Today: Repeated measurements - dependent data

Day 1:

- **Independent data**:
  - Oneway ANOVA (One classification variable)
  - Twoway ANOVA (Two classification variables, interaction)
- **Dependent data**:
  - MANOVA (More than one variable measured for each subject)

Today:

- **Repeated Measurements (dependent data)**:
  - For each subject **the same variable** is measured a fixed number of times at predefined time points (regions, doses, ...)

Today: Repeated Measurement Analysis

- Repeated measurement data
- Example: Growth of rats
- Hypotheses in repeated measurement analysis
- Multivariate repeated measurement analysis
- Multivariate repeated measurement analysis with missing values
- Example: VO_2 in CHF patients and healthy during exercise
- Repeated measurement analysis with many time points
- Example: The effect of cholagogues on gallbladder volume
- Univariate repeated measurement analysis

The organization of repeated measurements

In the simplest setting we have **two treatment groups** and a number of **subjects** in each group **measured at several time-points**.

		Time-point			
Group	Subject	1	2	$\cdots$	T
1	1	$\square$	$\square$	$\cdots$	$\square$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_1	$\square$	$\square$	$\cdots$	$\square$
2	1	$\square$	$\square$	$\cdots$	$\square$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_2	$\square$	$\square$	$\cdots$	$\square$

What we want to learn from repeated measurements

Typically, we want to compare the two treatment groups, and

- assess whether there is a **significant treatment effect** or not

However, usually there is a good reason for measuring the parameter of interest several times for each subject:

- The **timing of the treatment effect** may be of importance, or
- The **treatment effect may increase with time** since treatment initiation, or
- The **treatment effect may be uninfluenced by time** (constant treatment effect)

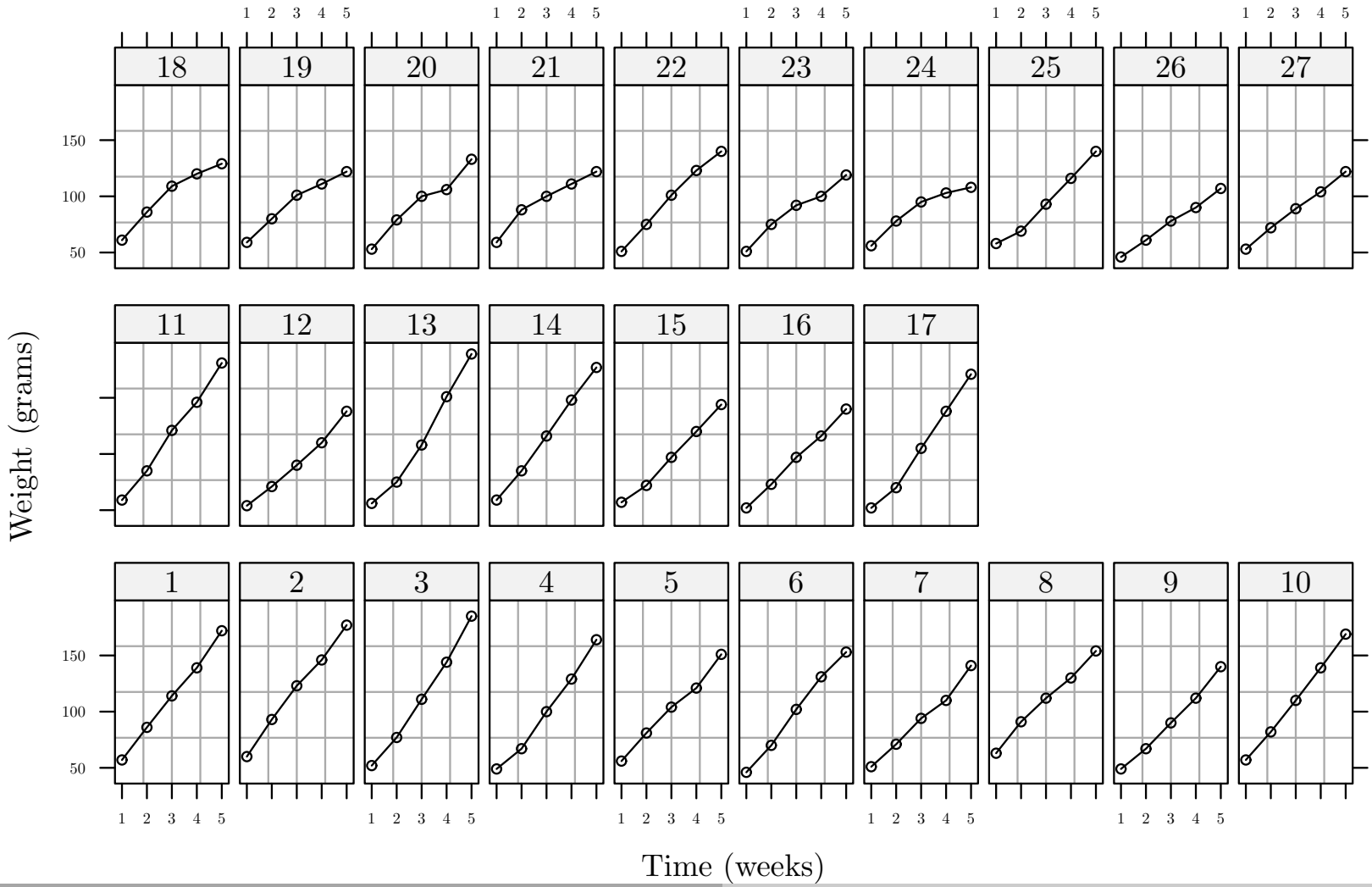
In other words, the **interplay between treatment and time** is of importance.

Growth of rats: The data

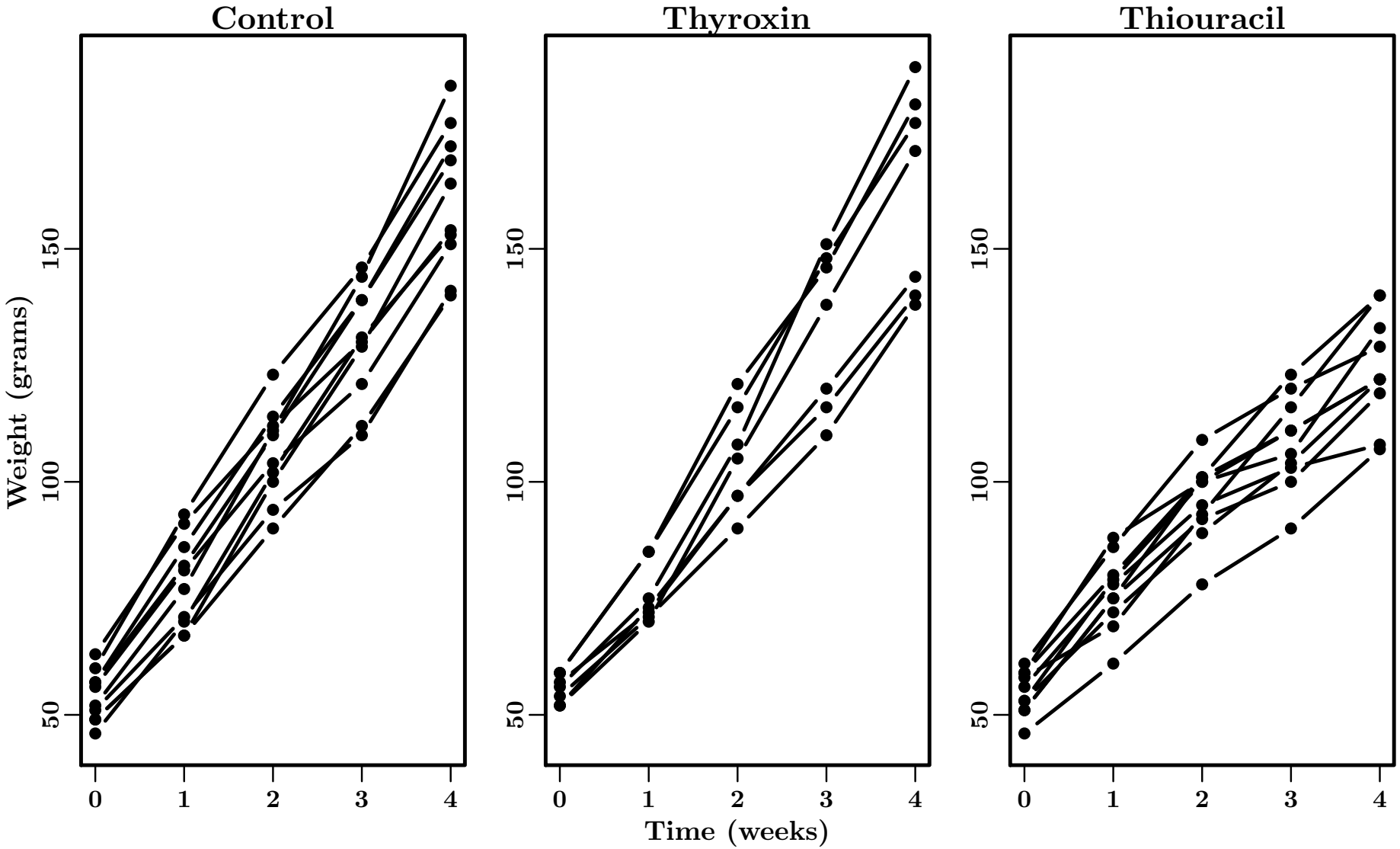
Data: Weight (grams) of 27 rats randomized to having either **nothing** (Group 1), **Thyroxin** (Group 2), or **Thiouracil** (Group 3) added to their drinking water immediately after the weight measurement at week 0.

Rat	Group	Week 0	Week 1	Week 2	Week 3	Week 4
1	1	57	86	114	139	172
2	1	60	93	123	146	177
3	1	52	77	111	144	185
4	1	49	67	100	129	164
5	1	56	81	104	121	151
6	1	46	70	102	131	153
7	1	51	71	94	110	141
8	1	63	91	112	130	154
9	1	49	67	90	112	140
10	1	57	82	110	139	169
11	2	59	85	121	146	181
12	2	54	71	90	110	138
13	2	56	75	108	151	189
14	2	59	85	116	148	177
15	2	57	72	97	120	144
16	2	52	73	97	116	140
17	2	52	70	105	138	171
18	3	61	86	109	120	129
19	3	59	80	101	111	122
20	3	53	79	100	106	133
21	3	59	88	100	111	122
22	3	51	75	101	123	140
23	3	51	75	92	100	119
24	3	56	78	95	103	108
25	3	58	69	93	116	140
26	3	46	61	78	90	107
27	3	53	72	89	104	122

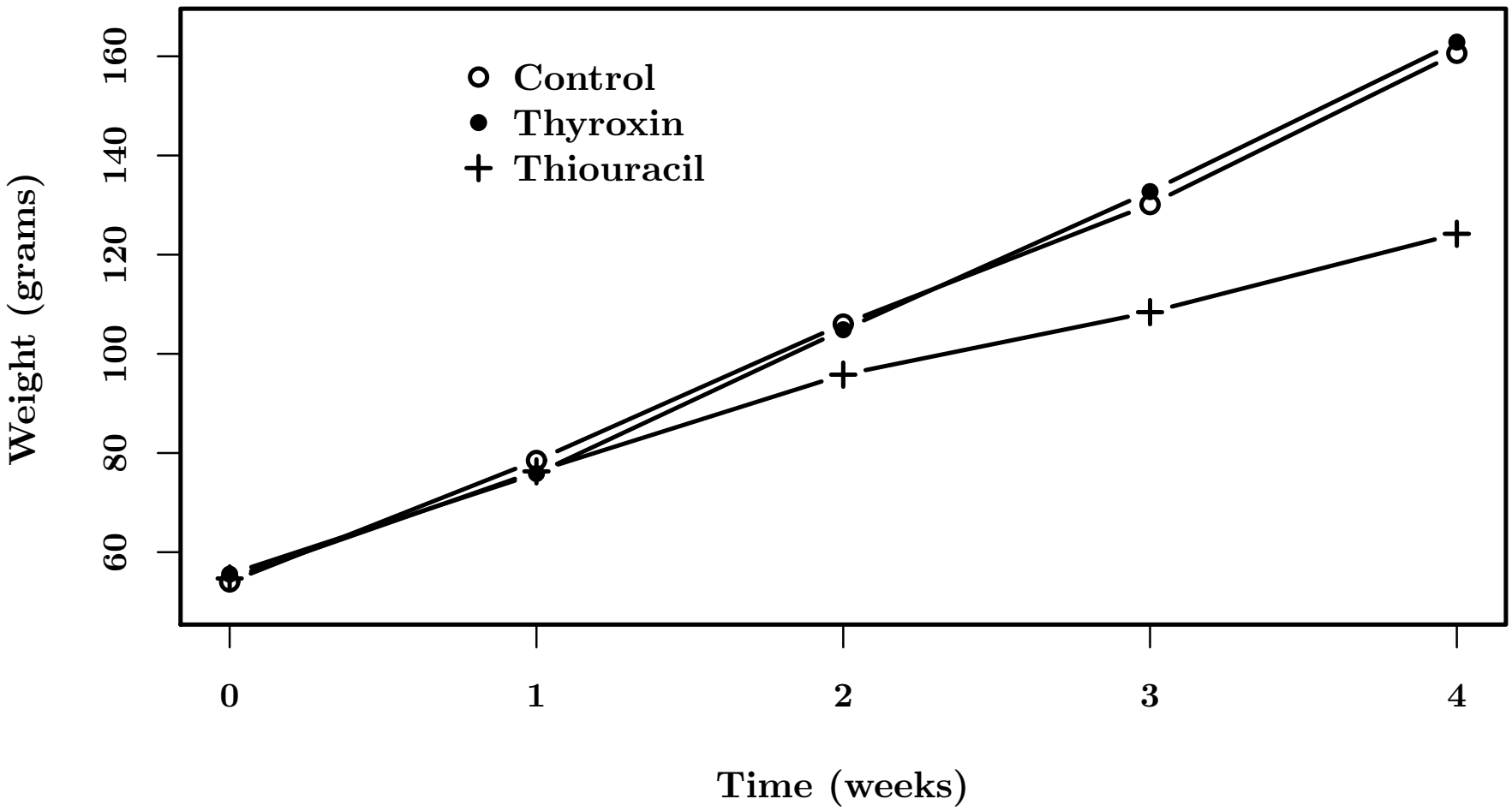
Growth of rats: Separate individual curves



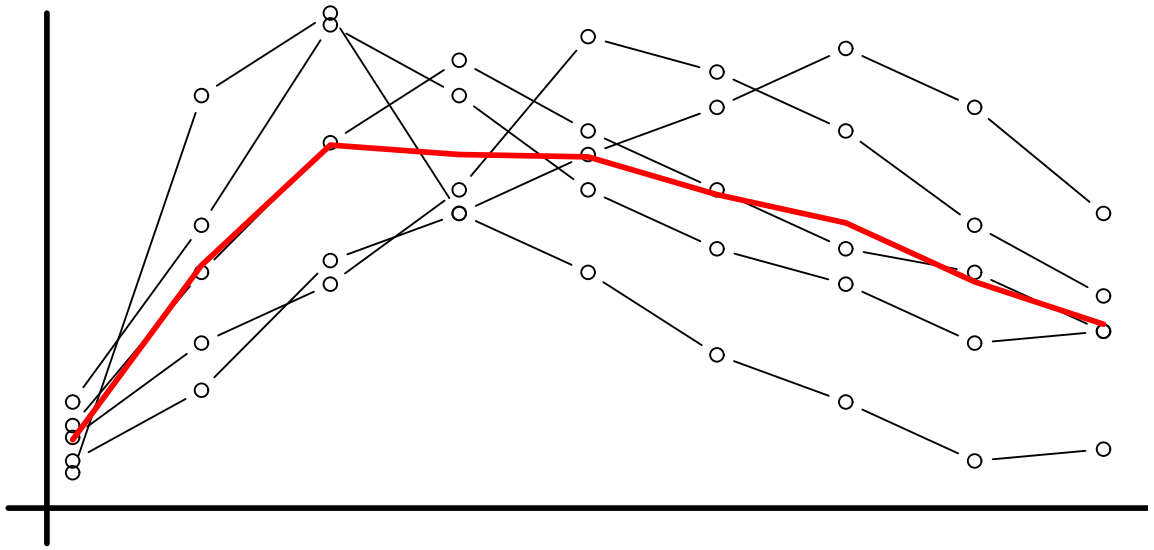
Growth of rats: Grouped individual curves



Growth of rats: Group mean curves



The mean curve should reflect the individual curves



- The **mean curve peaks at a lower level** than the individual curves
- The individual non-linear curves are **shifted horizontally** compared to each other
- This kind of curves is typical in pharmacokinetics/-dynamics, but the analysis of them is beyond the scope of this course (apart from summary statistics)
- **BASIC ASSUMPTION**: The shape of the mean curve is similar to that of the individual curves

Repeated measurements: Variation on two levels

When dealing with repeated measurements we need to consider variation on two levels:

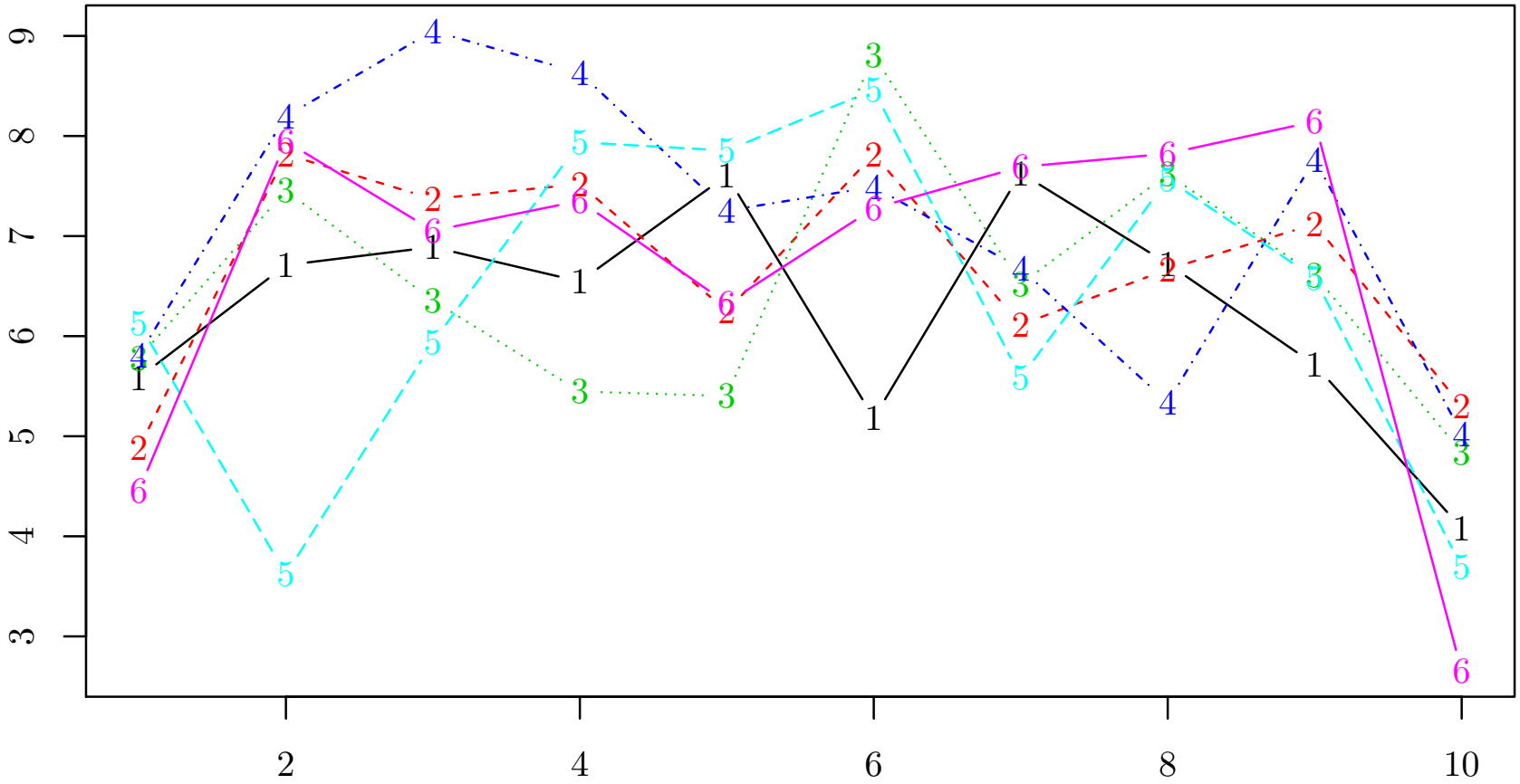
- **Within subject variation** - the variation between measurements on the same subject
- **Between subject variation** - the variation between levels for different subjects

The two types of variation are relevant for different purposes:

- **Within subject variation** - when analyzing **changes over time**
- **Between subject variation** - when analyzing **group differences**

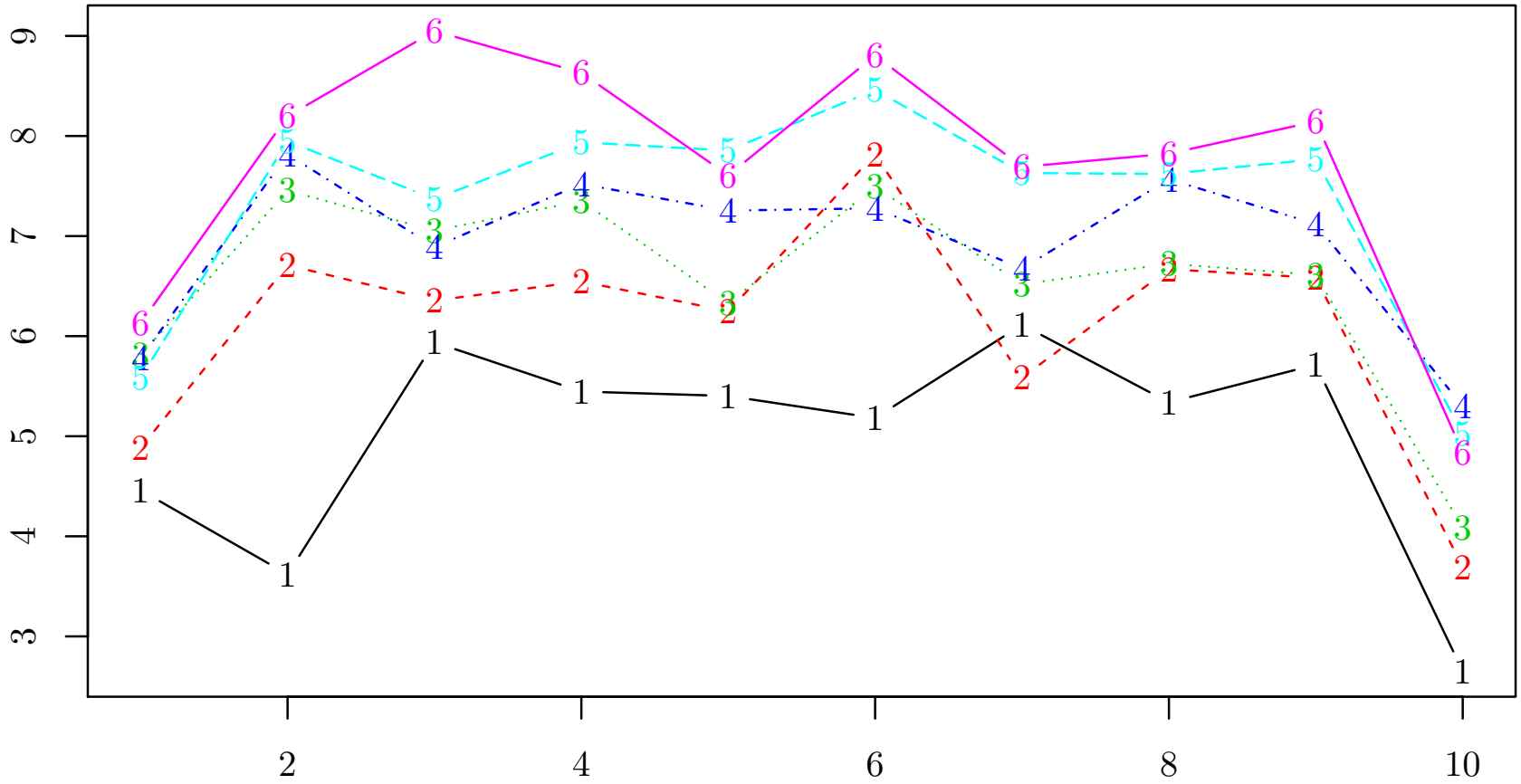
With only two time-points, the **between subject variation** is eliminated by considering the difference between the two observations for each subject.

Within and between subject variation: Data set I



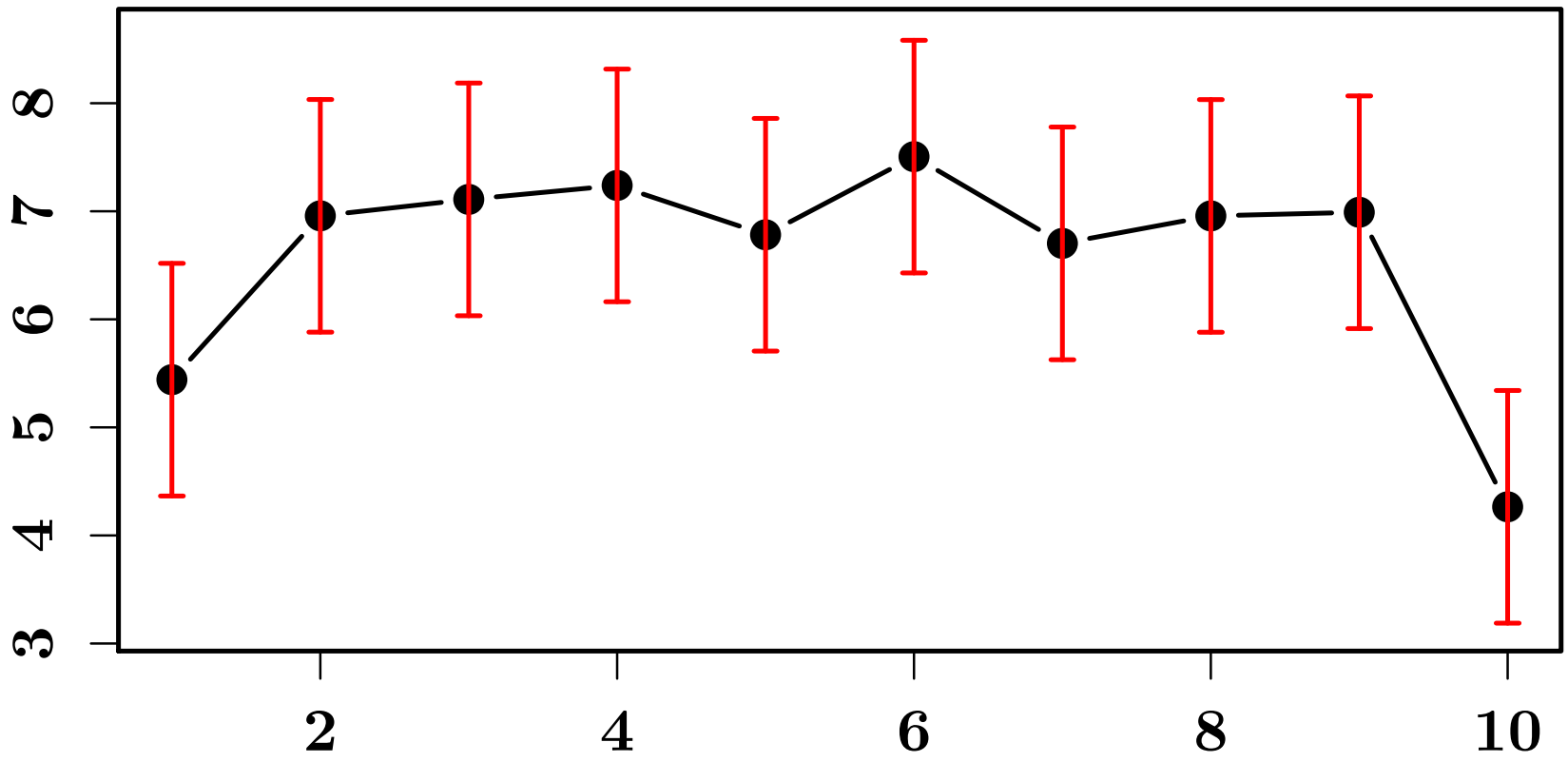
- Large **within subject variation** (100%)
- Small **between subject variation** (0%)

Within and between subject variation: Data set II



- Small **within subject variation** (22%)
- Large **between subject variation** (78%)

Within and between subject variation: Data set I and II



- Data set I and II have **equal mean curves**
- Data set I and II have **equal total standard deviation** (1.0767)

Repeated measurements: Comparing groups

As already mentioned, typically we want to compare treatment groups, and

- assess whether there is a **significant treatment effect** or not

What does the following statement tell you:

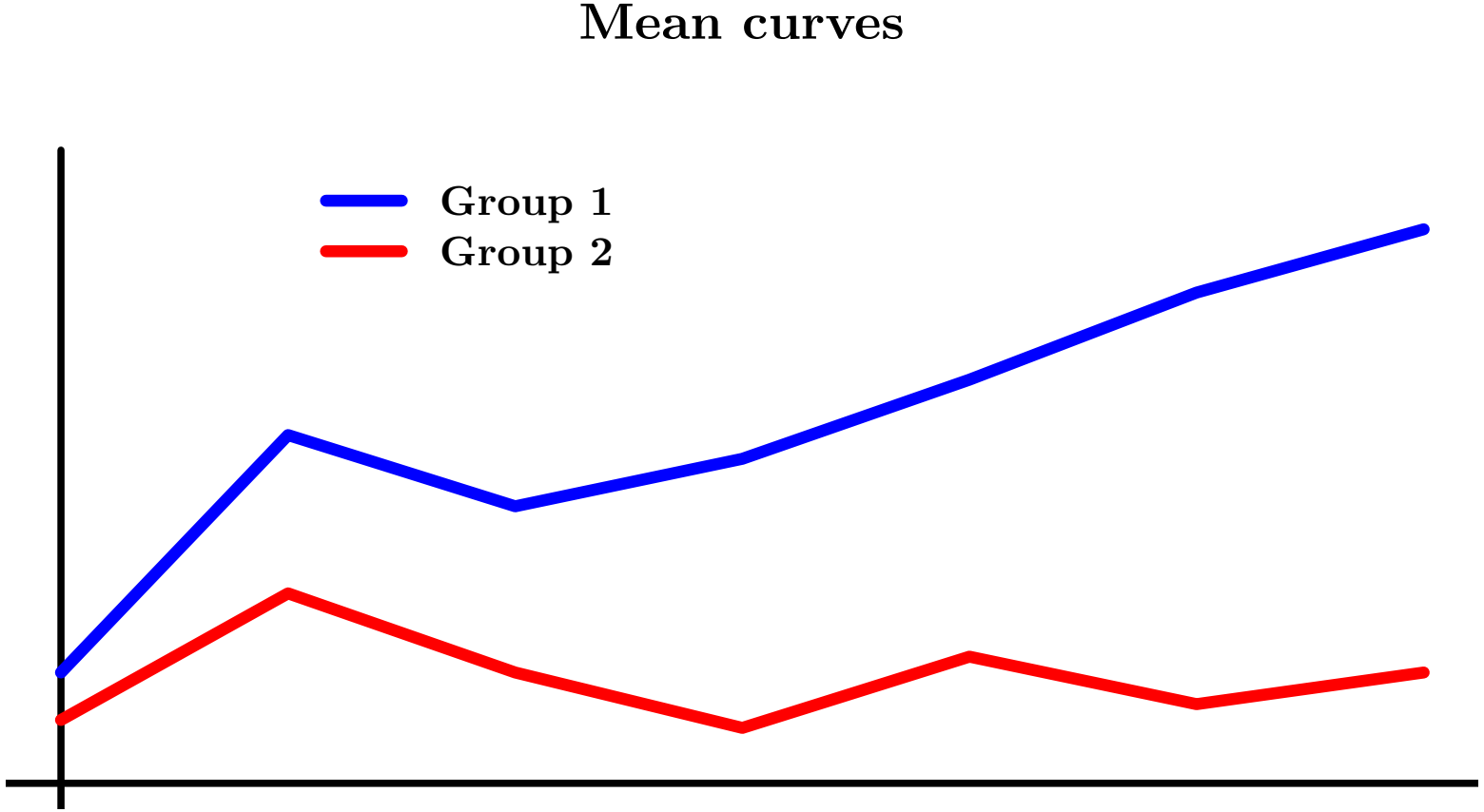
> Data were analyzed by repeated measurement analysis and there was a statistically significant difference between the groups (p=....).

??

There is more than one hypothesis of potential interest that can be tested, and it is not clear from the statement above to which hypothesis the p-value is referring.

In the following we will consider **5 different hypotheses**.

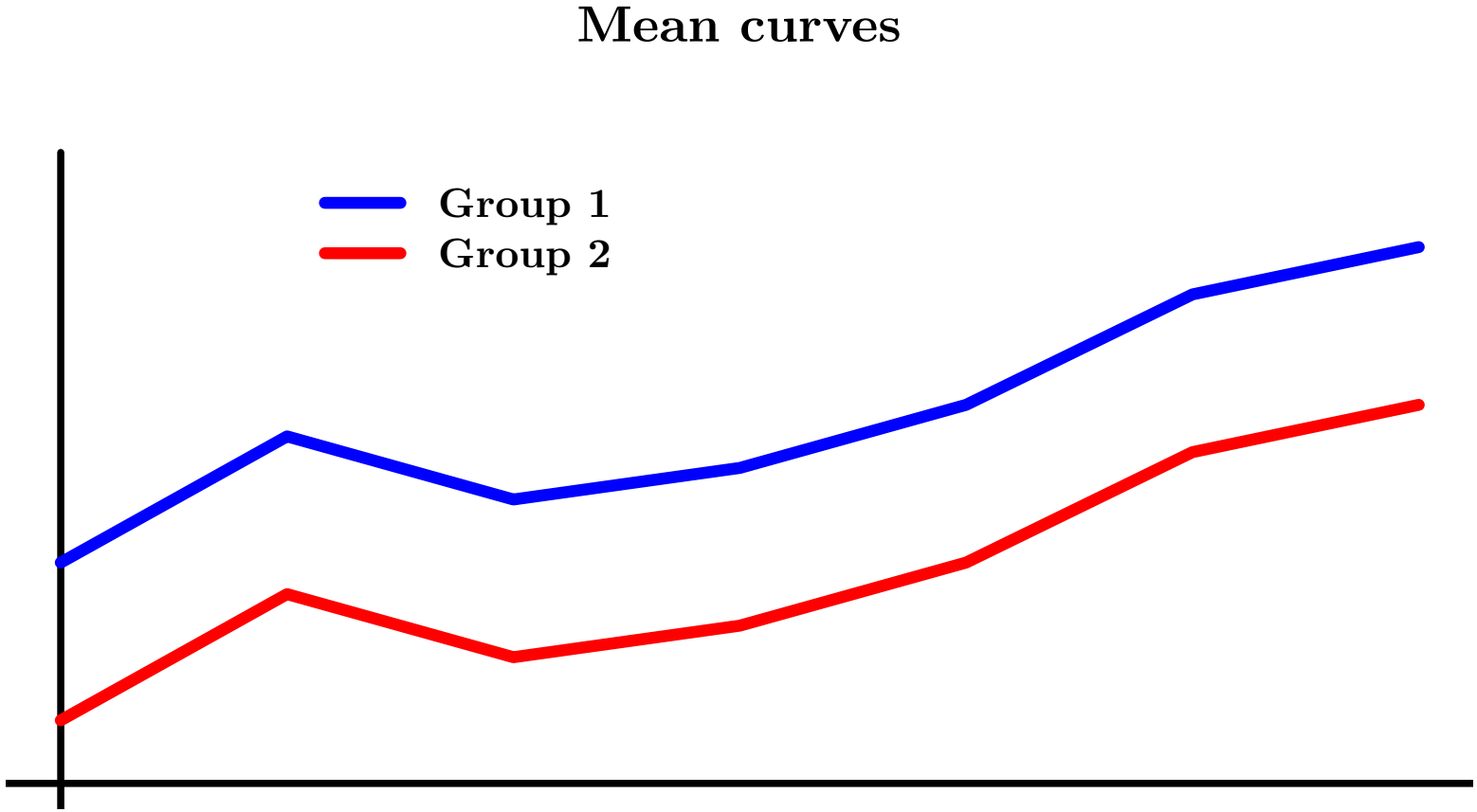
The starting point (H_1)



H_1: Two non-parallel mean curves

- The **development over time** is different in the two groups

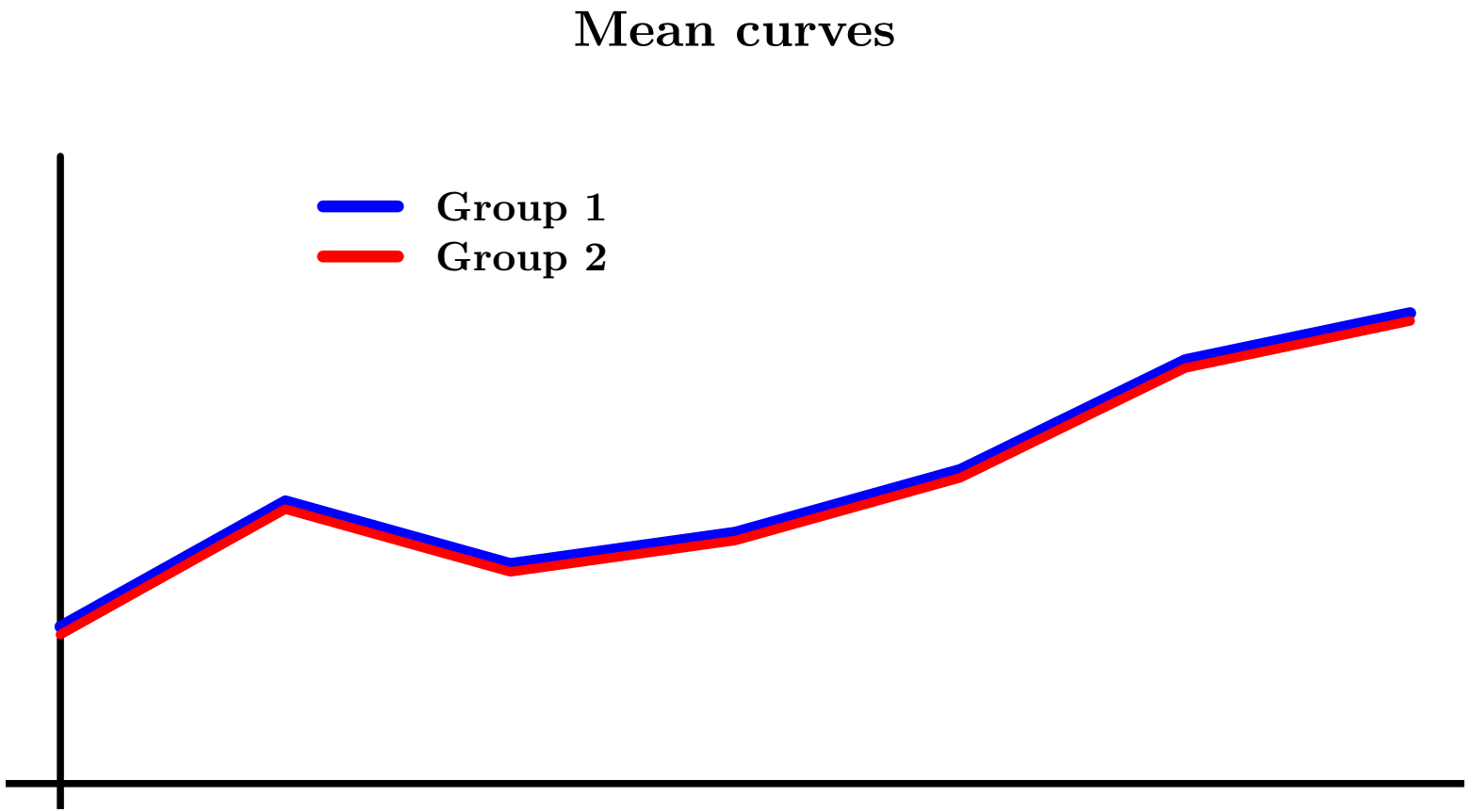
Same development over time in the two groups (H_2)



H_2: Two parallel mean curves

- Same **development over time** but different levels in the two groups

Equal mean curves in the two groups (H_3)

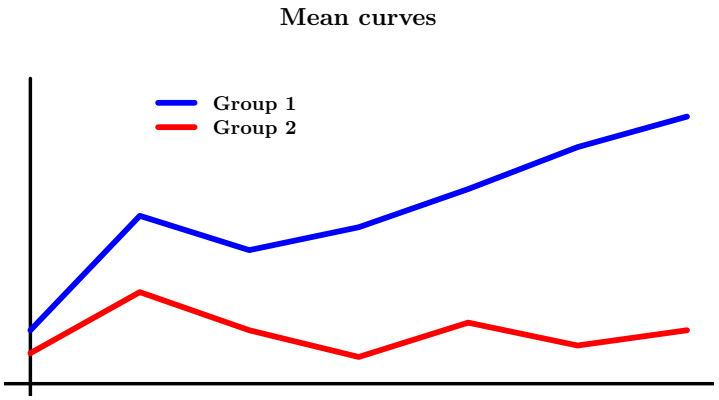


H_3: Equal mean curves

- Same **development over time** and levels in the two groups

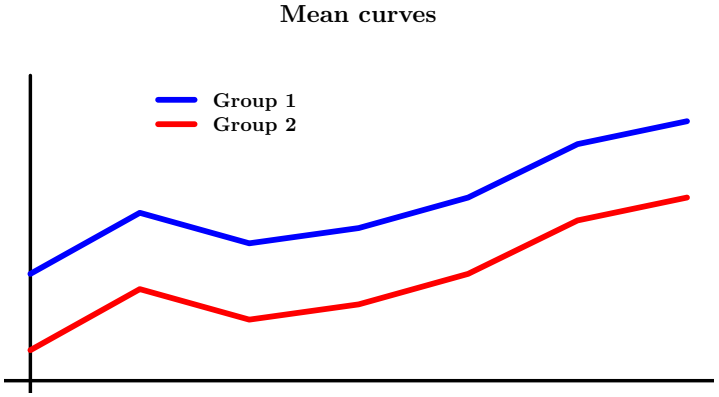
Testing for parallel mean curves

Test 1:



Assumption H_1:
Two non-parallel curves

$\Downarrow$

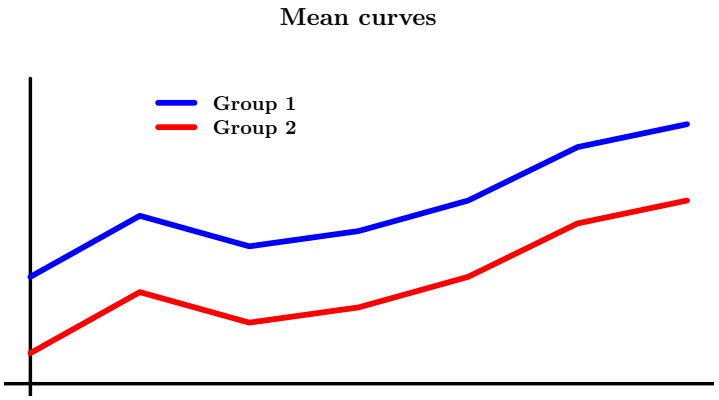


Hypothesis H_2:
Two parallel mean curves at different levels

The changes over time are similar in the two groups.

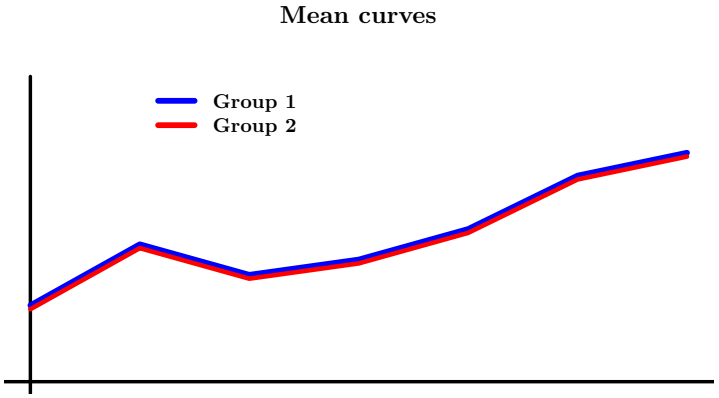
Testing from parallel mean curves to equal mean curves

Test 2:



Hypothesis H_2:
Two parallel mean curves at different levels

$\Downarrow$



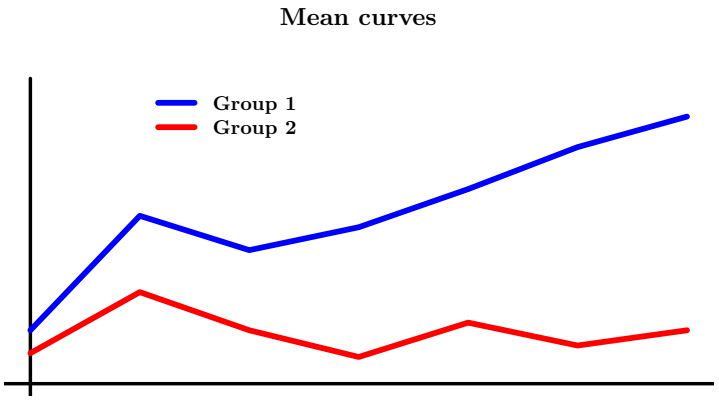
Hypothesis H_3:
Two equal mean curves

No difference between the two groups at any time point.

Testing from non-parallel mean curves to equal mean curves

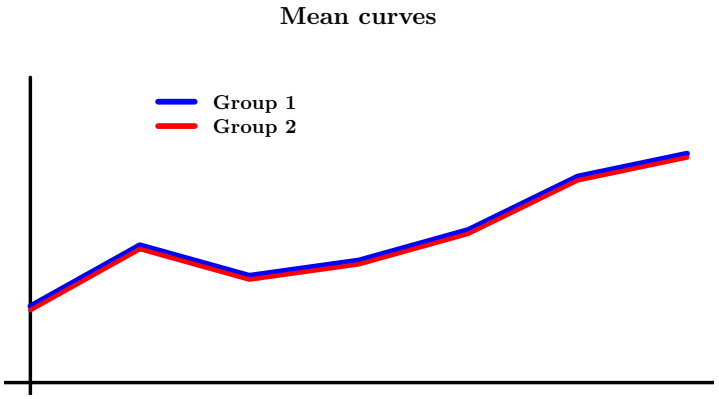
Test X:

Assumption H_1:
Two non-parallel mean curves



$\Downarrow$

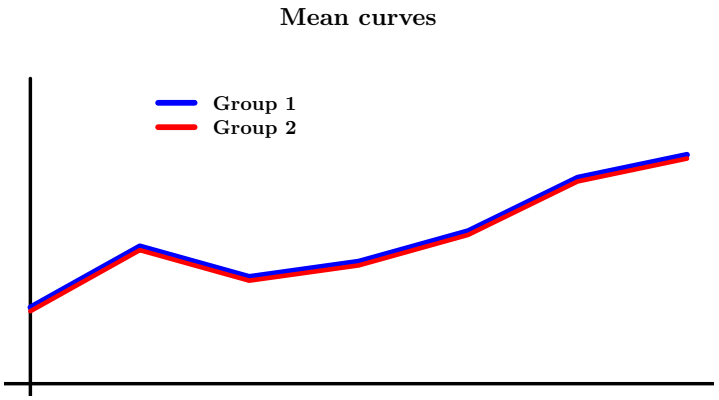
Hypothesis H_3:
Two equal mean curves



Going from arbitrary mean curves to no difference between the two groups at any time point.

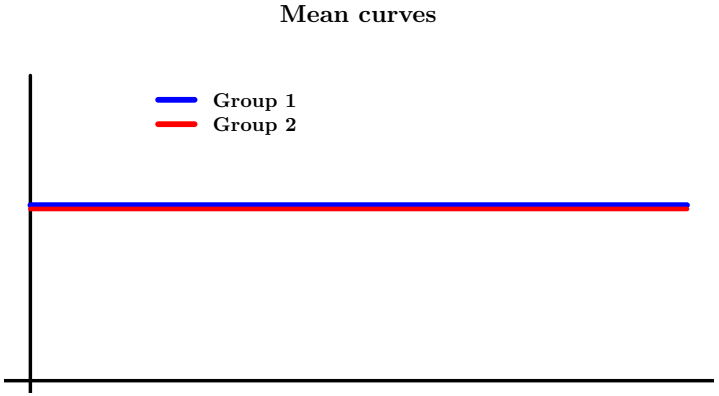
Testing for horizontal mean curves

Test 3:



Hypothesis H_3:
Two equal mean curves

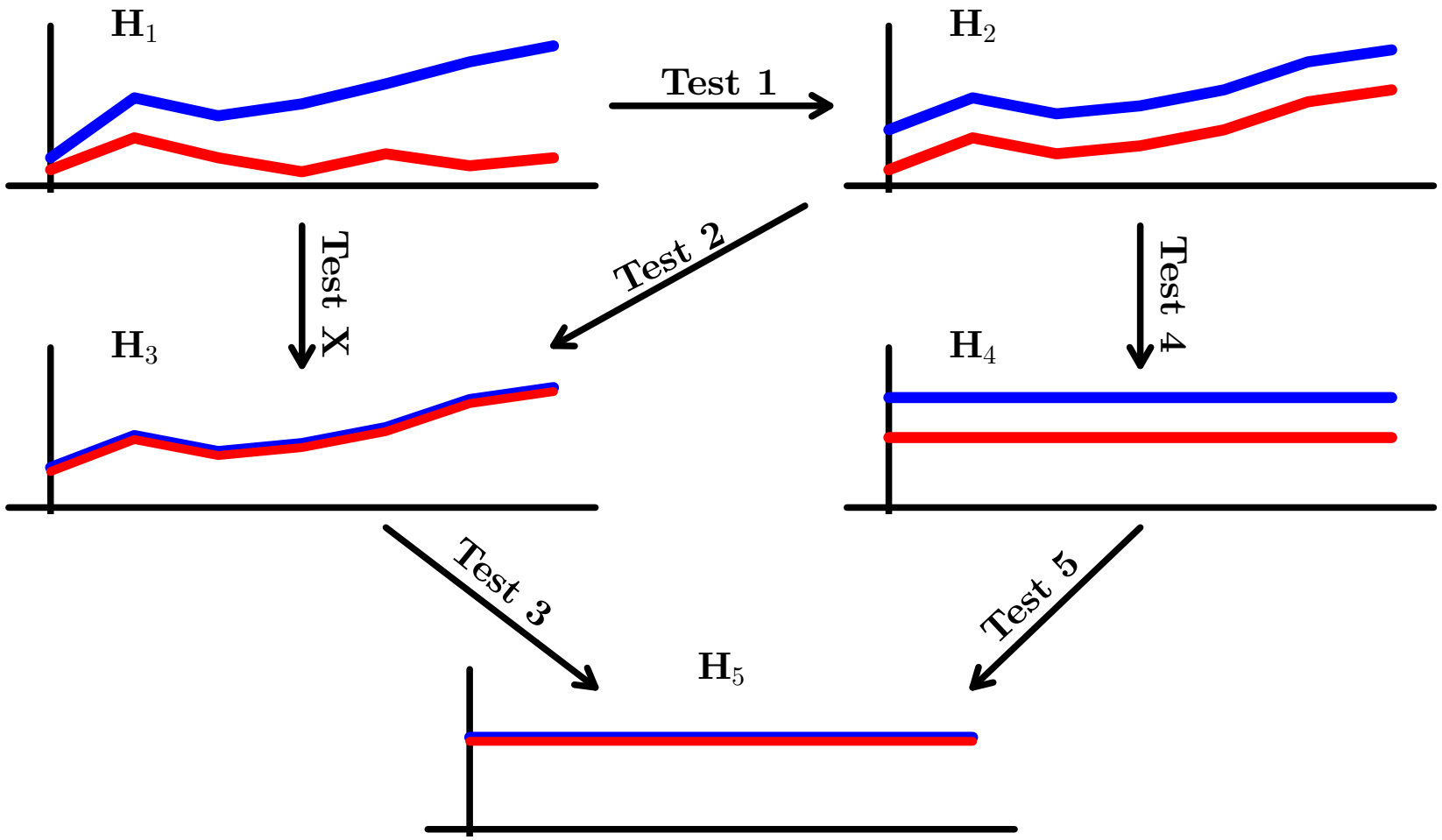
$\Downarrow$



Hypothesis H_5:
Two equal mean curves at a constant level

No development over time in any of the mean curves.

All the relevant hypotheses and tests



Make it clear which of the tests **the stated p-values correspond to**.

Underlying distributional assumptions

In order to make the different tests, we need to make some **assumptions about the distribution of the data**:

1. Measurements corresponding to different subjects are **independent**.
2. Measurements corresponding to the same subject are **multivariate normally distributed**.
3. **Standard deviations and correlations** between the different measurements on the same subject are the **same** in the different groups.

Comments:

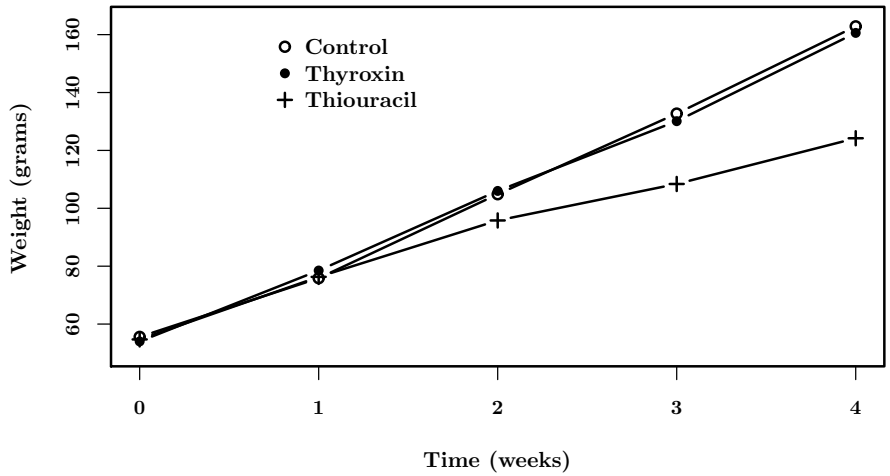
- The assumptions are identical to those behind **MANOVA** that we considered on day 1.
- Assumption 3 is only needed to achieve **exact tests**. We can test whether it is reasonable (like in the situation with two random samples from normal distributions), and if it is not accepted, we can still proceed but with approximate tests.

More on the underlying distributional assumptions

- Reordering the set of observations will **not** change the tests. That is, the fact that we measure the same quantity repeatedly over an **ordered variable** (time, dose, ...) is not used.
- **Test 1** (parallel curves), **Test 3** (no change over time), and **Test 4** (no change over time in either group) really only require that
  - **All pairwise differences** between any two time points should be multivariate normally distributed

  With k different time points it suffices to examine the $k-1$ differences: $(t_2 - t_1), (t_3 - t_2), \ldots, (t_k - t_{k-1})$.
- **Test 2** (from parallel to equal mean curves) and **Test 5** (same constant level in the two groups) really only require that
  - **The average** over time (or sum) is normally distributed (unpaired t-test)

Growth of rats: Test for parallel mean curves



Let us start by only considering group 1 and 3 (Control and Thiouracil). The usual strategy is first to **test for parallel mean curves** (Test 1):

$$\textbf{Test 1}: \qquad F = 9.76 \sim F(4, 15), \quad p = 0.0004$$

We conclude that there is clear evidence against the hypothesis of parallel mean curves. So there is a **significant treatment effect** but how large it is depends on time.

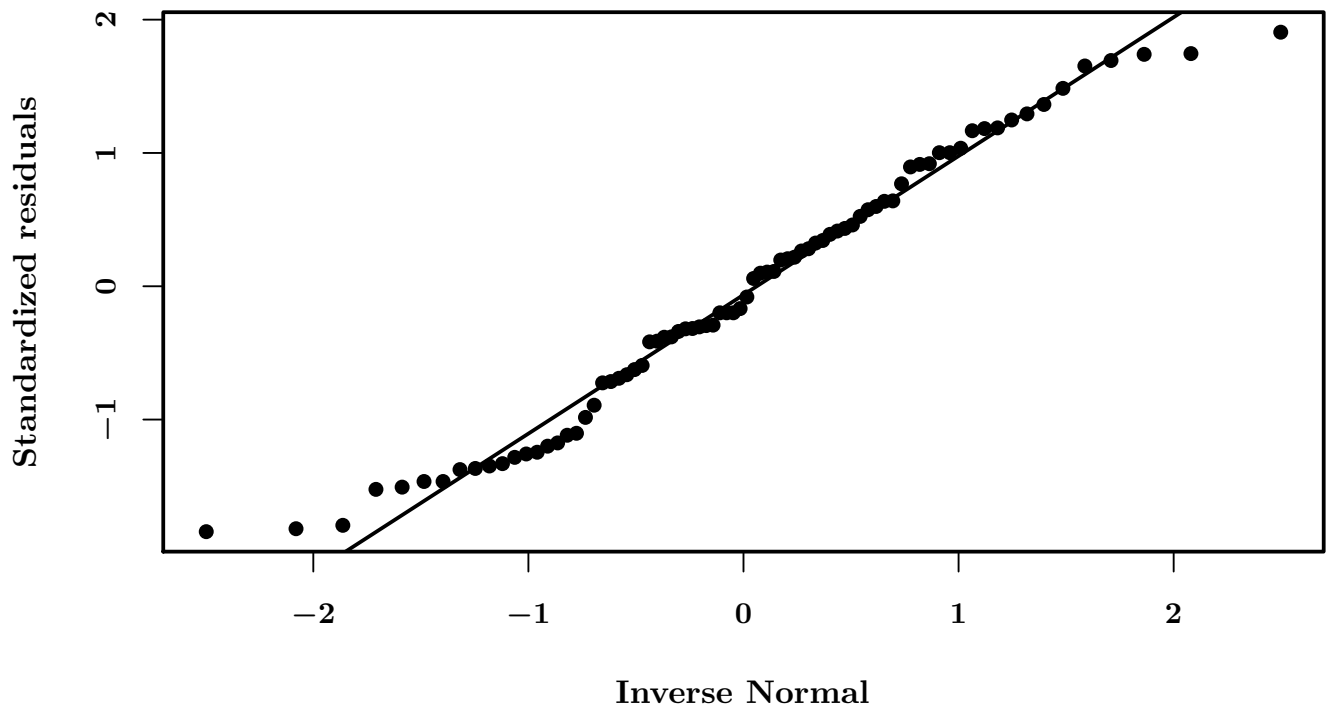
Could we have concluded that there was a significant treatment effect **if** we had accepted the hypothesis?

Growth of rats: Model validation

Remember that we have assumed that the time differences are **multivariate normally distributed with the same standard deviations and correlations** in the two groups.

We can test the hypothesis of equal standard deviations and correlations in the two groups:

$$p = 0.90$$



Growth of rats: Comparing Control and Thyroxin

Let us briefly consider group 1 and 2 (Control and Thyroxin). Testing for parallel mean curves yields

$$\textbf{Test 1}: \qquad F = 1.13 \sim F(4, 12), \quad p = 0.39$$

Having accepted that the mean curves in the two groups are parallel, we can then proceed and test for equal mean curves (Test 2):

$$\textbf{Test 2}: \qquad t = -0.11 \sim t(15), \quad p = 0.92$$

We conclude that there is no evidence against the hypothesis of equal growth curves in the two groups.

We could have tested the hypothesis of no effect of Thyroxin directly (Test X: $F = 3.18 \sim F(5, 11)$, $p = 0.05$), but **Test X has lower power** than Test 1 and 2, so:

- There is a greater risk of falsely concluding that there is no group effect
- We might overlook an easily interpretable description of the data (H_2)

Growth of rats: Analyses in Stata

In **Stata** we can use the command `mvtest` to test for parallel mean curves (after creating the differences `W10=Weight1-Weight0` and so on):

```
mvtest means W10 W21 W32 W43 if Group==1 | Group==3, by(Group)
```

The command `mvtest` can also be used to test for equal standard deviations and correlations in the two groups:

```
mvtest cov W10 W21 W32 W43 if Group==1 | Group==3, by(Group)
```

If we cannot accept the hypothesis of equal standard deviations and correlations, then we can make an approximate test of the hypothesis of parallel mean curves in the following way:

```
mvtest means W10 W21 W32 W43 if Group==1 | Group==3, by(Group) het
```

Problems in multivariate repeated measurements analyses

So far we have only considered **multivariate repeated measurements analyses** based on for example `mvtest` in **Stata**. This results in **exact F-tests** corresponding to the hypotheses of interest.

There are two major problems, however, that may arise in practice when using these methods:

1. **Missing values**: Often some individuals are missing measurements at certain time-points. The problem with multivariate repeated measurements analyses, as we have considered them so far, is that **all** measurements, for an individual with just one missing value, are left out of the analysis.
2. **Many time-points**: When there are more than about 6-7 time-points, the number of standard deviations and correlations becomes so great that it is not really feasible to make this kind of analysis.

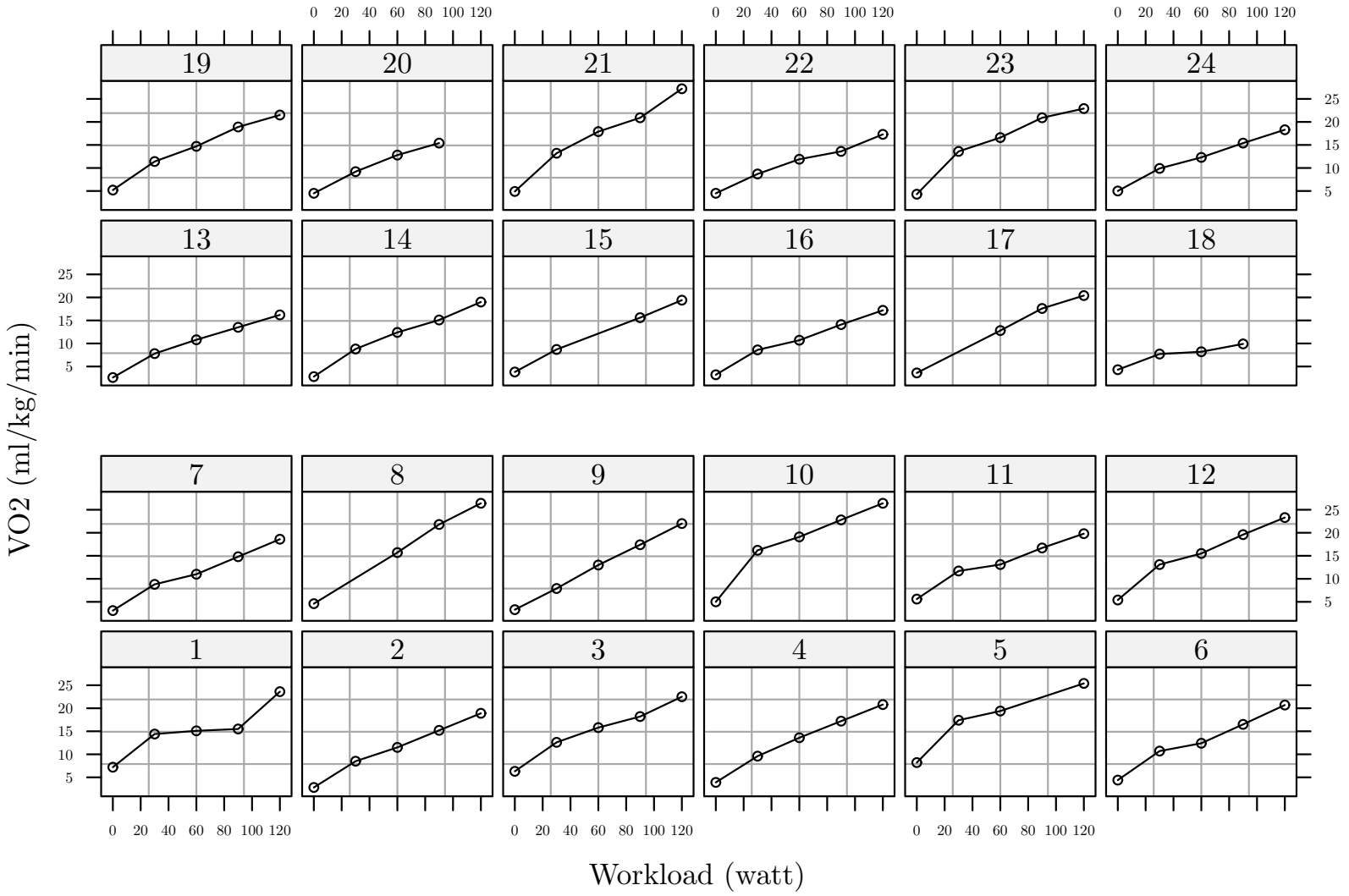
Let us consider an example that illustrates the first problem.

Example: VO_2 in CHF patients and healthy during exercise

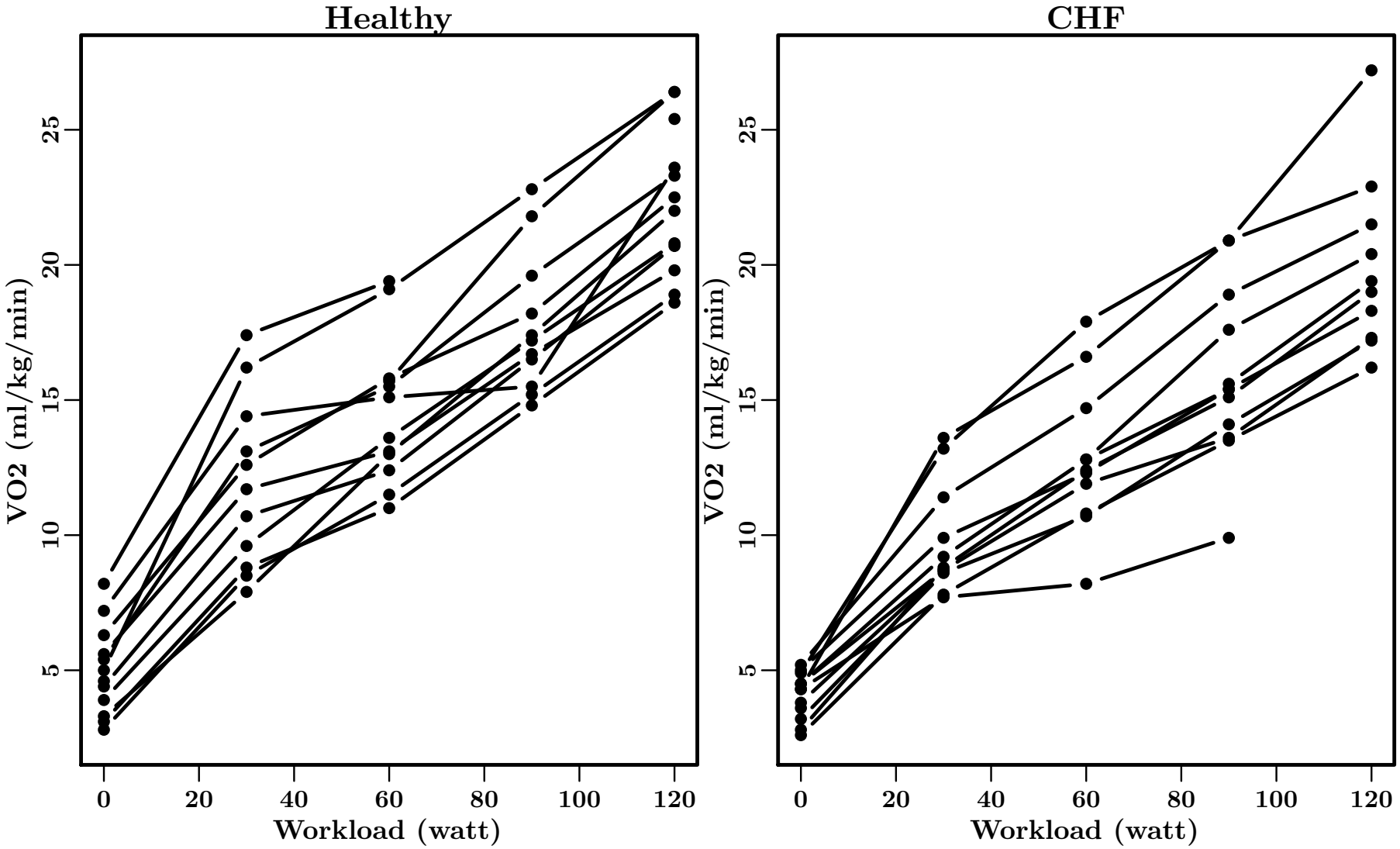
Data: In an experiment including 12 CHF patients and 12 healthy volunteers the individuals had their VO_2 (ml/kg/min) measured during increasing workload (0, 30, 60, 90, and 120 watt).

	Healthy						CHF				
Id	0	30	60	90	120	Id	0	30	60	90	120
1	7.2	14.4	15.1	15.5	23.6	1	2.6	7.8	10.8	13.5	16.2
2	2.8	8.5	11.5	15.2	18.9	2	2.8	8.8	12.4	15.1	19.0
3	6.3	12.6	15.8	18.2	22.5	3	3.8	8.7		15.6	19.4
4	3.9	9.6	13.6	17.2	20.8	4	3.2	8.6	10.7	14.1	17.2
5	8.2	17.4	19.4		25.4	5	3.6		12.8	17.6	20.4
6	4.4	10.7	12.4	16.5	20.7	6	4.3	7.7	8.2	9.9	
7	3.1	8.8	11.0	14.8	18.6	7	5.2	11.4	14.7	18.9	21.5
8	4.6		15.7	21.8	26.4	8	4.5	9.2	12.8	15.4	
9	3.3	7.9	13.0	17.4	22.0	9	4.9	13.2	17.9	20.9	27.2
10	5.0	16.2	19.1	22.8	26.4	10	4.5	8.7	11.9	13.6	17.3
11	5.6	11.7	13.1	16.7	19.8	11	4.3	13.6	16.6	20.9	22.9
12	5.4	13.1	15.5	19.6	23.3	12	5.0	9.9	12.3	15.4	18.3

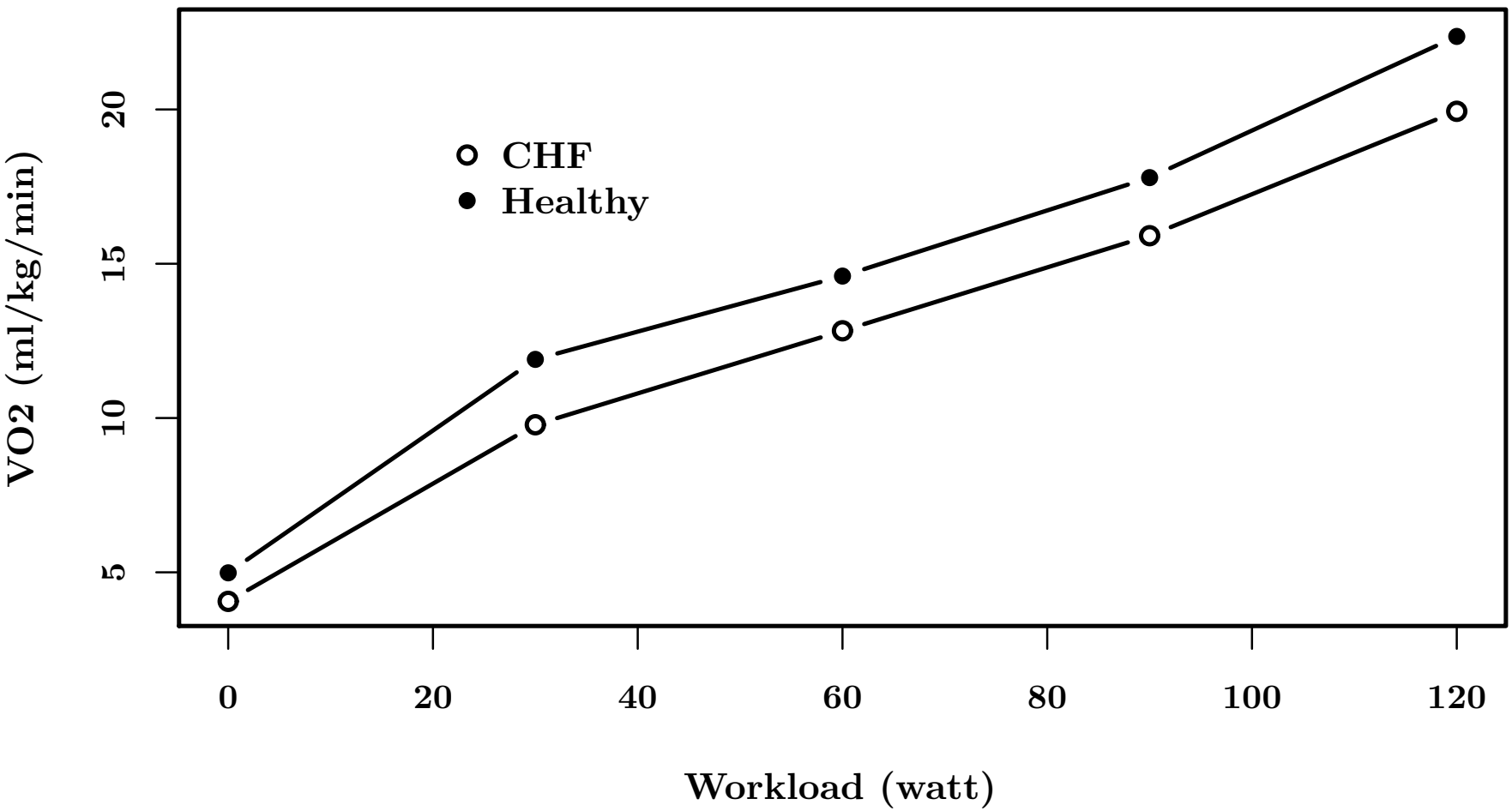
VO_2 in CHF patients and healthy: Individual curves



VO_2 in CHF patients and healthy: Grouped curves



VO_2 in CHF patients and healthy: Group mean curves



VO_2 in CHF patients and healthy: Parallel mean curves?

Testing for parallel mean curves using `mvtest` in **Stata** we get

$$\textbf{Test 1}: \qquad F = 2.05 \sim F(4, 13), \quad p = 0.15$$

The problem, of course, is that 6 out of the 24 subject have one missing value resulting in that about **25% of the observations are excluded** from this analysis.
It is possible to analyse all the data based on the multivariate repeated measurements model (in **Stata** this is done using `mixed`). The **price** is that we only get an **approximate test** (valid for large samples):

$$\textbf{Test 1}: \qquad \text{LR} = 9.54 \sim \chi^2(4), \quad p = 0.049$$

So, using all the information, we just conclude that there is evidence against the hypothesis of parallel mean curves, a conclusion that we did not reach using only observations from subjects with complete information.

VO_2 in CHF patients and healthy: Equal sd and correlations?

Observed standard deviations and correlations:

Healthy

$$\begin{pmatrix} 1.66 & & & & \\ 0.87 & 3.18 & & & \\ 0.73 & 0.92 & 2.70 & & \\ 0.21 & 0.66 & 0.85 & 2.62 & \\ 0.56 & 0.86 & 0.90 & 0.87 & 2.74 \end{pmatrix}$$

CHF

$$\begin{pmatrix} 0.86 & & & & \\ 0.54 & 2.05 & & & \\ 0.43 & 0.95 & 2.74 & & \\ 0.36 & 0.94 & 0.97 & 3.21 & \\ 0.54 & 0.90 & 0.97 & 0.93 & 3.27 \end{pmatrix}$$

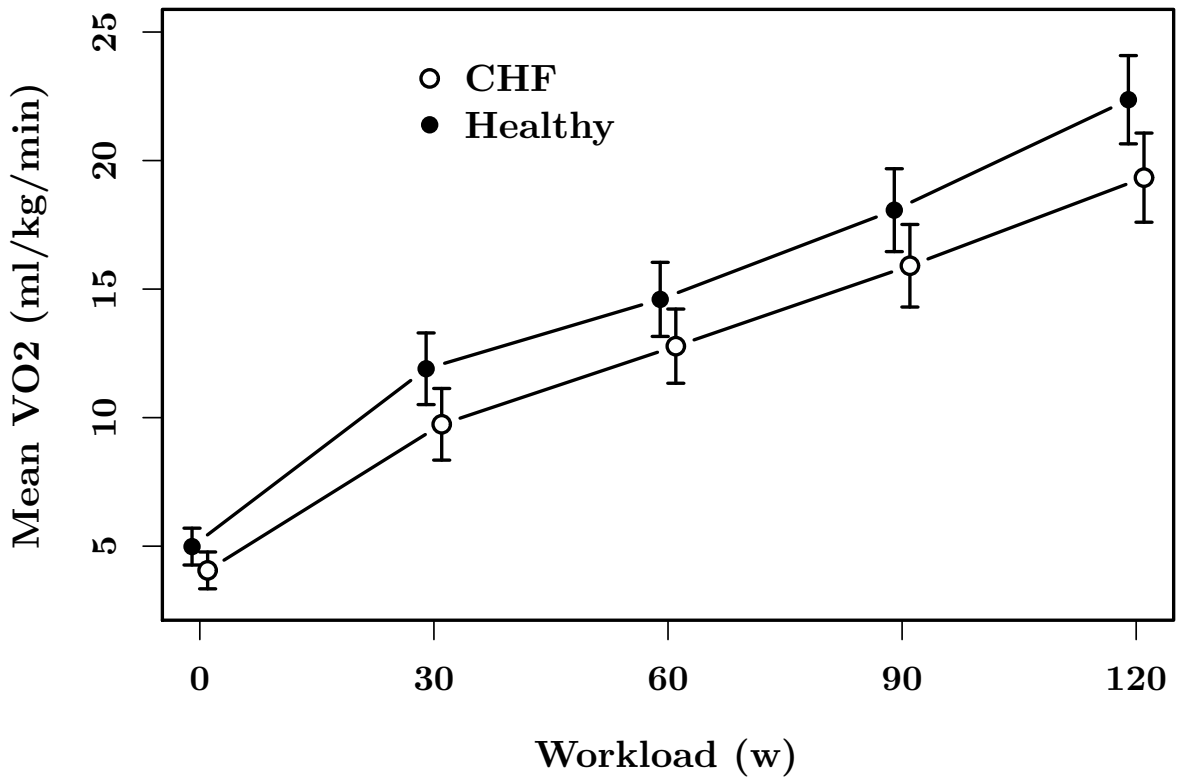
It is also possible to make an approximate test for the hypothesis of equal standard deviations and correlations in the two groups:

$$\text{LR} = 18.56 \sim \chi^2(15), \quad p = 0.23$$

If we had not been able to accept the hypothesis of equal standard deviations and correlations in the two groups, we could still have made an approximate test for parallel mean curves, but it would be a test with **low power**.

VO_2 in CHF patients and healthy: 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).

VO_2: Presenting estimated group differences

We have concluded that the group difference depends on the workload, so it makes sense to present the estimated group difference for each workload.

Workload (w)	Healthy−CHF (ml/kg/min)	95%-CI	p-value
0	0.9	[-0.1 , 1.9]	0.074
30	2.2	[0.2 , 4.1]	0.032
60	1.8	[-0.2 , 3.9]	0.080
90	2.2	[-0.1 , 4.4]	0.063
120	3.0	[0.6 , 5.5]	0.015

We see that the VO_2 is only slightly higher for healthy compared to CHF patients at rest (but almost significantly so because of the relatively small variation).

During exercise the difference between healthy and CHF patients increases, but the variation is large.

The problem with many time-points in repeated measurements analysis

Suppose that we consider a setup where each individual is measured at k time-points:

$$t_1, t_2, \ldots, t_k$$

If we want to analyse the data using a multivariate repeated measurements model, we need to estimate:

- k standard deviations
- $\frac{k(k-1)}{2}$ correlations

In total there are $\frac{k(k+1)}{2}$ standard deviations and correlations to be estimated, and if k is large (more than 6-7) this may not be possible in practice unless you have a very large number of patients.

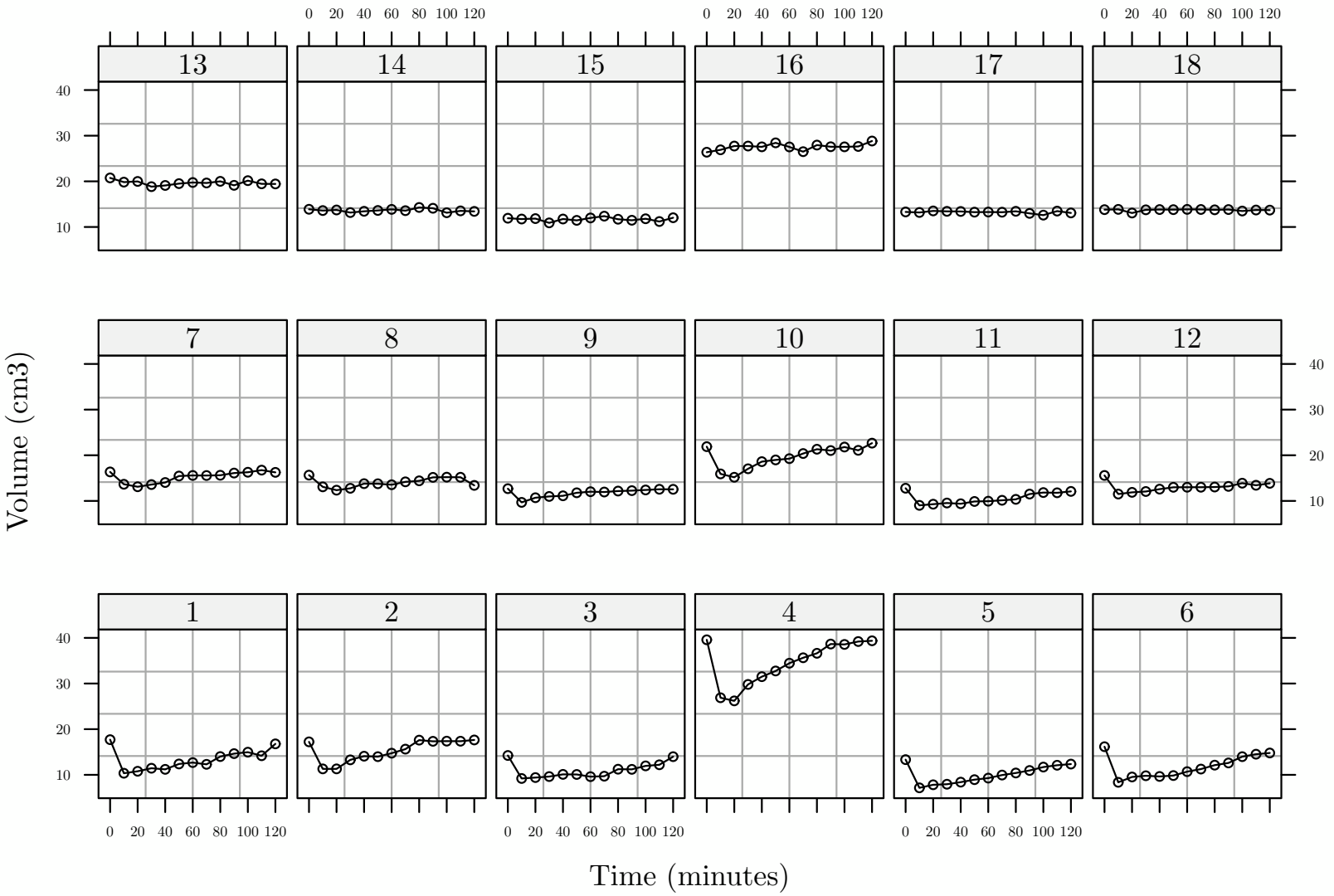
Let us consider an example that illustrates this problem.

Example: The effect of cholagogues on gallbladder volume

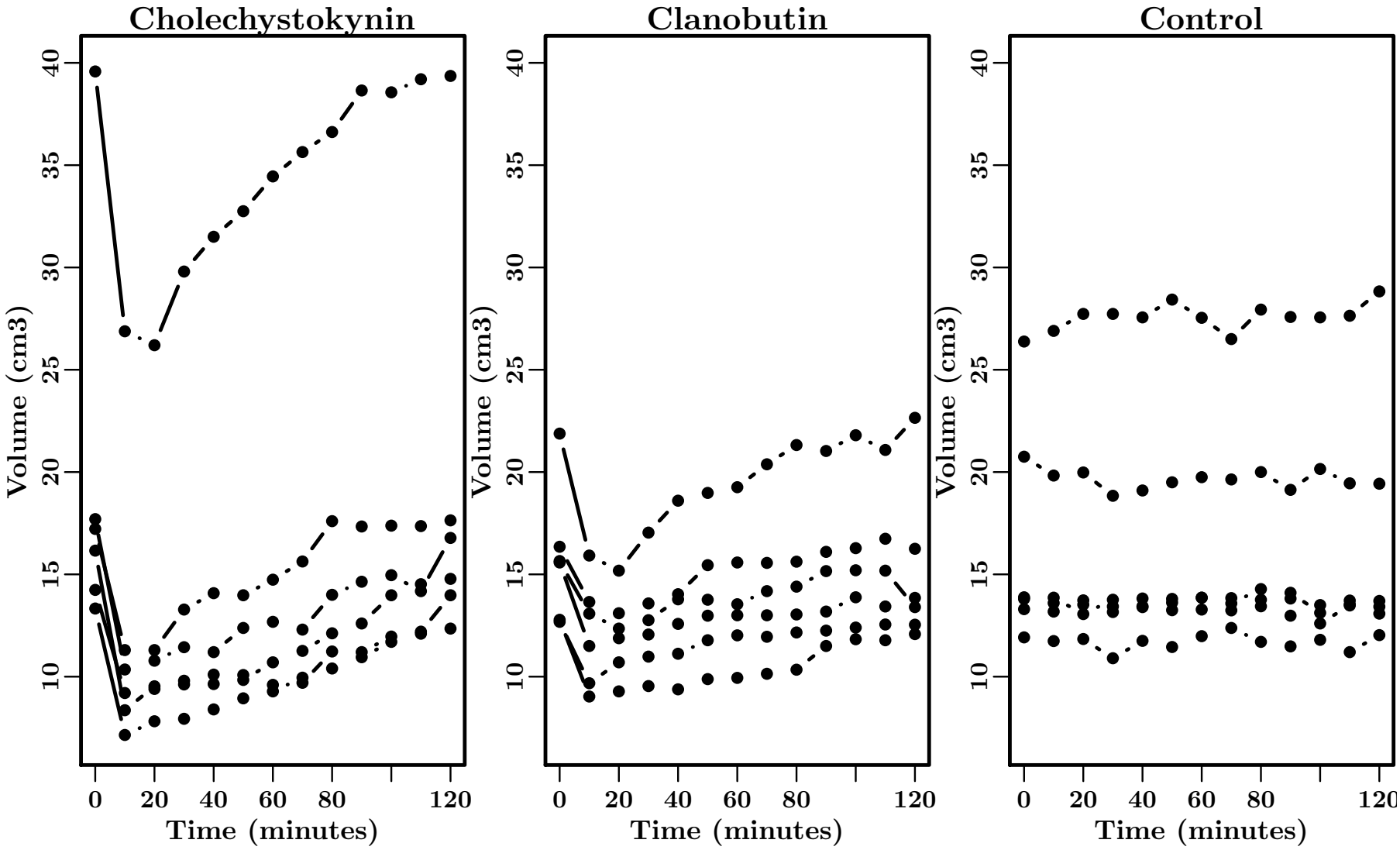
Data: At 10 minute intervals three groups of healthy dogs had the gallbladder volume measured (in fact the same 6 dogs in the three groups, but we will ignore this). Groups 1 and 2 were given **cholechystokynin** and **clanobutin** just after time 0, respectively. Group 3 was a **control** group.

Group	Dog	0	10	20	30	40	50	60	70	80	90	100	110	120
1	1	17.70	10.35	10.78	11.44	11.20	12.38	12.68	12.30	14.00	14.64	14.96	14.18	16.78
1	2	17.22	11.30	11.30	13.28	14.08	13.98	14.74	15.63	17.60	17.34	17.38	17.36	17.64
1	3	14.24	9.20	9.40	9.62	10.10	10.08	9.60	9.70	11.23	11.20	11.96	12.20	13.98
1	4	39.58	26.88	26.20	29.80	31.50	32.75	34.45	35.64	36.62	38.65	38.56	39.20	39.36
1	5	13.33	7.15	7.82	7.94	8.40	8.94	9.28	9.95	10.40	10.95	11.70	12.10	12.35
1	6	16.16	8.36	9.53	9.80	9.64	9.84	10.70	11.26	12.12	12.60	13.98	14.52	14.78
2	7	16.35	13.65	13.10	13.58	14.03	15.45	15.58	15.56	15.62	16.10	16.28	16.74	16.25
2	8	15.65	13.08	12.35	12.76	13.78	13.76	13.54	14.18	14.40	15.16	15.20	15.18	13.40
2	9	12.68	9.68	10.70	10.98	11.12	11.78	12.02	11.95	12.16	12.25	12.40	12.55	12.54
2	10	21.88	15.92	15.18	17.04	18.60	18.98	19.26	20.38	21.32	21.03	21.80	21.08	22.65
2	11	12.78	9.03	9.28	9.54	9.38	9.88	9.94	10.14	10.34	11.50	11.83	11.78	12.08
2	12	15.58	11.50	11.88	12.06	12.58	12.98	13.00	13.00	13.04	13.18	13.88	13.43	13.85
3	13	20.75	19.83	19.98	18.84	19.10	19.50	19.75	19.64	20.00	19.13	20.15	19.45	19.43
3	14	13.88	13.60	13.73	13.16	13.44	13.62	13.86	13.58	14.28	14.10	13.12	13.53	13.42
3	15	11.92	11.74	11.84	10.90	11.75	11.45	11.98	12.38	11.70	11.48	11.80	11.20	12.03
3	16	26.38	26.90	27.73	27.73	27.56	28.43	27.54	26.50	27.94	27.58	27.56	27.64	28.83
3	17	13.30	13.18	13.52	13.43	13.40	13.25	13.28	13.24	13.44	12.98	12.60	13.48	13.08
3	18	13.80	13.86	13.06	13.76	13.82	13.80	13.86	13.84	13.76	13.82	13.50	13.72	13.70

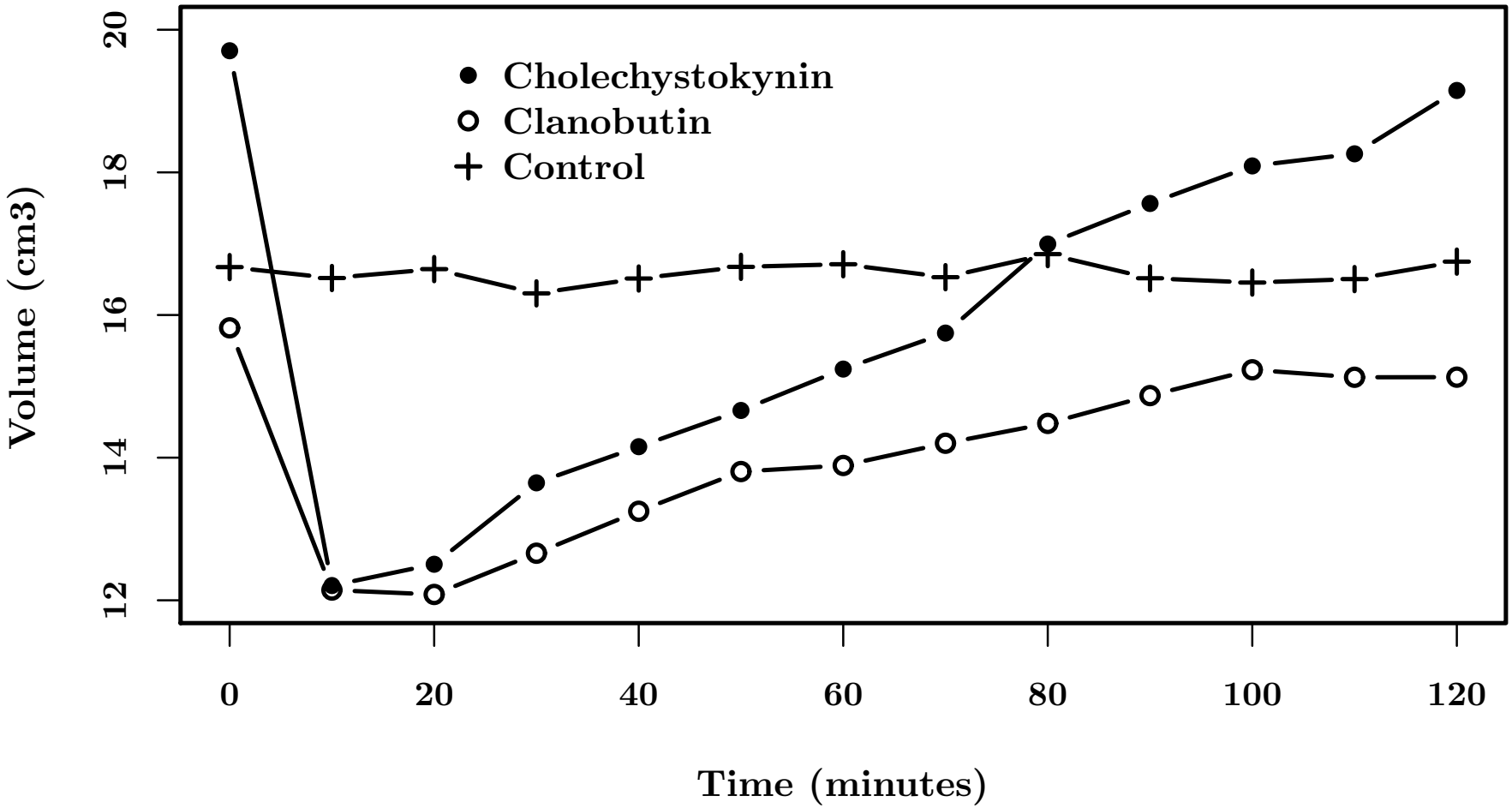
Gallbladder volume: Separate individual curves



Gallbladder volume: Grouped individual curves



Gallbladder volume: Grouped individual curves



Multivariate repeated measurements analysis: Problems

In this example we have $k = 13$ time-points and so a total of $\frac{13 \cdot 14}{2} = 91$ standard deviations and correlations to estimate **in each group**, but only $6 \cdot 13 = 78$ observations. That is of course a problem.
We could analyze the data using the multivariate repeated measurements ANOVA assuming **equal standard deviations and correlations** in the three groups, but:

- We cannot test whether this is reasonable (too many quantities to estimate)
- Any test based on this model will have **very low power** (almost all the information is used for estimating correlations)

Testing for parallel mean curves we get

$$\textbf{Test 1}: \qquad F = 1.95 \sim F(24, 8), \quad p = 0.17$$

So, a little surprisingly, we accept that we have parallel mean curves. With this many time-points, however, this test is almost **useless**!

Alternative to the multivariate repeated measurements analysis

We need to simplify the analysis (make some further assumptions) in order to reduce the number of standard deviations and correlations to be estimated.

A much used method to achieve this is the so-called **univariate repeated measurements ANOVA**.

The **univariate repeated measurements ANOVA** is the simplest statistical model that still takes into account that we have positively correlated observations corresponding to the same individual.

In the **univariate repeated measurements ANOVA** we only have **two parameters** describing standard deviations and correlations no matter how many time-points we have.

The univariate repeated measurement analysis

In the **univariate repeated measurements ANOVA** we additionally assume that

- The standard deviation is the same at all time-points

$$\sigma_{\mathrm{T}} = \text{The total standard deviation}$$

- The correlations between any two different measurements on the same subject are equal

$$\rho = \frac{\sigma_{\mathrm{B}}^2}{\sigma_{\mathrm{T}}^2}$$

Here σ_{B} is the standard deviation corresponding to the **between subject variation** and $\sigma_{\mathrm{T}}^2 = \sigma_{\mathrm{B}}^2 + \sigma_{\mathrm{W}}^2$:

The total variation $=$

The between subject variation $+$ **The within subject variation**

That is, it is to equivalent to consider σ_{B}, σ_{W} and σ_{T}, ρ.

Gallbladder volume: The univariate repeated measurements ANOVA

Testing for parallel mean curves using the **univariate repeated measurements ANOVA** results in

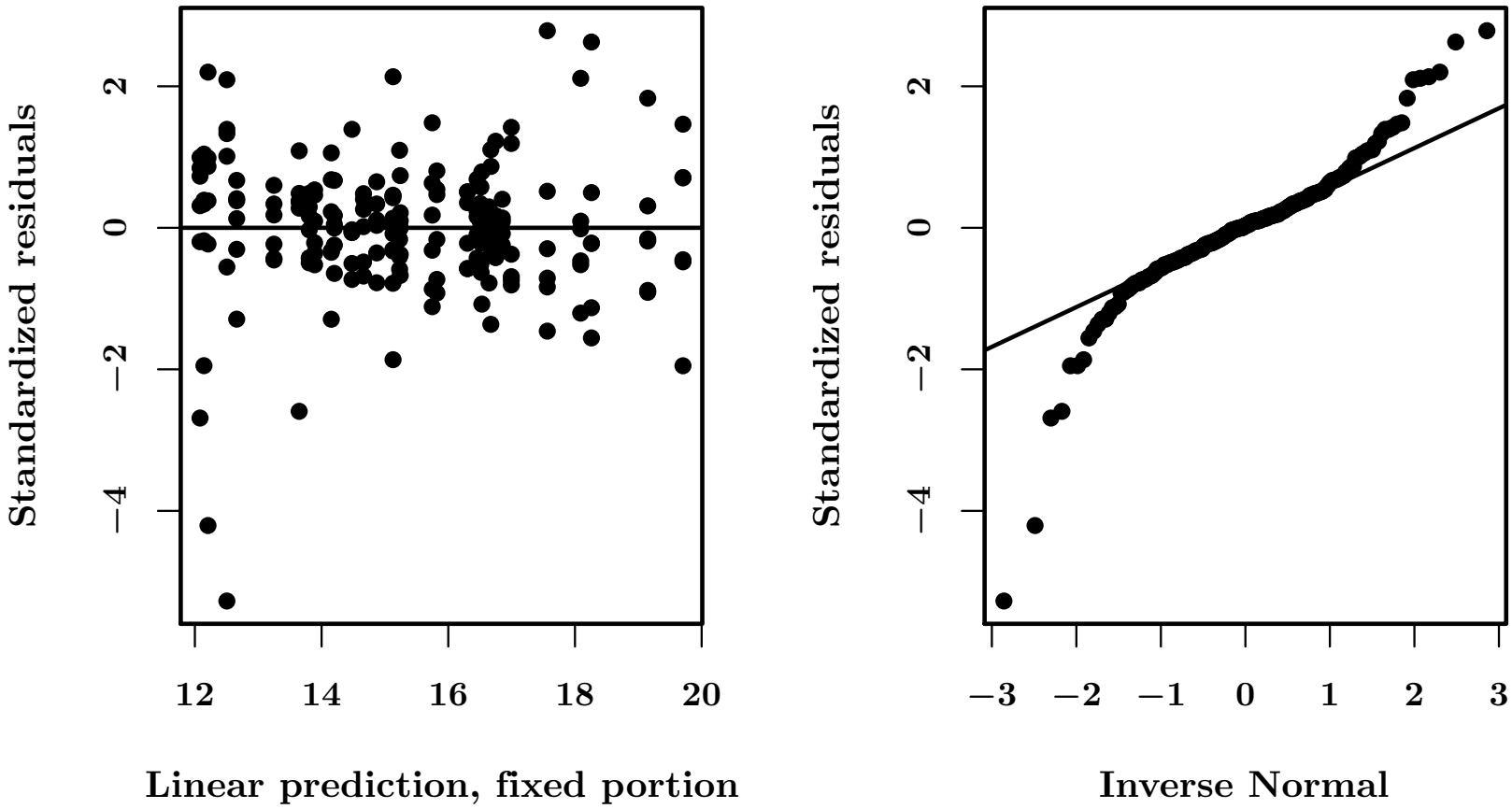
$$\textbf{Test 1}: \qquad F = 10.92 \sim F(24, 180), \quad p < 0.0001$$

We conclude that there is clear evidence against the hypothesis of parallel mean curves.

Several people have suggested corrections to the F-statistic above to take into account that the **univariate repeated measurements ANOVA** is probably too simple a model to adequately describe the data.

Stata, for example, provides three different corrections to the F-statistic corresponding to the test for parallel mean curves. In all cases the conclusion is unaltered.

Gallbladder volume: Model validation in the univariate repeated measurements ANOVA



There seems to be **some deviation** from the univariate repeated measurements model.

Gallbladder volume: Problems with the univariate repeated measurements ANOVA

Does it make sense to assume that the standard deviations and correlations are the same in the three groups? Let us look at the **standard deviations** in the three groups:

Group	0	10	20	30	40	50	60	70	80	90	100	110	120
1	9.9	7.3	6.8	8.1	8.7	9.1	9.6	10.0	9.9	10.6	10.2	10.4	10.1
2	3.4	2.6	2.0	2.6	3.1	3.2	3.2	3.6	3.8	3.5	3.6	3.4	4.0
3	5.7	5.8	6.1	6.2	6.0	6.4	5.9	5.5	6.1	6.0	6.2	6.1	6.5

Even though the standard deviations are fairly constant in each group they are very different between groups.
It is possible to make the univariate repeated measurements ANOVA but to allow the within and between subject variation to vary between groups.

Gallbladder volume: Testing in the univariate repeated measurements ANOVA with variations depending on group

Testing for parallel mean curves in the univariate repeated measurements ANOVA with the within and between subject variation varying between groups:

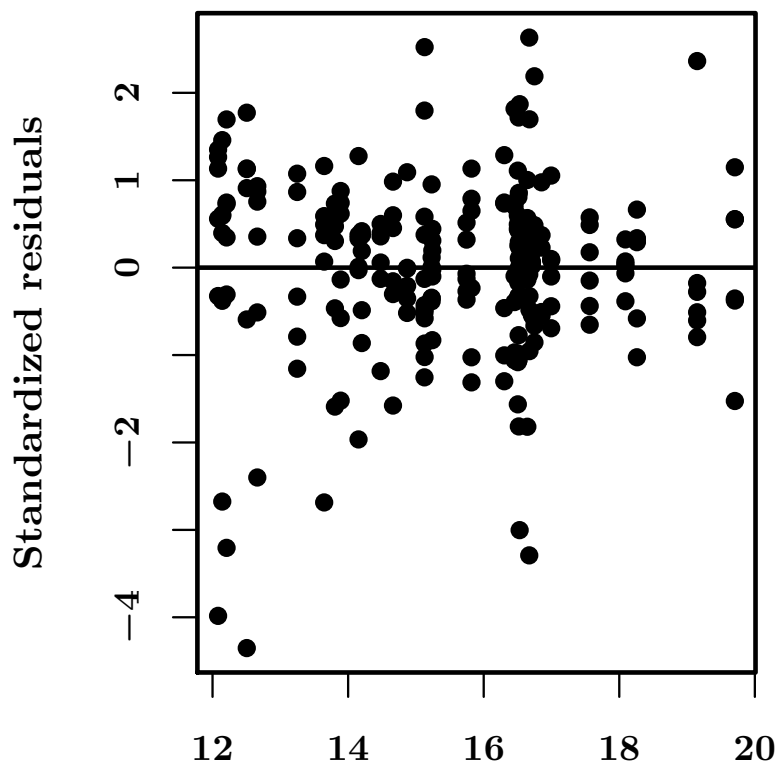
$$\textbf{Test 1}: \qquad \text{LR} = 213.27 \sim \chi^2(24), \quad p < 0.0001$$

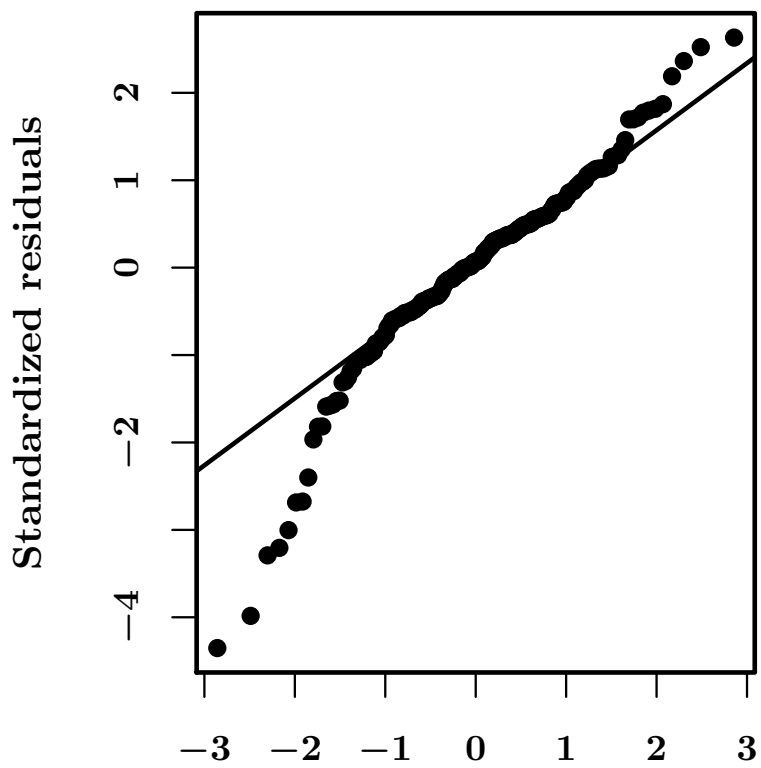
We conclude that there is clear evidence against the hypothesis of parallel mean curves.

Estimates for the within and between subject variations are given by:

Group	Between (σ_B)	Within (σ_W)	Total (σ_T)	Correlation
1	9.27	1.34	9.37	0.98
2	3.18	0.74	3.27	0.95
3	6.03	0.44	6.05	0.99

Gallbladder volume: Model validation when the variation depends on group





There is **less clear deviation** from the univariate repeated measurements model with within and between subject variations varying with group.

Gallbladder volume: Exponentially decreasing correlation between observations

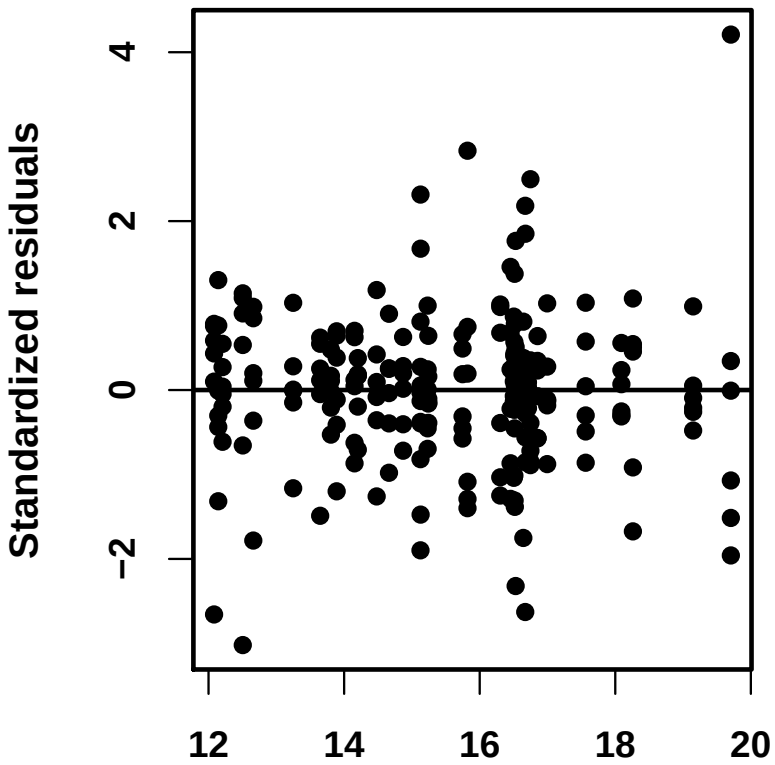
As already noted the **multivariate repeated measurement model** and the **univariate repeated measurement model** represent the two extremes in the spectre of statistical models that takes into account the possible correlation between observations on the same individual.

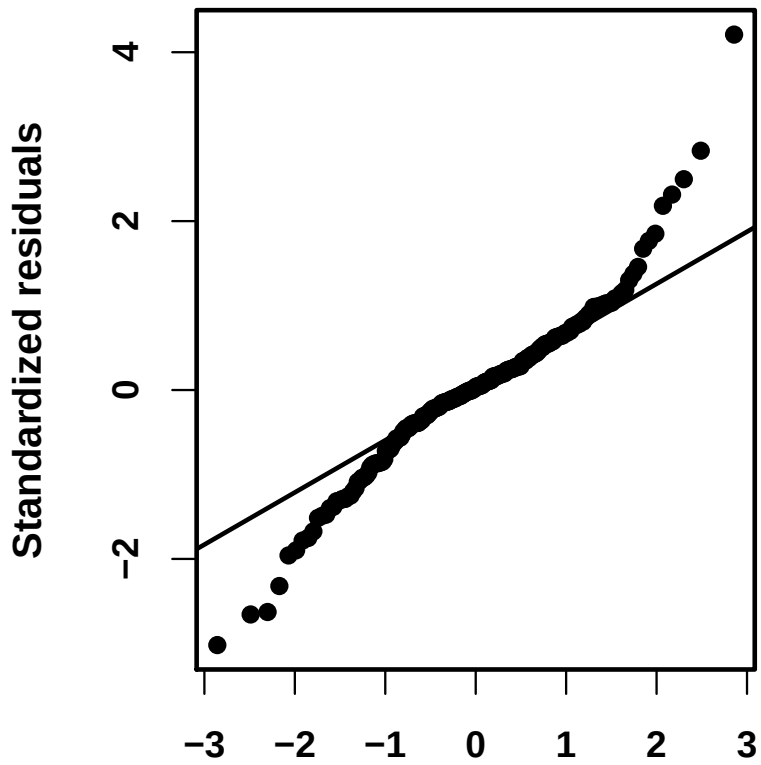
Multivariate Exponential Univariate

It is possible to do something in between:

- **Exponentially decreasing correlation** between observations.
- Observations close to each other in time are more correlated than observations far from each other.
- In this example we still get a clear rejection of parallel mean curves (test 1): LR $= 197.76$, p < 0.0001.

Gallbladder volume: Model validation with exponentially decreasing correlation depending on group





There is **no clear deviation** from the univariate repeated measurements model with exponentially decreasing correlation varying with group.