

## Last time and today

Last time: Time as a covariate and more complex designs

- ▶ The Random Coefficient Model (RCM):
  - ▶ **Linear regression** with repeated measurements
  - ▶ Random tilt and shift of population straight lines
- ▶ More **complex designs**: Cross-over trial with repeated measurements
  - ▶ Analysis of **summary statistics**
  - ▶ Taking **design variables** into account in the analysis

Today:

- ▶ **Variance components** (variation on different levels)
  - ▶ Example: Ear-thickness in wild-type and knockout mice
  - ▶ Example: Muscle strength in patients with multiple sclerosis
  - ▶ **Variation on different levels**
- ▶ Choice of model, and presentation of results
  - ▶ Example: Cardiac output in patients with chronic heart failure and healthy volunteers

## Example: Ear thickness for wild-type and knockout mice

**Data:** 48 hours after an oxazolone induced acute allergic reaction the ear thickness was measured (mm) for both ears of 8 wild-type and 8 MK2 knockout mice. MK2 is involved in the regulation of expressions of several pro-inflammatory cytokines. Data are from **Funding et al. (2008)**.

Knockout			Wild-type		
Mouse	Right	Left	Mouse	Right	Left
1	0.32	0.25	9	0.38	0.50
2	0.24	0.23	10	0.25	0.25
3	0.25	0.26	11	0.31	0.37
4	0.35	0.29	12	0.37	0.31
5	0.26	0.25	13	0.39	0.37
6	0.36	0.30	14	0.46	0.51
7	0.37	0.33	15	0.43	0.34
8	0.32	0.35	16	0.33	0.39

## Ear thickness: Questions of interest

The primary question of interest is whether there is a **systematic** difference between the two types. Is there an association between the knocked out gene and the size of the allergic reaction?

However, it is also of interest to identify how large a part of the **random variation** that is due to

- ▶ The **variation between mice**
- ▶ The **variation between ears (within mice)**

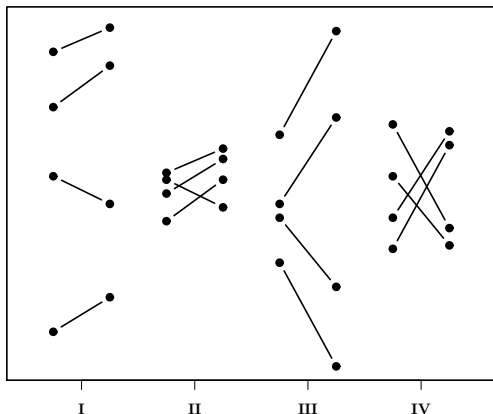
Note that this is **repeated measurements** as we measure the same quantity twice on each mouse, **but** we measure it at **different locations**.

Furthermore, it is typically not of interest to estimate the ear thickness for these particular mice, but rather to make statements about the level (and variation) that we might expect to encounter in the populations of which the mice in this study are a random sample.

## Illustrating variation on two levels

Consider just **one population** and **four subjects** with two measurements on each. We can then have one of the following situations:

- I. Small within and large between subject variation
- II. Small within and between subject variation
- III. Large within and between subject variation
- IV. Large within and small between subject variation

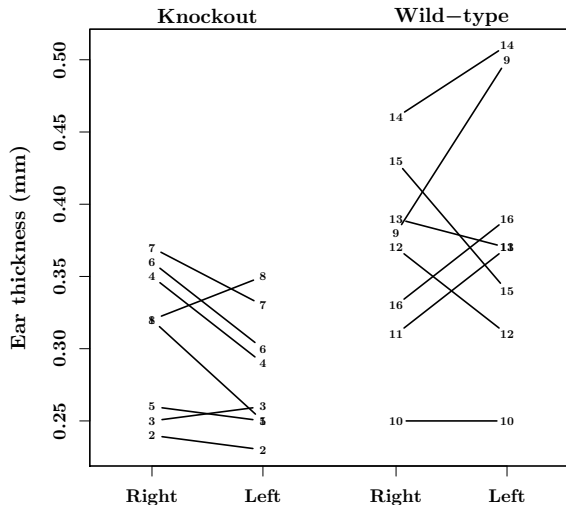


## Ear thickness: Variation in the data

The problem is, of course, that it can be very hard to distinguish the different types of variation in real data when both of them are present to some extent:

**We need a statistical model**

The model should be able to reflect variation on both levels.



# The analysis of variance model with random effects

If we just had ears from 16 different mice in each the two groups we would just assume:

- ▶ Normally distributed ear thickness in each group
- ▶ Independent ears
- ▶ Same variation between ears,  $\sigma$

$$\text{Ear thickness} = \text{type mean} + \text{random ear effect}$$

Now that each mouse contributes with two ears we also need a **random mouse effect**:

$$\begin{aligned} \text{Ear thickness} = & \text{type mean} + \text{random mouse effect} \\ & + \text{random ear within mouse effect} \end{aligned}$$

## Random variation on two levels

Two random effects results in random variation on two levels:

- ▶ Variation **between mice**:  $\sigma_M$
- ▶ Variation **between ears within mouse**:  $\sigma$

### Comments:

- ▶ We have seen this before in the **univariate repeated measurements ANOVA**, where we had a **between** and **within** subject variation
- ▶  $\sigma_M$  and  $\sigma$  are called **variance components** as the total variation is the sum of these two components:

$$\sigma_T^2 = \sigma_M^2 + \sigma^2$$

- ▶ The correlation between measurements on the two ears from the same mouse is

$$\frac{\sigma_M^2}{\sigma^2 + \sigma_M^2}$$

## Ear thickness: Estimates and tests

In our example we get the following estimates for the systematic effect of type:

$$\text{KO} = 0.30, \quad [0.25, 0.34]$$

$$\text{WT} = 0.37, \quad [0.33, 0.42]$$

The  $t$ -test statistic for the hypothesis of no systematic difference between the two types is given by

$$t = 2.62 \sim t(14), \quad p = 0.02$$

We conclude that there is some evidence against the hypothesis of no systematic difference between the two types.

Estimates for the variance components are:

$$\hat{\sigma}^2 = 0.0016, \quad \hat{\sigma}_M^2 = 0.0026, \quad \hat{\sigma}^2 + \hat{\sigma}_M^2 = 0.0042$$

We see that 62% of the variation seen in the measurements is due to the variation between mice.



# Ear thickness: The analyzes in Stata I

Random effects models can be fitted to data using mixed i **Stata**:

```
mixed thickness bn.type, nocons || mouse: , reml var dfmethod(kroger)
```

thickness	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
type						
Knockout	.295625	.0207175	14.27	0.000	.2511905	.3400595
Wild-type	.3725	.0207175	17.98	0.000	.3280655	.4169345

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
-----+-----				
mouse: Identity				
var(_cons)	.0026353	.0013282	.0009813	.0070766
-----+-----				
var(Residual)	.0015969	.0005646	.0007986	.0031931

## Ear thickness: The analyzes in Stata II

The **Stata**-command `pwcompare` gives the exact test for no systematic difference between types.

```
pwcompare type, eff small
estat icc
```

	Contrast	Std. Err.	t	P> t	[95% Conf. Interval]
thickness					
type					
Wild-type vs Knockout	.076875	.0292989	2.62	0.020	.0140351 .1397149

Level	ICC	Std. Err.	[95% Conf. Interval]
mouse	.6226794	.15844	.3055984 .860884

## Ear thickness: Analyzing mice means

In this simple example the t-test for the hypothesis of no difference between the two types can be obtained by analyzing the mice means,

$$\text{Mouse mean} = \frac{1}{2}(\text{ear}_1 + \text{ear}_2),$$

using a two sample t-test.

- ▶ This gives the right t-test ( $t = 2.62 \sim t(14)$ ,  $p = 0.02$ ), but not the correct confidence intervals for the mean parameters.
- ▶ We also get the correct estimate and confidence interval for the type-difference

$$\text{WT-KO} : 0.077 \text{ mm}, 95\% - \text{CI} : 0.014 - 0.140 \text{ mm}$$

- ▶ Obviously it is not possible to say anything about the **within and between** mice variations.
- ▶ If we had more than two groups we would get the correct  $F$ -test by analyzing the subject means using a **one-way ANOVA** model. If the design is just slightly more complicated than that, we cannot get the right test by analyzing a summary statistic.

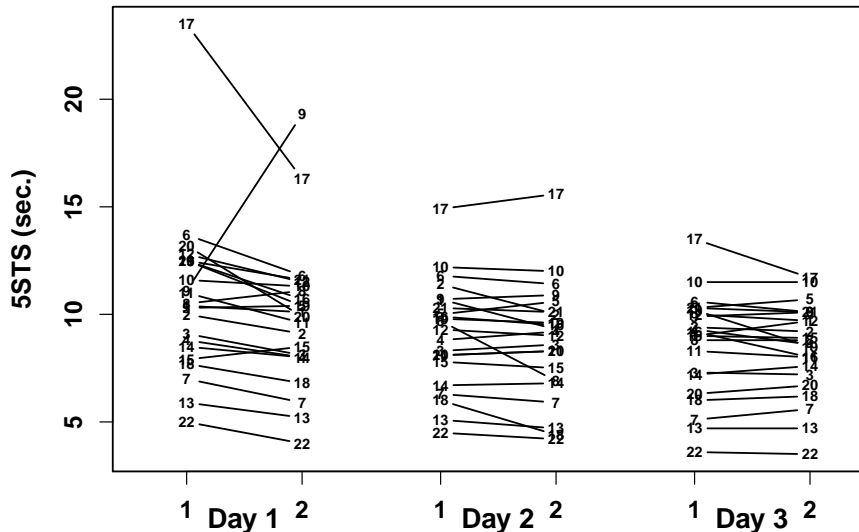
## Example: 5STS patients with multiple sclerosis

**Data:** We consider 22 patients with multiple sclerosis who completed a 5-repetition sit-to-stand test on three days with two trials each day. In the table below we have the time in seconds. The data are from **Møller et al (2012)**.

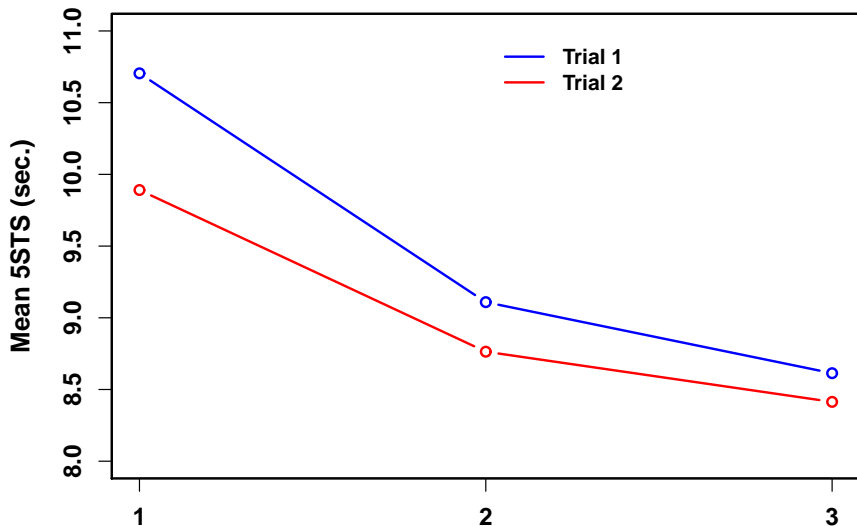
Patient	Day 1		Day 2		Day 3		Patient	Day 1		Day 2		Day 3	
	1	2	1	2	1	2		1	2	1	2	1	2
1	10.4	10.1	10.7	9.3	9.0	9.7	12	12.8	11.5	9.3	9.0	10.0	9.7
2	10.0	9.1	11.4	10.0	9.4	9.2	13	5.9	5.2	5.1	4.7	4.7	4.7
3	9.1	8.1	8.3	8.6	7.3	7.2	14	8.5	8.0	6.7	6.8	7.2	7.6
4	8.8	8.0	8.8	9.2	9.3	8.6	15	7.9	8.5	7.8	7.5	9.0	8.9
5	10.3	10.4	10.0	10.6	10.3	10.7	16	12.5	10.7	9.9	9.5	9.2	8.0
6	13.7	11.8	11.8	11.4	10.6	10.1	17	23.5	16.3	14.9	15.6	13.5	11.7
7	7.0	5.9	6.3	5.9	5.1	5.6	18	7.7	6.8	6.0	4.4	6.0	6.2
8	10.5	11.1	9.7	6.9	8.8	8.8	19	12.5	10.4	9.8	9.6	10.2	8.5
9	11.1	19.3	10.7	10.9	9.9	10.1	20	13.2	9.9	8.1	8.3	6.3	6.7
10	11.6	11.3	12.2	12.0	11.5	11.5	21	12.5	11.6	10.3	10.1	10.3	10.1
11	11.0	9.6	8.1	8.3	8.3	8.0	22	5.0	4.0	4.5	4.2	3.6	3.5

**Questions:** Is there a systematic effect of day and trial (familiarity, fatigue), and what is the variation between subjects and between subjects tested on the same day?

# Scatter plot of the times for the 5-STS



# Mean curves of the times for the 5-STS



## 5-STS times: Factors to include in the analysis

We are interested in the **systematic effect** of:

- ▶ **Day**
- ▶ **Trial**

We want to include a **random patient effect** like in the previous example, and:

- ▶ **A random patient and day interaction**

You could also include a random patient and trial interaction, but that is maybe not so obvious.

We have the following three variance components:

- ▶  $\sigma_P^2$  - the variation between patients
- ▶  $\sigma_{PD}^2$  - the variation between trials
- ▶  $\sigma^2$  - the residual variation

## 5-STS times: Consequences of the statistical model

The total variation is as usual the sum of the variance components:

$$\sigma_T^2 = \sigma_P^2 + \sigma_{PD}^2 + \sigma^2$$

A consequence of the statistical model, which seems natural, is that for measurements on the **same patient on different days**:

$$\text{corr} = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_{PD}^2 + \sigma^2}$$

whereas for measurements on the **same patient on the same day**:

$$\text{corr} = \frac{\sigma_P^2 + \sigma_{PD}^2}{\sigma_P^2 + \sigma_{PD}^2 + \sigma^2}$$

Observations corresponding to different patients are not correlated.



## Recommendations for setting up a statistical model

The following approach can be used if you are in doubt about how to set up a statistical model:

1. Make a list of all factors associated with your experiment
2. Consider which should be included as systematic (fixed) effects and which should be random:
  - ▶ **Systematic:** The ones that you are typically interested in
  - ▶ **Random:** Factors defining a random sample from a population that you would like infer about (typically subjects)
3. In principle consider all interactions between your factors but include only those that you can interpret
4. Interactions where at least one of the factors is random should be random otherwise they should be systematic

## 5-STS times: Two-way ANOVA with random effects

Regarding the factors of interest (day and trial) it seems natural also to include an interaction between the two and therefore in essence consider a two-way ANOVA with random effects of patient and the interaction between patient and day.

The  $F$ -test statistic corresponding to the hypothesis of no interaction is given by

$$F = 0.90$$

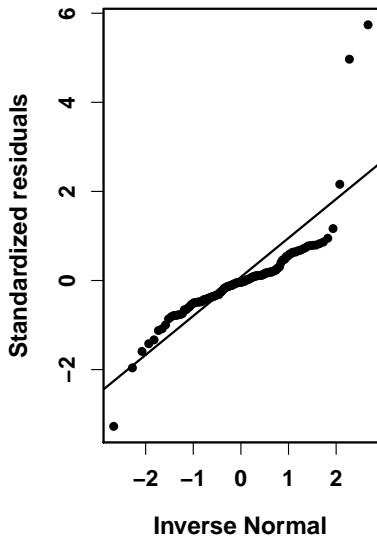
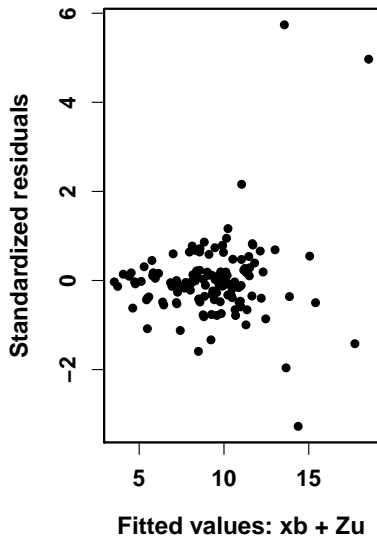
which compared to an  $F(2, 63)$ -distribution results in a  $p$ -value of 0.42.

Testing for no main effect of day and trial:

$$\text{Day: } F = 15.88 \sim F(2, 42), \quad p < 0.0001$$

$$\text{Trial: } F = 5.34 \sim F(1, 65), \quad p = 0.0240$$

## 5-STS times: Model validation



## 5-STS times: Analysis on the log-scale

Now the  $F$ -test statistic corresponding to the hypothesis of no interaction is given by

$$F = 1.58$$

which compared to an  $F(2, 63)$ -distribution results in a  $p$ -value of 0.21.

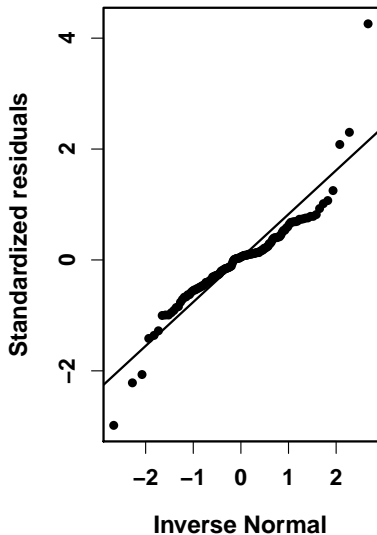
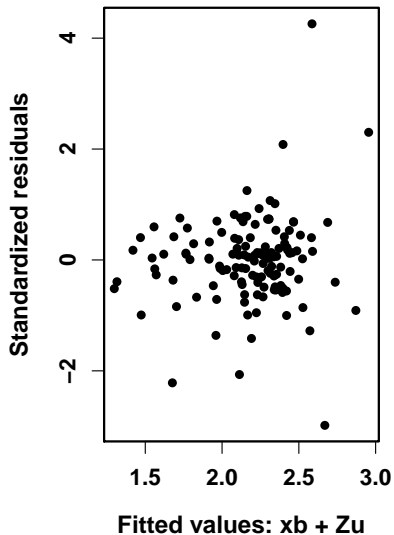
Testing for no main effect of day and trial:

$$\text{Day: } F = 19.84 \sim F(2, 42), \quad p < 0.0001$$

$$\text{Trial: } F = 10.47 \sim F(1, 65), \quad p = 0.0019$$

In other words we get almost identical conclusions: No significant interaction between day and trial but a clear main effect of day and trial.

## 5-STS times: Model validation on the log-scale



The estimated effects are

**Day:**

Effect	Estimate	95%-confidence interval	<i>p</i> -value
2 / 1	0.88	[0.83 , 0.93]	< 0.0001
3 / 1	0.84	[0.79 , 0.89]	< 0.0001
3 / 2	0.96	[0.90 , 1.01]	0.13

**Trial:**

$$2 / 1: 0.95, [0.92, 0.98], p = 0.002$$

The estimated variance components:

Variance component	Estimate	Percent
P	0.0889	87
PD	0.0054	5
Residual	0.0077	8
Total	0.1020	100

## 5-STS times: The analyzes in Stata I

In **Stata** we use `mixed` with two `||` to obtain the two random effects. Here it means that we get a random effect corresponding to patient and one for day within patient.

The hypothesis of no interaction between day and trial is tested using `testparm`.

```
mixed ltime bn.day##bn.trial, nocons || patient: || day: , reml dfmethod(kroger)
testparm bn.day#bn.trial, small
```

```
( 1) [ltime]1bn.day#1bn.trial = 0
( 2) [ltime]2.day#1bn.trial = 0
```

```
F( 2, 63.00) = 1.58
Prob > F = 0.2140
```

## 5-STS times: The analyzes in Stata II

The **Stata**-command `pwcompare` is used to obtain a pairwise comparison of the three days and a comparison of the two trials.

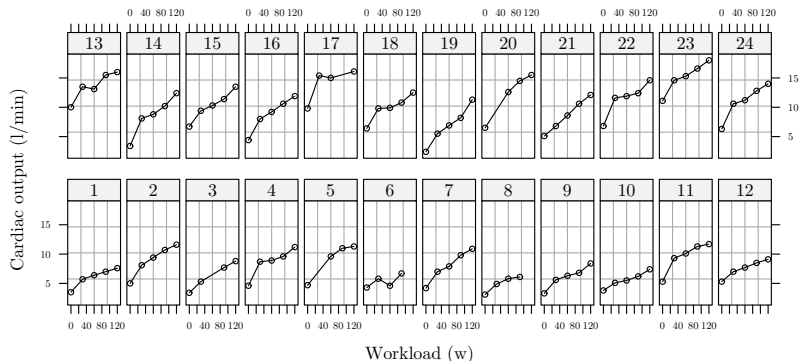
```
pwcompare day trial, eff eform small
```

	exp(b)	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
ltime						
day						
2 vs 1	.87694	.0255019	-4.52	0.000	.8269561	.9299452
3 vs 1	.8383825	.0243807	-6.06	0.000	.7905963	.8890571
3 vs 2	.9560318	.027802	-1.55	0.130	.9015397	1.013817
trial						
2 vs 1	.9512344	.0146964	-3.24	0.002	.9223318	.9810426



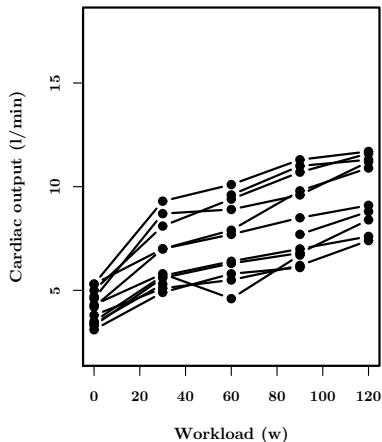
## Example: Cardiac output for CHF patients and healthy

**Data:** In exercise 7 we considered data on cardiac output (l/min) measured for 12 CHF patients (1 – 12) and 12 healthy volunteers (13 – 24). The measurements are obtained using the inert gas re-breathing technique under increasing workload (0, 30, 60, 90, and 120 watt).

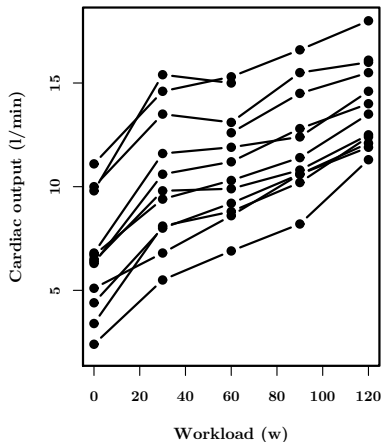


# The variation between subjects in the two groups

CHF

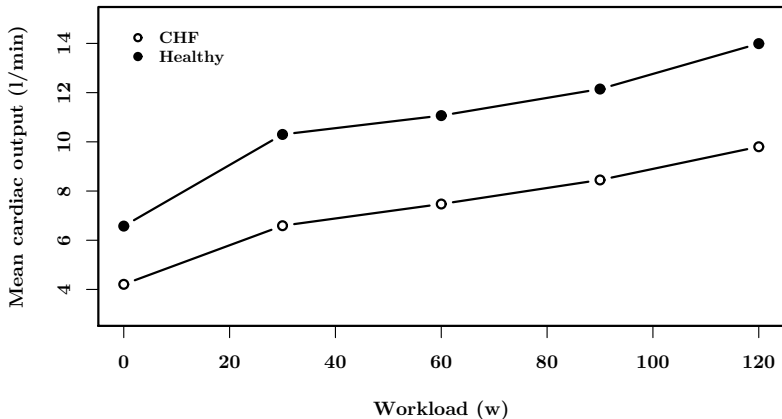


Healthy



# Cardiac output for CHF patients and healthy: Mean curves

We want to compare the two groups. Is there any difference between the cardiac output for CHF patients compared to healthy individuals, and if so does it depend on the workload?



## How should we analyze the data?

So far we have considered several ways of analyzing **repeated measurements** (listed in decreasing order of complexity regarding the correlation between measurements on the same subject):

- ▶ Multivariate repeated measurements analysis
- ▶ Random coefficient models (RCM)
- ▶ Univariate repeated measurements analysis

With **fewer than 6-7 time-points** but several **missing observations** it seems appropriate to consider the **multivariate repeated measurement analysis** for example using `mixed` in **Stata**.

This produces an **approximate** test (likelihood ratio test) for the hypothesis of parallel mean curves in the two groups,

$$\text{Test 1: } LR = 19.45 \sim \chi^2(4), \quad p = 0.0006$$

## Model diagnostics in the multivariate analysis

One way to evaluate the model is to compare the observed and expected standard deviations and correlations. Observed standard deviations and correlations in the two groups:

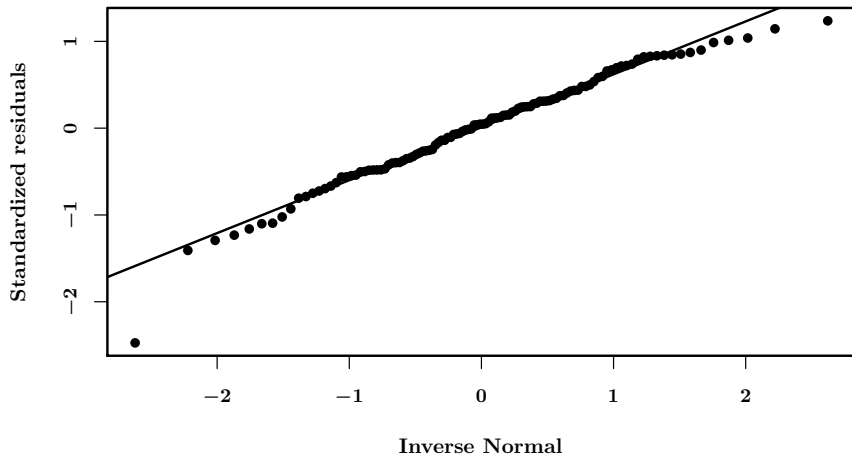
$$\begin{array}{cc}
 \textbf{Healthy} & \textbf{CHF} \\
 \left( \begin{array}{ccccc} 2.65 & & & & \\ 0.94 & 3.21 & & & \\ 0.94 & 0.99 & 2.60 & & \\ 0.92 & 0.96 & 0.98 & 2.52 & \\ 0.92 & 0.94 & 0.97 & 0.97 & 2.08 \end{array} \right) & \left( \begin{array}{ccccc} 0.79 & & & & \\ 0.84 & 1.53 & & & \\ 0.72 & 0.93 & 1.87 & & \\ 0.78 & 0.93 & 0.96 & 1.95 & \\ 0.72 & 0.91 & 0.95 & 0.97 & 1.71 \end{array} \right)
 \end{array}$$

These are to be compared to each other (they should be similar) and to what is implied by the model ( $p = 0.0035$  for the hypothesis of equal standard deviations and correlations in the two groups):

$$\left( \begin{array}{ccccc} 1.95 & & & & \\ 0.91 & 2.46 & & & \\ 0.86 & 0.96 & 2.24 & & \\ 0.85 & 0.94 & 0.97 & 2.32 & \\ 0.79 & 0.90 & 0.96 & 0.96 & 1.95 \end{array} \right)$$

Except for the correlations with the cardiac output at workload 0 in the CHF-group the correlations look similar and are well described by the model. The sd's, however, are not that similar.

## Cardiac output: Model diagnostics (QQ-plot)



## Cardiac output: Excluding the data at workload 0

In the multivariate analysis we saw, particularly in the CHF-patients group, that the measurements corresponding to workload 0 were special compared to the rest of the measurements.

If we exclude the measurements corresponding to workload 0 (and only consider cardiac output during exercise), we get

$$p = 0.07$$

for the hypothesis of equal standard errors and correlations in the two groups.

In the multivariate repeated measurements model there is some evidence against the hypothesis of **parallel curves** ( $p = 0.0397$ ).

Of course you should not exclude data from your analysis just like that, but it might be **instructive** to present the **results of both analyses**.

## Cardiac output: Analyses with unequal standard deviations and correlations

Back to considering all the data:

In **Stata** it is possible to test the hypothesis of parallel mean curves in the two groups without assuming that the standard deviations and correlations are the same in the two groups.

```
mixed co bn.workload#bn.group || subj: , nocons
residual(un, t(workload) by(group))
```

This results in an approximate test,

$$\text{Test 1: } LR = 16.75 \sim \chi^2(4), \quad p = 0.0022$$

**Conclusion:** There is still clear evidence against the hypothesis that the difference between the two groups is the same for all workloads.



## Cardiac output: Observed and expected standard deviations and correlations

Observed standard deviations and correlations in the two groups:

$$\begin{array}{c} \textbf{Healthy} \\ \left( \begin{array}{ccccc} 2.65 & & & & \\ 0.94 & 3.21 & & & \\ 0.94 & 0.99 & 2.60 & & \\ 0.92 & 0.96 & 0.98 & 2.52 & \\ 0.92 & 0.94 & 0.97 & 0.97 & 2.08 \end{array} \right) \end{array}$$

$$\begin{array}{c} \textbf{CHF} \\ \left( \begin{array}{cccccc} 0.79 & & & & & \\ 0.84 & 1.53 & & & & \\ 0.72 & 0.93 & 1.87 & & & \\ 0.78 & 0.93 & 0.96 & 1.95 & & \\ 0.72 & 0.91 & 0.95 & 0.97 & 1.71 & \end{array} \right) \end{array}$$

Expected standard deviations and correlations in the two groups:

$$\begin{array}{c} \textbf{Healthy} \\ \left( \begin{array}{ccccc} 2.65 & & & & \\ 0.93 & 3.10 & & & \\ 0.94 & 0.99 & 2.60 & & \\ 0.93 & 0.96 & 0.98 & 2.65 & \\ 0.92 & 0.94 & 0.97 & 0.97 & 2.08 \end{array} \right) \end{array}$$

$$\begin{array}{c} \textbf{CHF} \\ \left( \begin{array}{cccccc} 0.79 & & & & & \\ 0.84 & 1.53 & & & & \\ 0.72 & 0.92 & 1.81 & & & \\ 0.78 & 0.93 & 0.96 & 1.95 & & \\ 0.75 & 0.93 & 0.91 & 0.97 & 1.74 & \end{array} \right) \end{array}$$

**Conclusion:** The observed standard deviations and correlations are well described by the model.

## Presenting the results from the statistical analysis

Once the data have been analyzed by an appropriate statistical model the focus should be on how to present the results. Usually this involves:

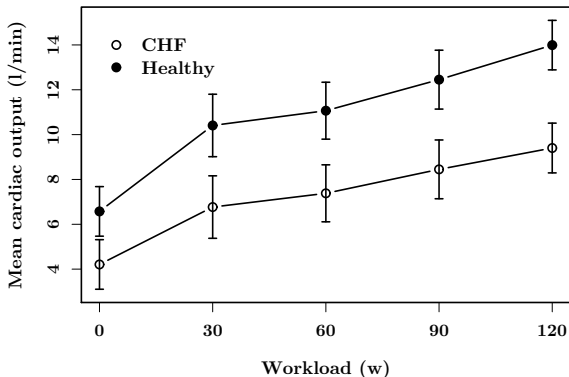
- ▶ Describing how the **data were analyzed**
- ▶ Presenting **estimated means** with 95%-confidence intervals
- ▶ Presenting **estimated effect sizes** (group differences) with confidence intervals (and possibly  $p$ -values)
- ▶ Describing how **model validation** was performed

In repeated measurements analyzes it is most common to focus on potential group differences at the different time-points, but other aspects may be of interest:

- ▶ Is the difference between group A and B more pronounced at time 2 than at time 1?

## Cardiac output: Presenting the estimated means

In this example we get the following estimated means with 95%-confidence intervals:



**A word of caution:** The confidence intervals can be **very wide if the variation between subjects is considerable** (and therefore not very informative regarding changes over time).

## Cardiac output: Presenting relevant estimated differences

In the cardiac output data it looks as if the difference between CHF-patients and healthy gets more pronounced with increasing workload.

Workload (w)	Healthy–CHF (l/min)	95%-CI	<i>p</i> -value
0	2.4	[0.8, 3.9]	0.003
30	3.7	[1.8, 5.7]	0.000
60	3.7	[1.9, 5.5]	0.000
90	4.0	[2.1, 5.9]	0.000
120	4.5	[3.0, 6.0]	0.000

To see whether the difference is significantly bigger at workload 30 compared to workload 0, we can in **Stata** do the following:

```
lincom (30.workload#1.group - 30.workload#2.group) -
      (0.workload#1.group - 0.workload#2.group)
```

## Cardiac output: Comparing group differences

Comparing successive group differences:

Workload (w)	Healthy–CHF (l/min)	95%-CI	<i>p</i> -value
30 – 0	1.36	[ 0.50 , 2.22]	0.002
60 – 30	–0.05	[–0.62 , 0.52]	0.870
90 – 60	0.33	[–0.12 , 0.79]	0.153
120 – 90	0.50	[–0.06 , 1.05]	0.082

We can summarize the difference between CHF-patients and healthy, when it comes to cardiac output, in the following way:

- ▶ The biggest increase in the group difference is seen when going from no workload to a moderate workload (0 – 30 w)
- ▶ At moderate workloads (30 – 90 w) the difference between CHF-patients and healthy is more or less constant
- ▶ When going from a moderate to a high workload (90 – 120 w) the difference between CHF-patients and healthy is once again increased (though not significantly)

## Cardiac output: Comparing workload differences

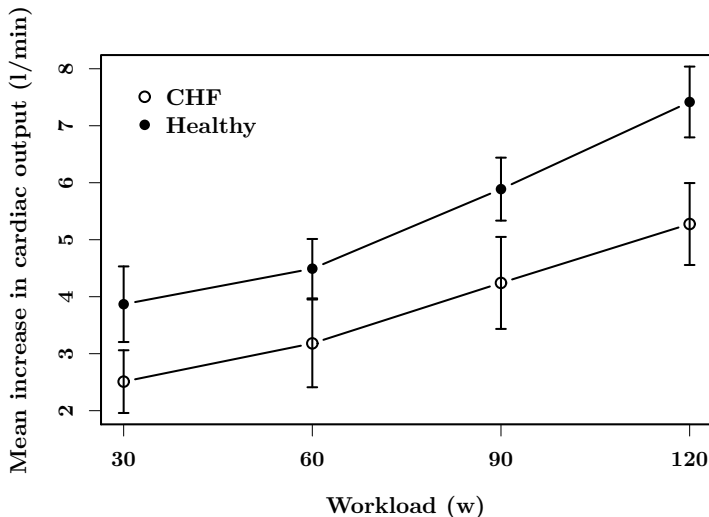
As for the group differences we can give the estimated effect of increasing the workload in the two groups.

Workload (w)	CHF	95%-CI	<i>p</i> -value	Healthy	95%-CI	<i>p</i> -value
30 – 0	2.5	[2.0, 3.1]	0.000	3.9	[3.2, 4.5]	0.000
60 – 30	0.7	[0.3, 1.1]	0.001	0.6	[0.2, 1.0]	0.002
90 – 60	1.1	[0.7, 1.4]	0.000	1.4	[1.1, 1.7]	0.000
120 – 90	1.0	[0.7, 1.3]	0.000	1.5	[1.1, 2.0]	0.000

We can summarize the effect of increasing the workload in the following way:

- ▶ The biggest increase in cardiac output is seen from 0 to 30 w in both groups
- ▶ With each increase in workload we expect to see an increase in cardiac output
- ▶ The increase in cardiac output (with increasing workload) is biggest in the healthy volunteer group (except from 30 to 60 w).

# Increase (from rest) in mean cardiac output for each group



## Cardiac output: Describing how the data were analyzed

A description of the data analysis should contain enough information to enable the reader to achieve the same results if he/she had access to the data and the statistical software used. In this example we might use the formulation:

- ▶ The cardiac output data were analyzed using a multivariate repeated measurements ANOVA with workload and group (and the interaction between them) as factors.
- ▶ The unequal standard errors and correlations in the two groups was taken into account in the analysis.
- ▶ The data were analyzed using **Stata** version 14.2.

### Comments:

- ▶ There is no reason to mention the other simpler models that were considered but did not adequately describe the data.
- ▶ Different statistical packages can do different analyzes and can produce slightly different results when performing the same statistical analysis.
- ▶ Don't just write: " $p < 0.05$  was considered statistically significant" and then say significant / not significant. **Give the p-values.**



## Cardiac output: Describing the test results

The results of the statistical tests can be formulated in the following way:

- ▶ There was clear evidence against the hypothesis of parallel curves in the two groups (identical to the hypothesis of no interaction between workload and group),  $p = 0.0022$ .
- ▶ If the data corresponding to workload 0 were excluded, the  $p$ -value was 0.0397.
- ▶ At all workloads there was a clear difference between the two groups (all  $p$ -values less than 0.003).
- ▶ The hypotheses of equal differences between CHF-patients and healthy volunteers at workload 30, 60, and 90 w were accepted ( $p$ -values over 0.15). At workload 0 w the difference was significantly smaller ( $p = 0.002$ ), and at workload 120 w larger (though not significantly,  $p = 0.082$ ).
- ▶ There were clearly significant increases in cardiac output with increasing workload in both groups (all  $p$ -values less than 0.003).

It may not be relevant to report all these test results.

## Cardiac output: Describing the model validation

Of course we have to assume that people have checked the validity of the statistical analysis even though they do not report what they have done.

Often, though, it is fairly easy to reassure the readers that the results can be trusted by shortly describing what was done to validate the model.

In this example we could simply write the following:

- ▶ Model validation was performed by comparing observed and expected within subject standard deviations and correlations and by inspecting QQ-plots.
- ▶ There was some evidence against the hypothesis of equal standard deviations and correlations in the two groups ( $p = 0.0035$ ), but this was confined to the measurements corresponding to workload 0 w in the CHF-patients group. If the data corresponding to workload 0 w were excluded, the  $p$ -value was 0.07 and all conclusions were unchanged.