).

Randomized sequence (recommended):

- Subjects are **randomized** to **different orders**, e.g. AB or BA.
- **Advantage:** We can separate treatment and period effects.
- With more than two treatments we get many orderings, e.g. ABC, ACB, BAC, BCA, CAB and CBA with three treatments.
- If only some sequences are included, there may be problems separating treatment and period effects (see exercise 4).

Single measures vs repeated measurements

Most cross-over studies collect single measures:

- That is one of each outcome from each period.
- Statistical analysis is straightforward.

But some collect repeated measurements:

- E.g. Series of measurements from each period.
- It it not clear how to model these series optimally:
  - Which interactions should be included as fixed effects?
  - Which covariance pattern is apropriate?
- If you have all but a few repetitions, my recommendation is to **reduce to summary statistics** (e.g. AUC).
- Ideally, summary statistics and model should be specified in the protocol to prevent manipulation of the evidence.

Period effect

EMA and FDA guidelines recommend adjusting for a potential period effect when analysing crossover studies.

Differences between periods ought to be random when the washout is sufficient and the disease is in steady state. Nevertheless, we sometimes find a period effect due to:

- Seasonal variation.
- Differences in calibration.
- Compliance issues (worrysome, may affect treatment effect).

In case there is a period effect:

- You gain a bit of statistical power by adjusting for it.

(and if there is no period effect, adjusting for it is harmless).

Carry-over effect

If the treatment effect differs between the two periods, then we have an **interaction** between the treatment and the period effect.

- This is often interpreted as a **carry-over effect**.
- I.e. an enhanced or diminished effect of a treatment caused by the treatment(s) from the previous period(s).

BUT: Crossover studies aim at an **overall** treatment effect.

- This is what the studies are **powered** for.
- Not various treatment effects depending on what treatment patients did or did not receive previously.

Carry-over effects should be **ruled out** in the **planning stage** by ensuring a **sufficiently long wash-out** between the treatments.

- If you cannot rule out a carry-over effect, then you should use a different design for your study.

Test of carry-over effect

Historically, statisticians have recommended testing for carry-over effect before evaluating the treatment effect.

BUT: Studies are usually not powered for testing the interaction.

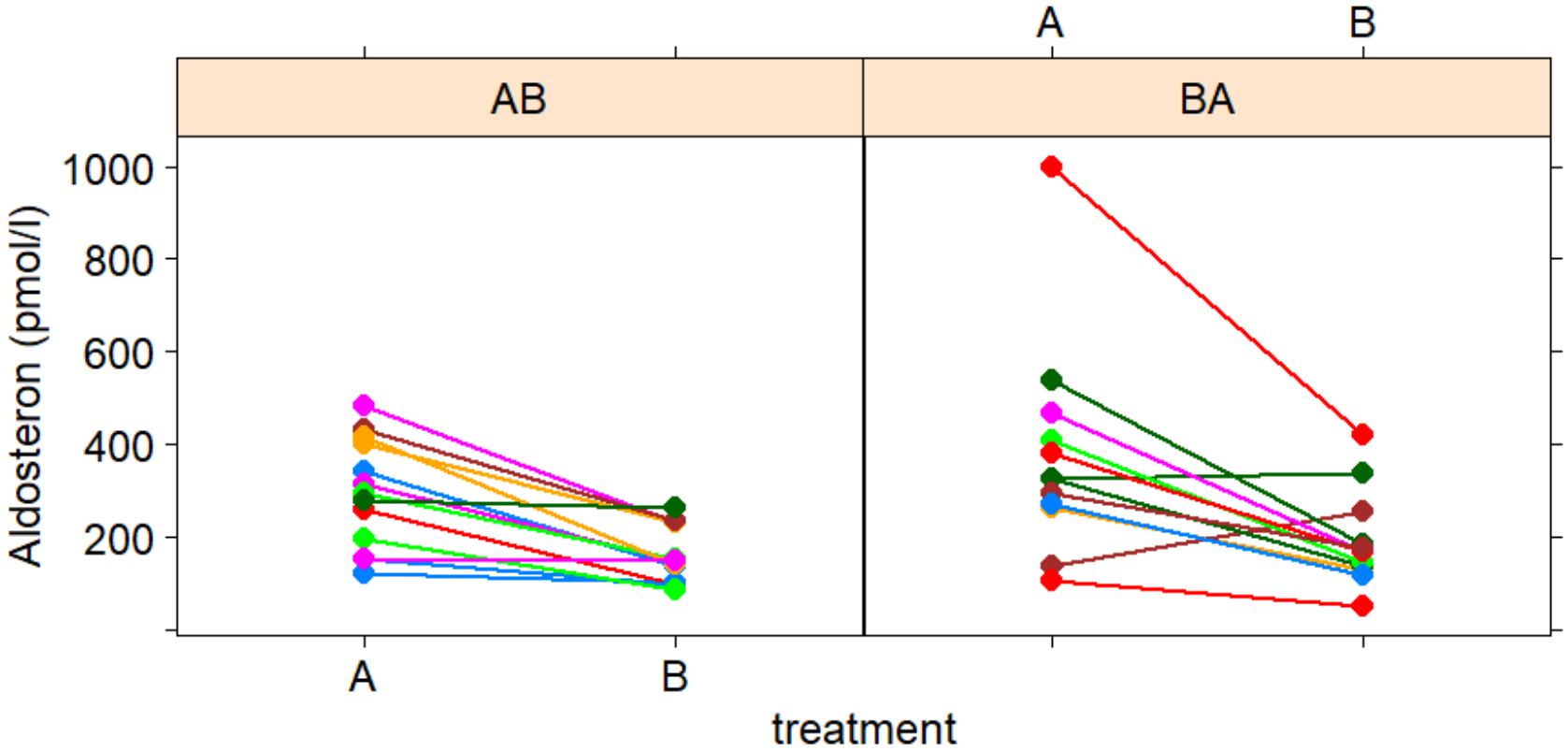
- If there is a genuine carry-over effect it is likely that the test **won't find it** (high risk of type II error)
- If there is no carry-over effect, one in twenty tests will come out with a **false positive** (5% risk of type I error).

EMA and FDA guidelines **dis**recommends testing for carry-over.

Outline

Case study: Single measures

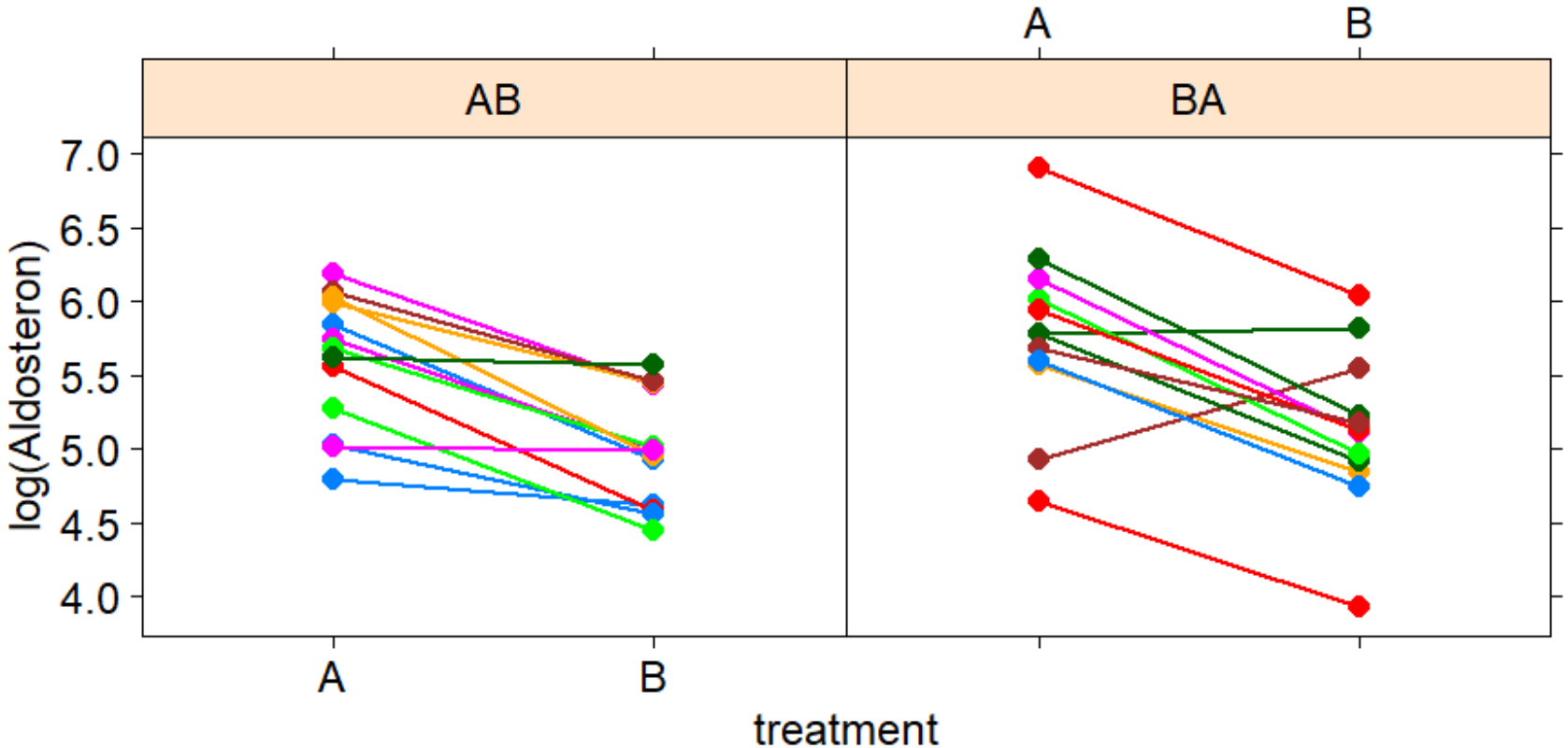
A) potassium supplement, **B)** placebo.



Concentrations have a **skew distribution** and this has also been found by other studies in the litterature.

Case study: Data on log-scale

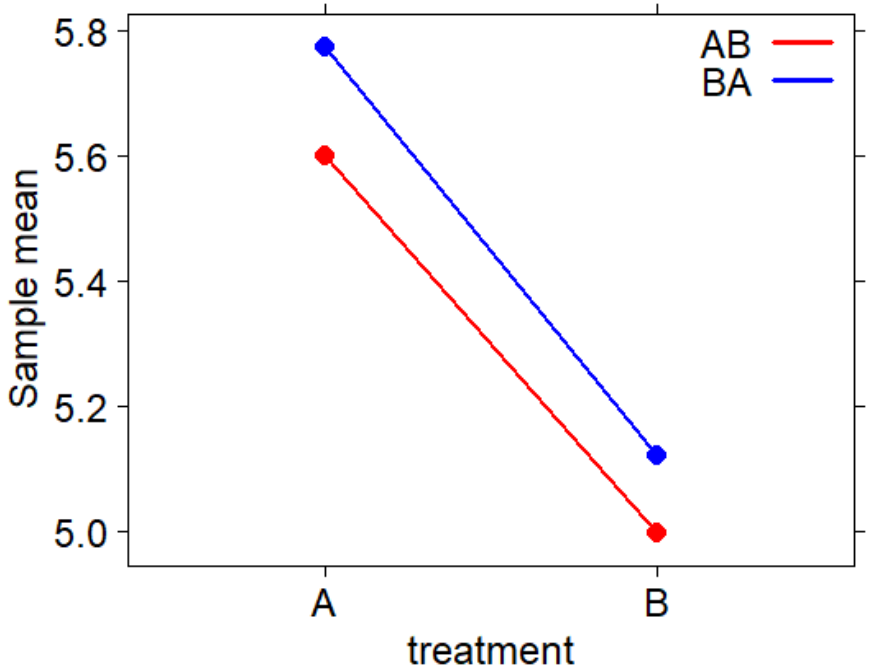
After log-transformation:



The distribution of the concentrations looks more symmetrical.

Trends in summary statistics

treatment -period		N	Mean	SD
AB	A-1	13	5.60	0.45
	B-2	13	5.00	0.38
BA	A-2	12	5.77	0.59
	B-1	12	5.12	0.54



- Means tend to be higher with treatment A compared to B, and somewhat higher for sequence BA compared to AB
- Same trend in the SDs.
- No apparent effect of period (curves are parallel).

Model for single measures

Fixed effects:

- `treatment` and `period`
- **NOT** the `treatment*period`-interaction.

Covariance structure:

- Potentially outcomes could be more variable with one treatment than with the other.
- If we had more than two treatments, outcomes from some particular treatments could be stronger correlated than others.
- Hence we assume an unstructured covariance pattern with `treatment` as the replication factor.

Two-way ANOVA like model

Systematic effect of `treatment` and `period` and **no interaction**:

Mean	treatment=B (placebo)	treatment=A (potassium)
period=1	α	$\alpha + \beta$
period=2	$\alpha + \gamma$	$\alpha + \beta + \gamma$

Default regression parameters:

- Intercept (mean with placebo in period 1, reference point).
- Effect of potassium (assumed the same in both periods)
- Effect of period (assumed the same for both treatments)

Individual deviations from the population mean are assumed 2D-normally distributed with possibly treatment dependent variances.

R-code and output

```
fit.main <- lmm(logaldo~period+treatment,
                repetition=~treatment|id,
                structure="UN",
                data=long)
```

```
Fixed effects: logaldo ~ period + treatment

            estimate    se     df  lower upper p.value
(Intercept)    5.049 0.102  30.81   4.84 5.258  <0.001 ***
period2        0.015 0.085 22.758 -0.161 0.191   0.861
treatmentA     0.627 0.085  23.01  0.452 0.802  <0.001 ***
```

Estimated treatment effect (ratio between geometric means):

- $\exp(0.627) \simeq 1.87$ (95% CI: 1.57-2.23).

more output

The cross-over study attains higher power with fewer subjects compared to a two-sample design, but only if the correlation between the paired outcomes is reasonably high.

```
Residual variance-covariance: unstructured

  - correlation structure: ~0 + treatment
        B     A
    B 1.000 0.755
    A 0.755 1.000

  - variance structure: ~treatment
            standard.deviation ratio
    sigma.B              0.542 1.000
    sigma.A              0.502 0.927
```

- Was the study design succesful?

Alternative approach based on t-test

Compute the difference in outcomes from period 1 and 2 and compare these between AB and BA using Students t-test.

- **AB:** Mean$(\Delta) = (\alpha + \beta) - (\alpha + \gamma) \; = \; \beta - \gamma$
- **BA:** Mean$(\Delta) = \alpha - (\alpha + \beta + \gamma) \; = \; -\beta - \gamma$
- The difference in mean Δ's is 2β (2 x treatment effect).

```
> t.test(delta~sequence, data=wide, var.equal=TRUE)

  t = 7.4018, df = 23, p-value = 1.585e-07

  95 percent confidence interval:
   0.9039391 1.6051951
```

- 0.9039/2=0.452 and 1.6052/2=0.802 (same as `lmm`)

Mixed model vs t-test

In case the data is complete (no missing values):

- You get the same results unless sample size is so small that the mixed model approximation to the degrees of freedom is poor.

In case there are missing data:

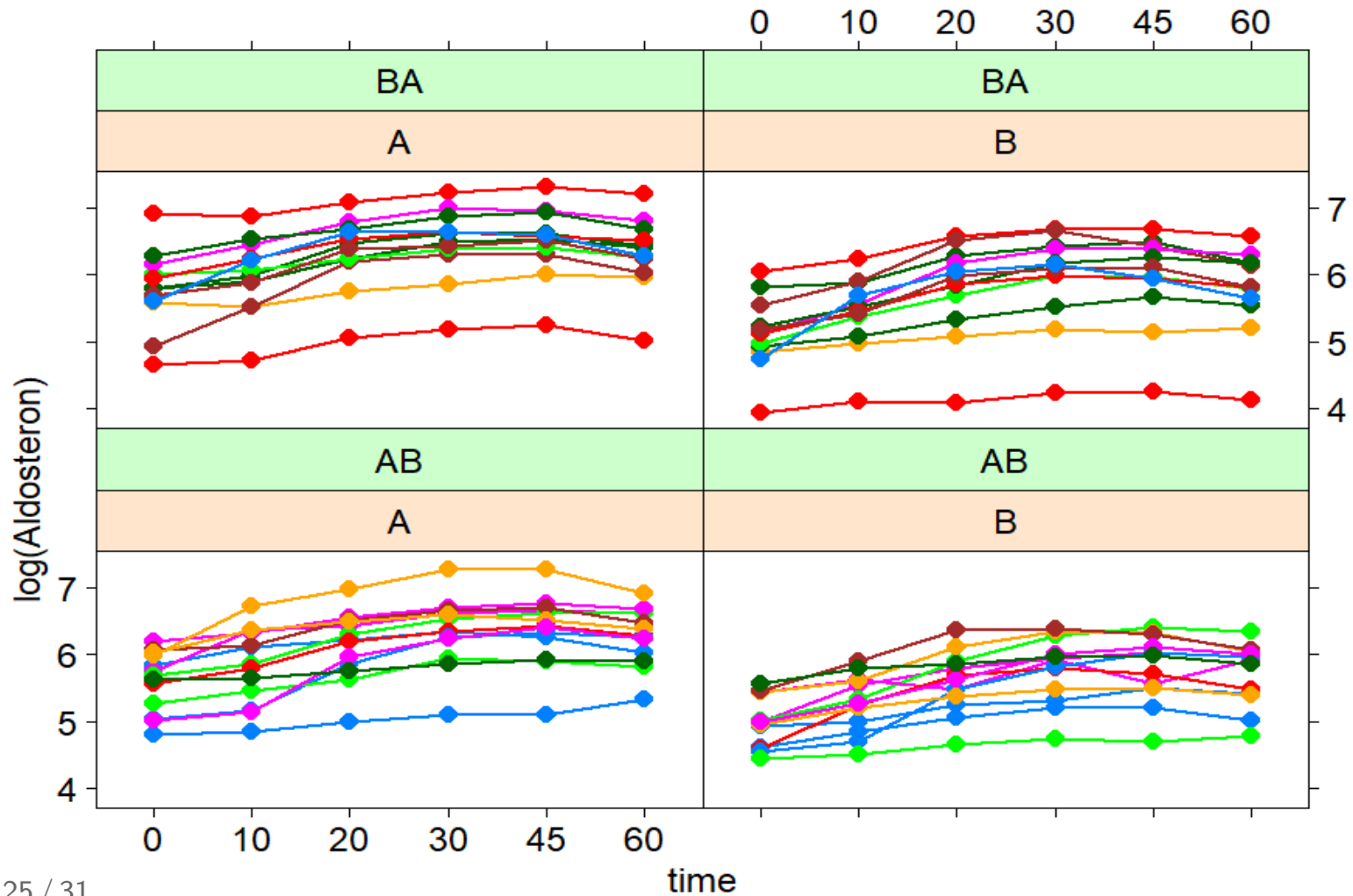
- The t-test only use complete cases which may be sub-optimal.
- The mixed model makes more optimal use of the data when missing values are *missing at random* (see lecture 6).
- Nevertheless, complete case is often a sensible choice for the primary analysis in a randomized crossover trial:

  We compare the treatment effect in those patients who are able to complete both treatments according to protocol.

Outline

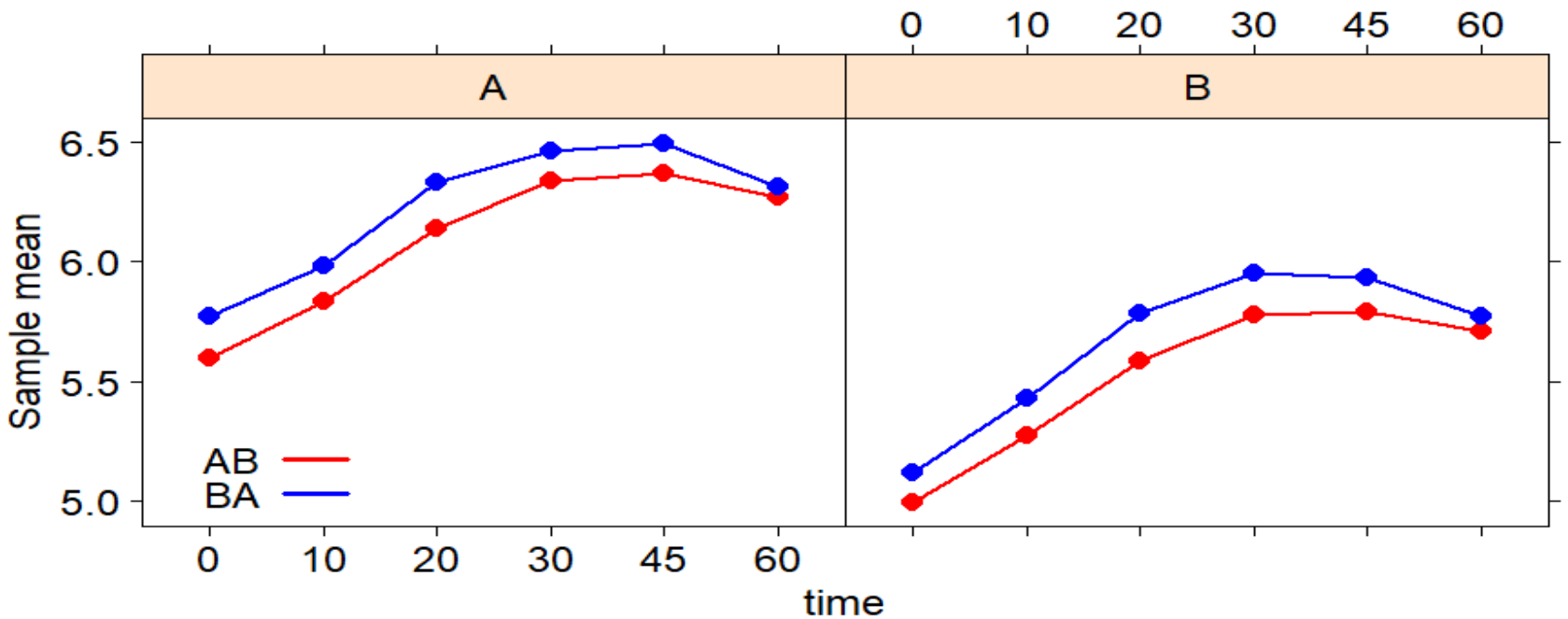
Case study: Repeated measurements

Log-transformed concentrations of aldosteron in response to 30 minutes angiotensin stimulation:



Trends in summary statistics

Compute sample means for each combination of `time` and `treatment` within each `sequence` gruop:



Seeming effects of: `time` and `treatment`.

No obvious effects of: `period` or `time*treatment` interaction.

Model for repeated measurements

Fixed effects:

- `period`, `treatment`, `time`, and `treatment*time`
  (we are interested in how treatment affects the time effect).

Covariance:

- We expect that observations within the same treatment-period are more strongly correlated when they are close in time.
- We also expect correlation between the paired observations across the two periods, but likely not as strong.
- Variances may differ between treatment and time points.
- **Use an unstructured covariance pattern!**
  (Note that `lmm` takes a while to run with 12 replicates).

Output: Estimates on log-scale

```
Fixed effects: logaldo ~ period + treatment + time + treatment:time

                    estimate    se   df  lower upper  p.value
(Intercept)            5.03 0.096 29.1  4.834 5.226  < 2e-16 ***
period2               0.052 0.047  3.3 -0.091 0.195    0.346
treatmentA            0.628 0.083 23.1  0.456 0.801 1.18e-07 ***
time10                0.292  0.04 22.9  0.209 0.374 2.09e-07 ***
time20                0.624 0.065   24  0.489 0.759 1.26e-09 ***
time30                0.804 0.069   24  0.662 0.945 1.99e-11 ***
time45                0.801 0.072 23.8  0.652 0.951 6.96e-11 ***
time60                0.683 0.065 23.6  0.549 0.817 2.20e-10 ***
treatmentA:time10    -0.068 0.044 21.9 -0.159 0.023    0.137
treatmentA:time20    -0.075 0.054 23.5 -0.186 0.036    0.177
treatmentA:time30     -0.09 0.056 23.7 -0.207 0.026    0.122
treatmentA:time45    -0.056 0.061 23.6 -0.183 0.071    0.370
treatmentA:time60    -0.079 0.055 23.6 -0.194 0.035    0.164
```

Results: Increase from baseline (time=0)

- Switch reference to get the estimated time effect for the other treatment and back-transform with exp for interpretation.

Ratio increase in median aldosteron (pmol/l) from baseline to each follow-up during 30 minute angiotensin stimulation

	Traeatment A	Treatment B	P-value
time	Estimate (95\% CI)	Estimate (95\% CI)	A vs B
10	1.25 (1.15;1.36)	1.34 (1.23;1.45)	0.14
20	1.73 (1.53;1.96)	1.87 (1.63;2.14)	0.18
30	2.04 (1.79;2.33)	2.23 (1.94;2.57)	0.12
45	2.11 (1.84;2.41)	2.23 (1.92;2.59)	0.37
60	1.83 (1.62;2.06)	1.99 (1.73;2.26)	0.16

The response to the stimulation is slightly stronger with placebo, but not significantly so at any follow-up time.

Alternative results: Relative differences in concentration

- Remove intercept and main effect of time from model formula to get treatment differences at each separate time point.

Ratio difference in median aldosteron (pmol/l) for A vs B before and during 30 minute angiotensin stimulation

time	Estimate (95\% CI)	P-value
0	1.87 (1.58;2.23)	<.0001
10	1.75 (1.47;2.09)	<.0001
20	1.74 (1.48;2.04)	<.0001
30	1.71 (1.45;2.02)	<.0001
45	1.77 (1.51;2.08)	<.0001
60	1.73 (1.53;1.96)	<.0001

Concentrations are overall higher with A compared to B and the relative difference seems to be stable during stimulation.

But then again...

Are all of these comparisons clinically relevant - ?!?

- Concentrations at each time point (6 comparisons)?
- Increments since time 0 (5 comparisons)?
- Increments over all pairs of time points (15 comparisons)?

BE CAREFUL

- Increased risk of false positives due to multiple testing.
- Don't drown yourself/your reader in statistical results.

BETTER: A few carefully selected summary statistics.

Relative difference	Estimate (95\% CI)	P-value
before stimulation	1.87 (1.57;2.23)	<0.0001
incremental AUC	1.75 (1.50;2.04)	<0.0001