

Análisis de datos longitudinales

Grado en Estadística

Tema 3 – Sesión 8

Análisis de datos continuos (I)

Juan R González

Departamento de Matemáticas, UAB

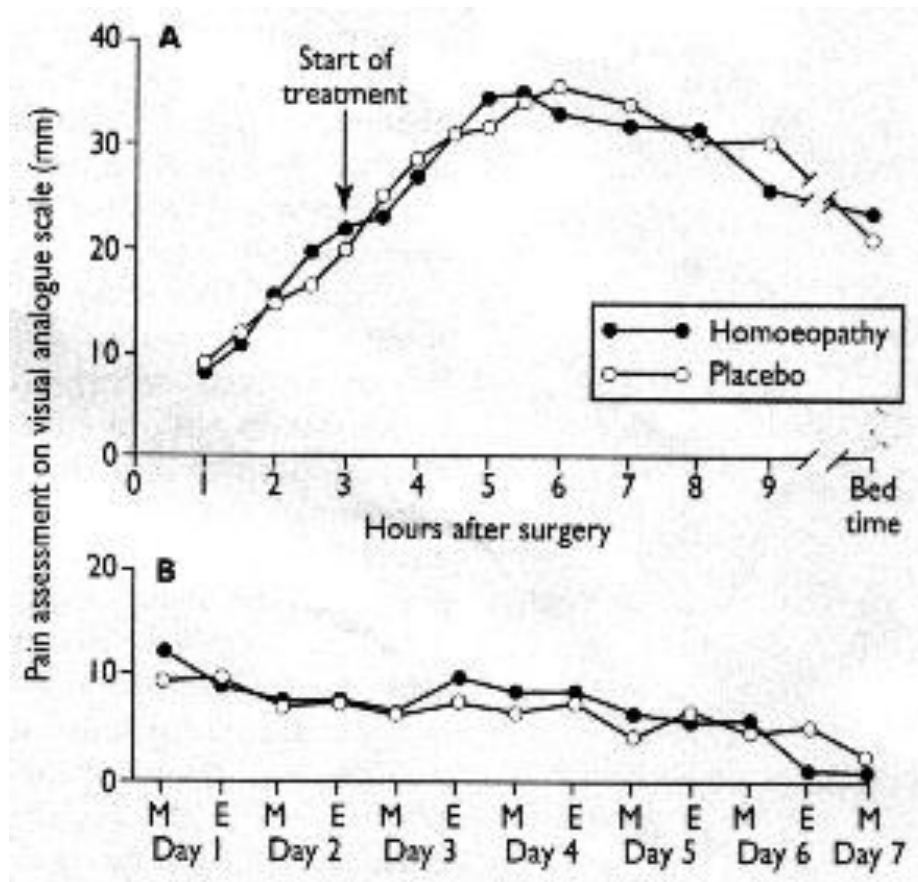
Insituto de Salud Global Barcelona, ISGlobal

Esquema

- Ejemplos de análisis
- Visualización de datos
- Hipótesis del modelo
- Missing data
- Estratégias de análisis
 - ANCOVA para la medida final, ajustando por diferencias basales (`end-point analysis)
 - ANOVA con medidas repetidas: “aproximación univariante”
 - ANOVA “Multivariante” (MANOVA)
 - GEE
 - Modelos mixtos
 - Modelización del cambio
- Interpretación de resultados

Ejemplos

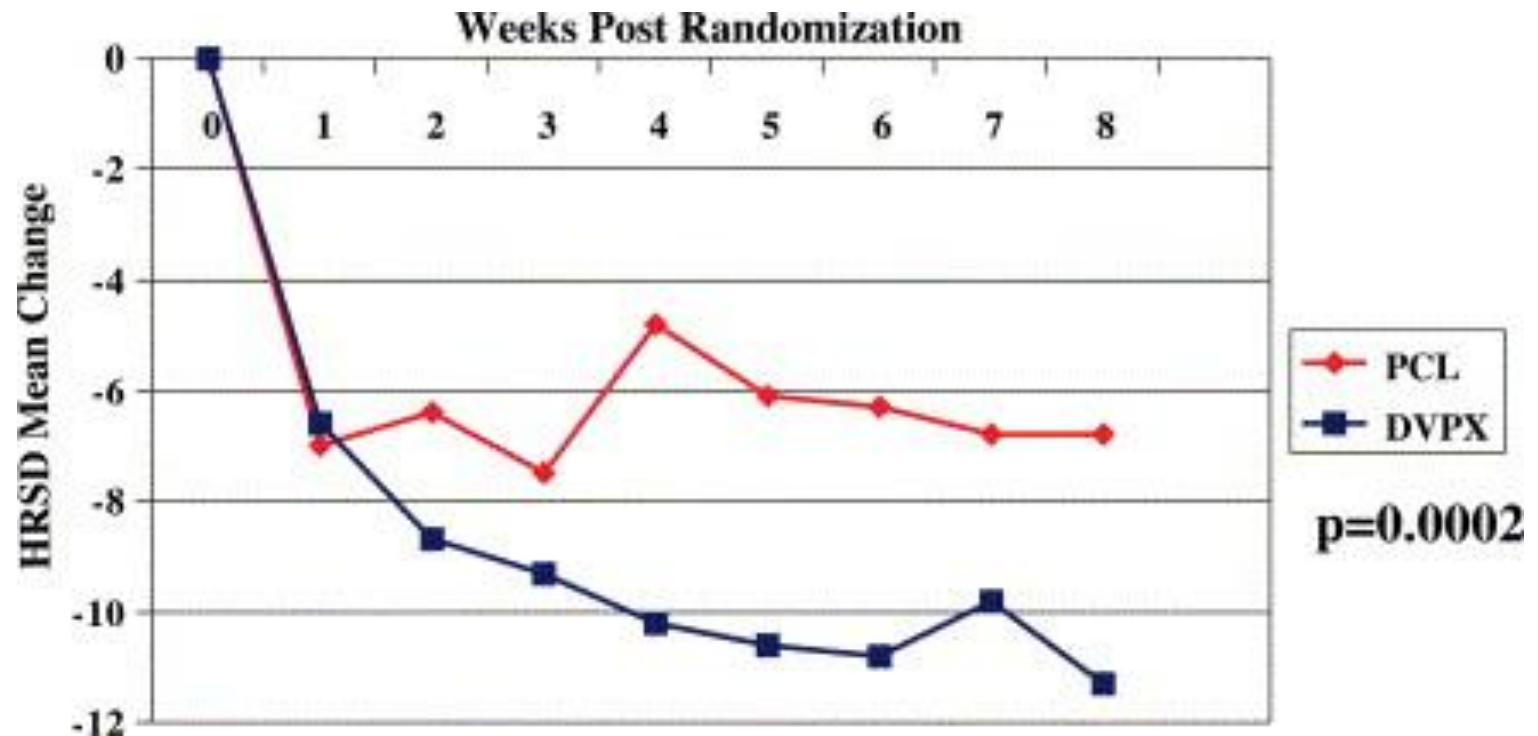
Mean pain assessments by visual analogue scales (VAS)



Day of surgery

Days 1-7 after surgery
(morning and evening)

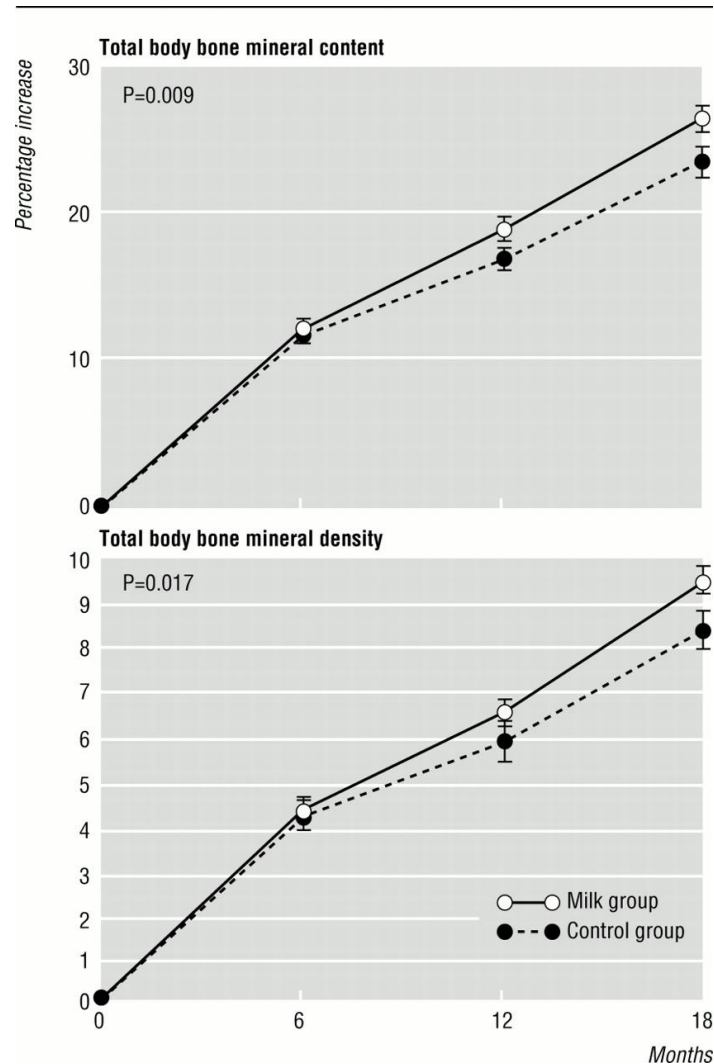
Ejemplos



Davis et al. "Divalproex in the treatment of bipolar depression: A placebo controlled study." J Affective Disorders 85 (2005) 259-266.

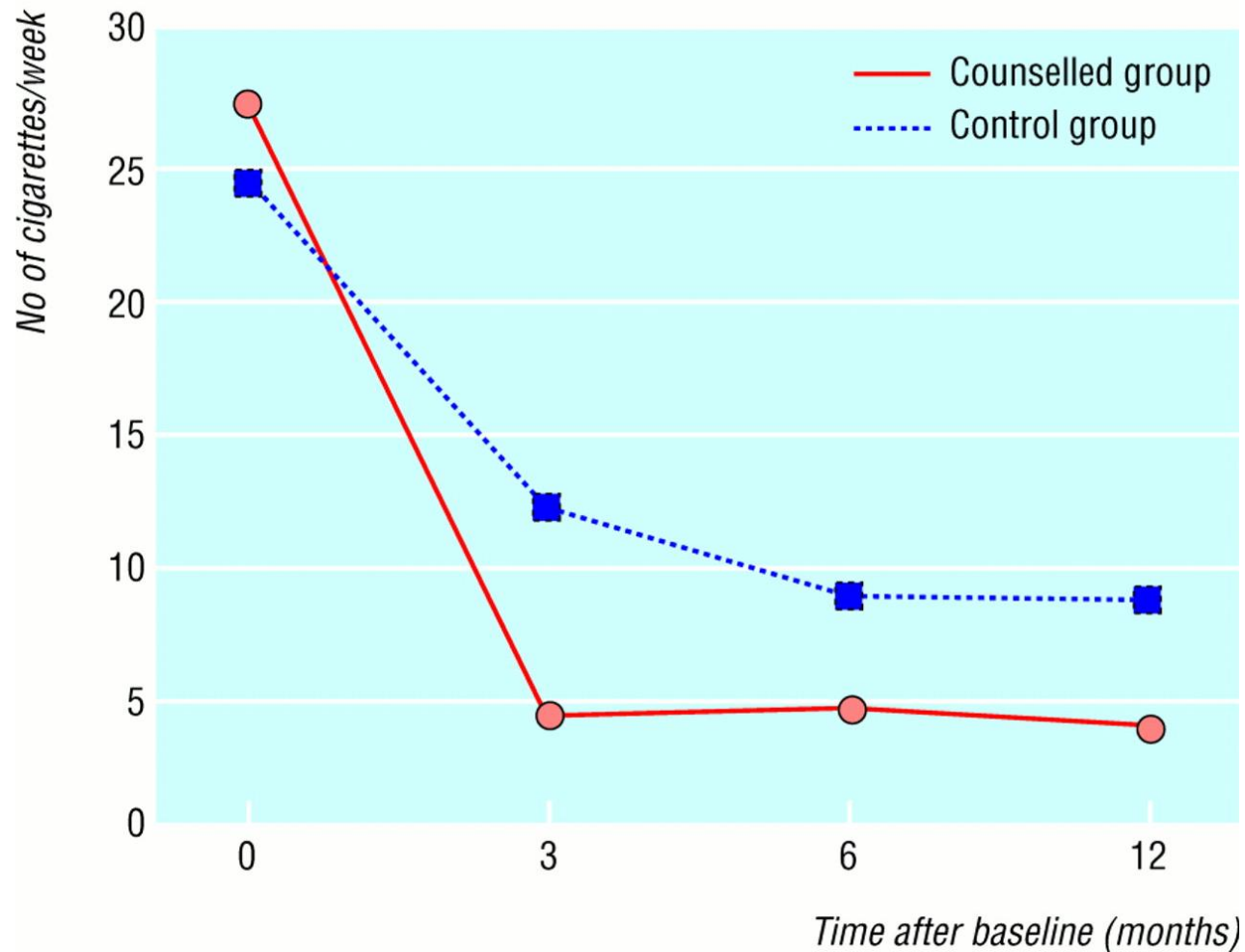
Ejemplos

P values are for the differences between groups by repeated measures analysis of variance



Mean (SE) percentage increases in total body bone mineral and bone density over 18 months.

Ejemplos



Datos Longitudinales: 'formato ancho'

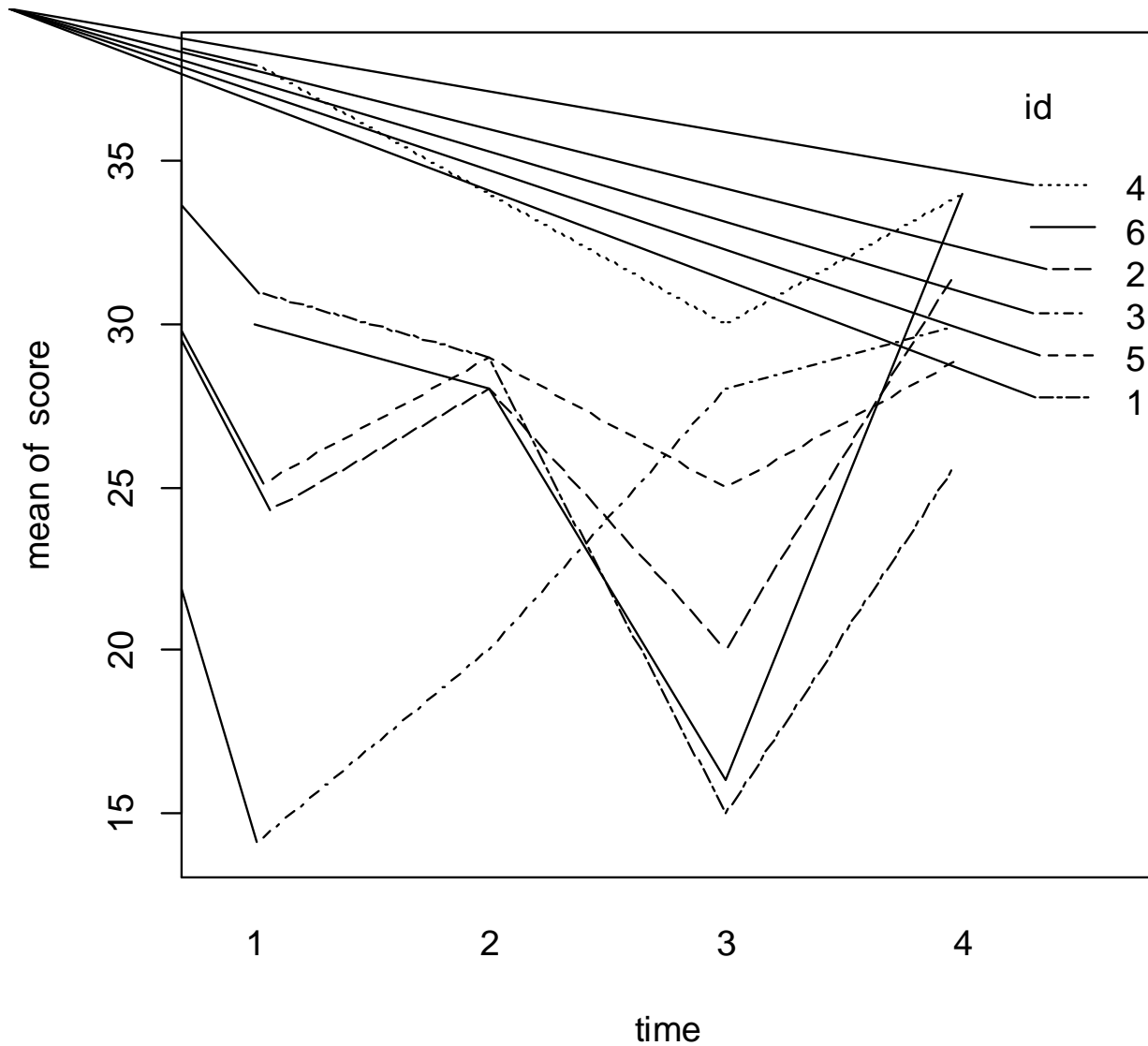
id	time1	time2	time3	time4
1	31	29	15	26
2	24	28	20	32
3	14	20	28	30
4	38	34	30	34
5	25	29	25	29
6	30	28	16	34

Datos Longitudinales: 'formato largo'

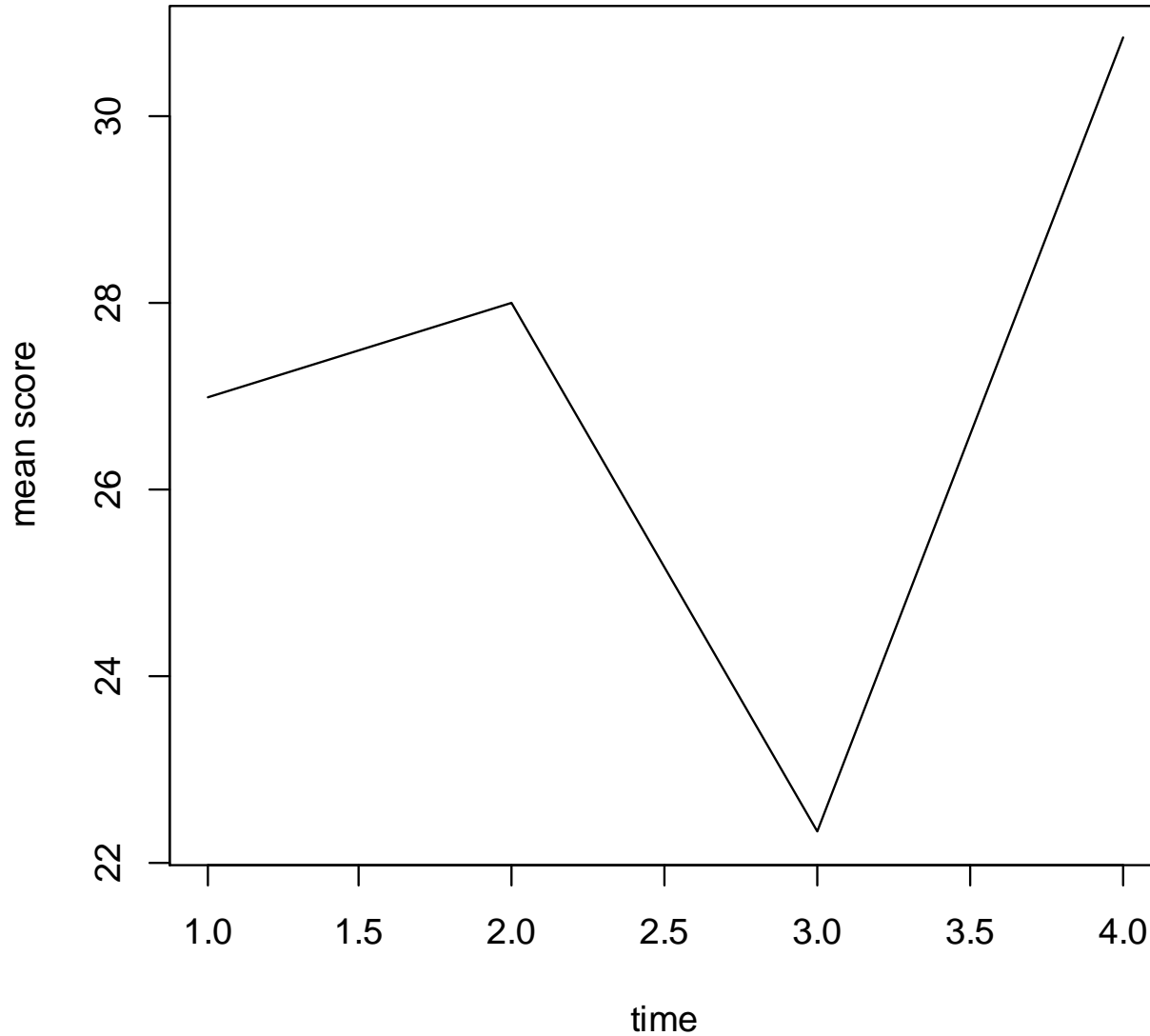
id	time	score
1	1	31
1	2	29
1	3	15
1	4	26
2	1	24
2	2	28
2	3	20
2	4	32
3	1	14
3	2	20
3	3	28
3	4	30

id	time	score
4	1	38
4	2	34
4	3	30
4	4	34
5	1	25
5	2	29
5	3	25
5	4	29
6	1	30
6	2	28
6	3	16
6	4	34

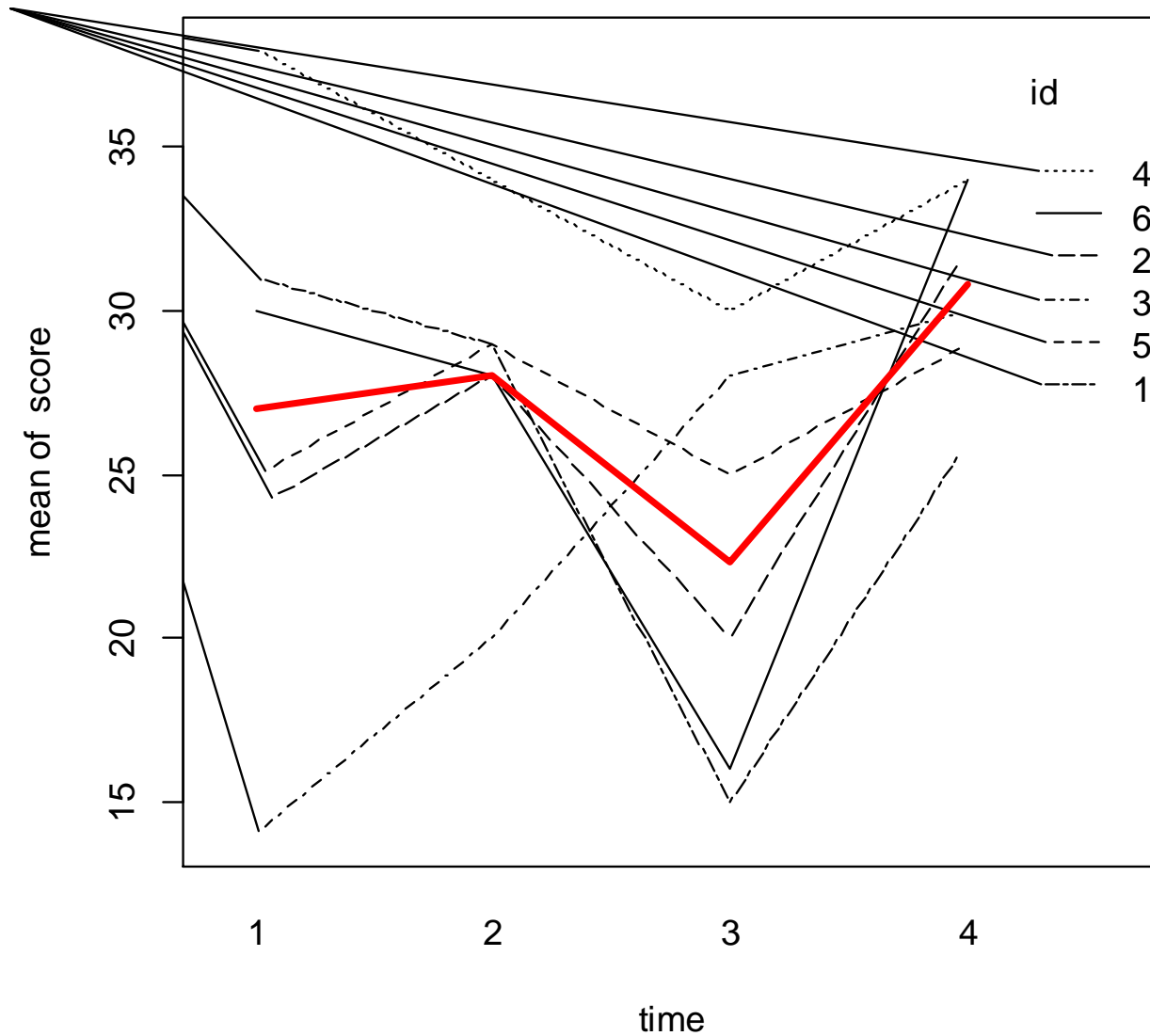
Datos: 'profile plot'



Datos: 'mean plot'



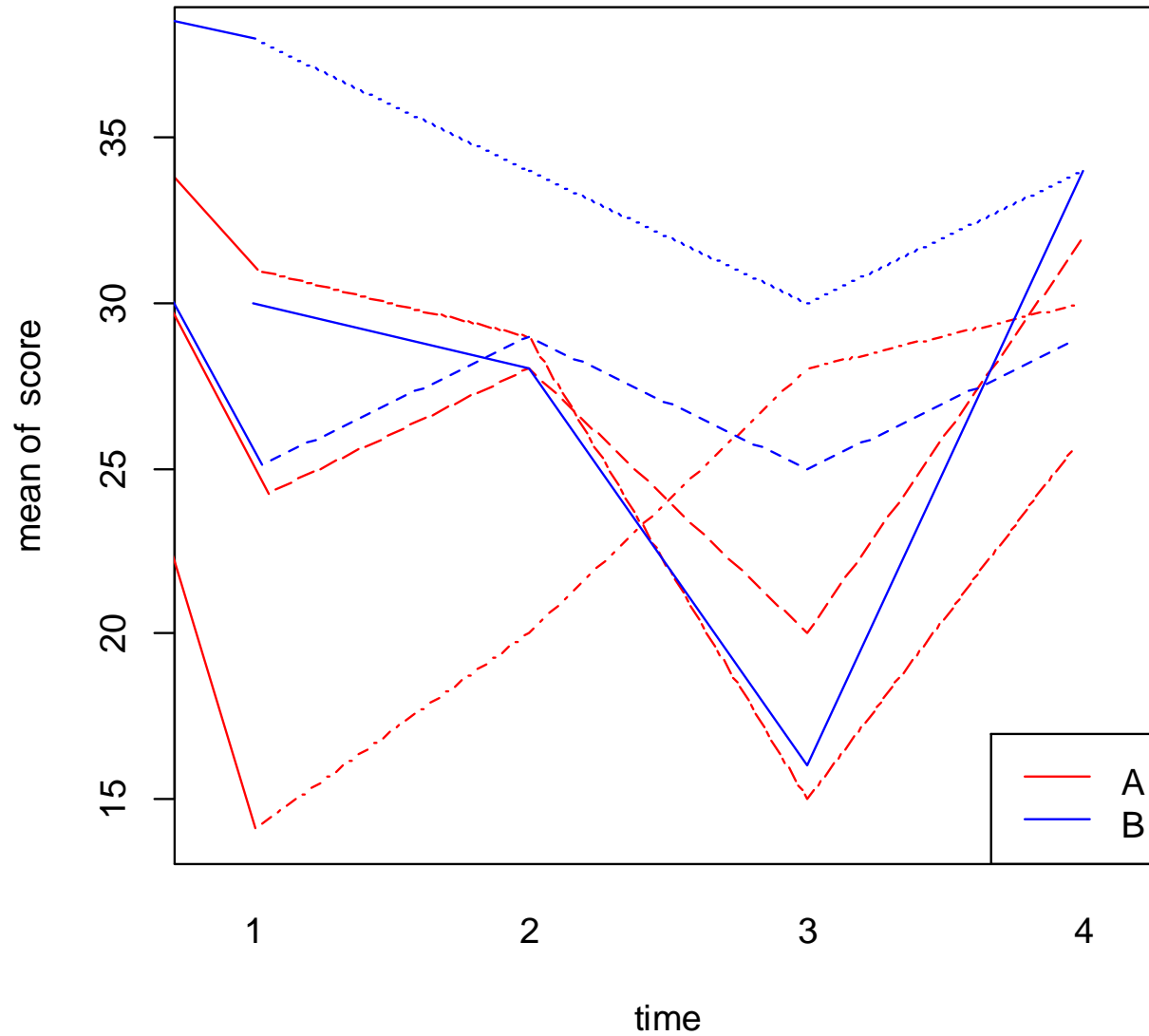
Datos: 'superimposed'



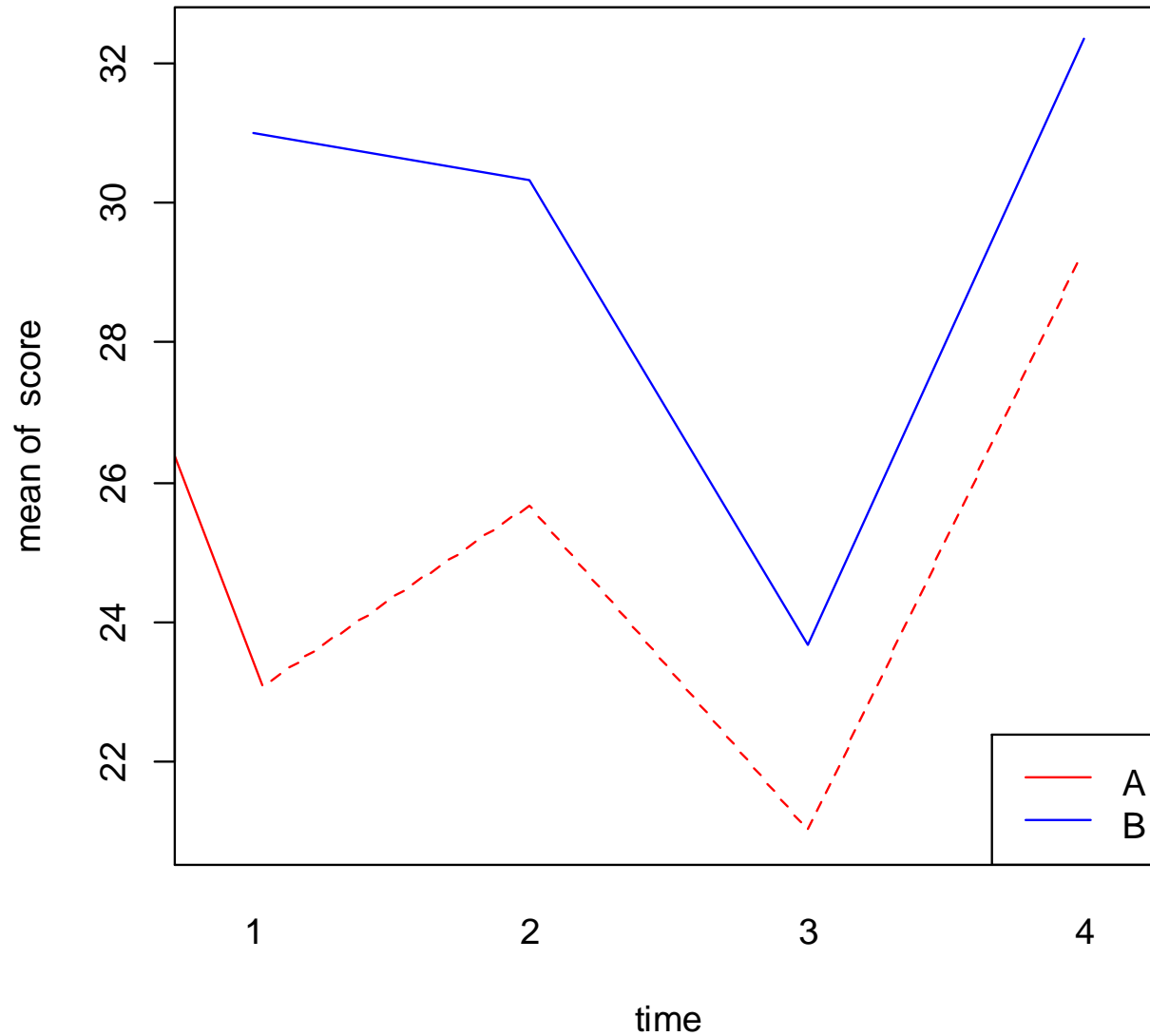
Datos: 'dos grupos'

id	group	time1	time2	time3	time4
1	A	31	29	15	26
2	A	24	28	20	32
3	A	14	20	28	30
4	B	38	34	30	34
5	B	25	29	25	29
6	B	30	28	16	34

Datos: 'dos grupos'



Datos: 'dos grupos'



Posibles preguntas científicas

- En promedio, ¿Existen diferencias estadísticamente significativas entre cada tiempo?
- Según el gráfico parece que hay diferencias entre los puntos 3 y 4
 - En promedio, ¿hay cambios significativos desde el punto inicial (baseline)?
 - ¿Son diferentes los dos grupos en algún tiempo?
- ¿Difieren los dos grupos en la respuesta a lo largo del tiempo? [IMP]
- Según el gráfico el perfil de respuesta es similar a lo largo del tiempo, aunque los grupos son más similares al final del estudio

Estrategias de análisis

Tradicionales

Estrategia 1: ANCOVA para la medida final, ajustando por diferencias basales ('end-point analysis')

Estrategia 2: ANOVA con medidas repetidas: "aproximación univariante"

Estrategia 3: ANOVA "Multivariante" (MANOVA)

Nuevos

Estrategia 4: GEE

Estrategia 5: Modelos mixtos

Estrategia 6: Modelización del cambio

Estrategias de análisis

Table 1. Comparison of Traditional and Mixed-Effects Approaches for the Analysis of Repeated-Measures Data

	End-Point Analysis	rANOVA	rMANOVA	Mixed-Effects Analysis
Complete data required on every subject	Yes	No*	Yes	No
Possible effect of omitting subjects with missing values	Sample bias	Sample bias	Sample bias	Not applicable†
Possible effects of imputation of missing data	Estimation bias	Estimation bias	Estimation bias	Not applicable†
Subjects measured at different time points	Yes	No	No	Yes
Description of time effect	Simple	Flexible	Flexible	Flexible
Estimation of individual trends	No	No	No	Yes
Restrictive assumptions about correlation pattern	Not applicable	Yes	No	No
Time-dependent covariates	No	Yes	No	Yes
Ease of implementation	Very easy	Easy	Easy	Hard
Computational complexity	Low	Low	Medium	High

Abbreviations: rANOVA, univariate repeated-measures analysis of variance; rMANOVA, multivariate repeated-measures analysis of variance.

*Subjects with missing data are often omitted from the analysis.

†It is not necessary to omit subjects with missing values from the analysis or to impute missing values.

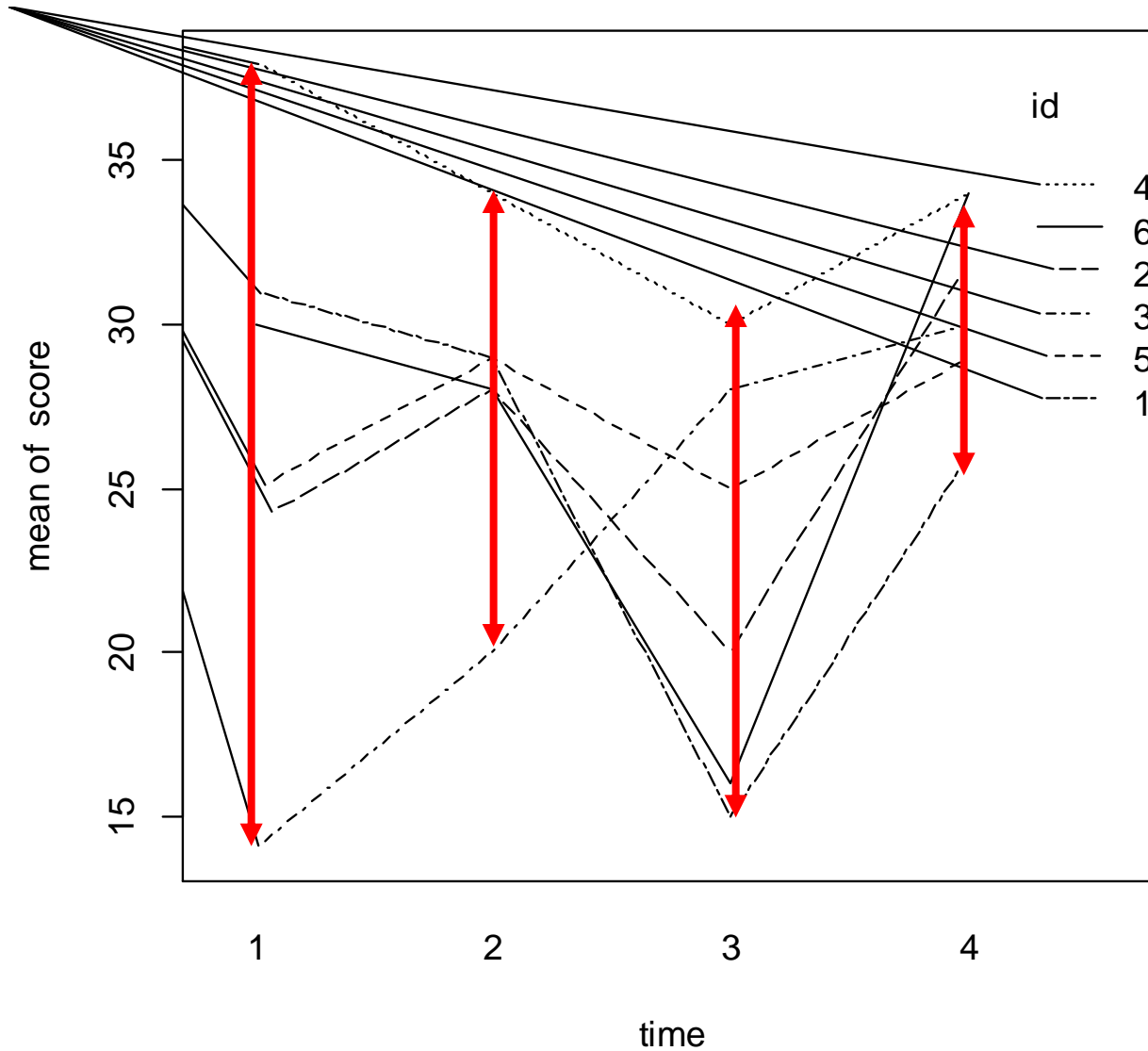
Consideraciones previas

- Espaciado tiempos intervalos
 - ANOVA y MANOVA con medidas repetidas requieren que todos los sujetos estén medidos en los mismos intervalos de tiempos. Los gráficos que hemos hecho también asumen esta hipótesis
 - MANOVA pondera todos los intervalos de la misma forma (equiespaciados)
- Hipótesis del modelo
 - TODAS las estrategias asumen que los datos están normalmente distribuidos y homogeneidad de las varianzas
 - Aunque todas las estrategias son robustas para set de datos grandes
 - ANOVA univariante para medidas repetidas asume esfericidad o **compound symmetry**
- Missing Data
 - Todos los métodos tradicionales requieren la imputación de los datos faltantes (o el análisis de datos completos)

‘Compound symmetry’

- Compound symmetry requiere:
 - a) La varianza de la variable respuesta debe ser la misma en cada punto temporal
 - a) La correlación entre las medidas repetidas debe ser igual según el intervalo de tiempo entre medidas


a) Igual varianza



a) Igual varianza

id	time1	time2	time3	time4
1	31	29	15	26
2	24	28	20	32
3	14	20	28	30
4	38	34	30	34
5	25	29	25	29
6	30	28	16	34

65.60 20.40 39.47 9.77



The diagram shows four arrows originating from the columns of the table above. The first arrow points from the 'time1' column to the value 65.60. The second arrow points from the 'time2' column to the value 20.40. The third arrow points from the 'time3' column to the value 39.47. The fourth arrow points from the 'time4' column to the value 9.77.

b) Igual correlación

	time1	time2	time3	time4
time1	1.00000	0.94035	-0.14150	0.28445
time2	0.94035	1.00000	-0.02819	0.26921
time3	-0.14150	-0.02819	1.00000	0.27844
time4	0.28445	0.26921	0.27844	1.00000

NO tenemos correlaciones iguales!

time1 y time2 están altamente correlacionados, pero ...

time1 y time3 están *inversamente* correlacionados!

‘Compound symmetry’ sería ...

time1	time2	time3	time4	
time1	1.00000	-0.04878	-0.04878	-0.04878
time2	-0.04878	1.00000	-0.04878	-0.04878
time3	-0.04878	-0.04878	1.00000	-0.04878
time4	-0.04878	-0.04878	-0.04878	1.00000

Missing data

- Importante imputar los datos. En caso contrario, hay que eliminar toda esa observación
- Con datos faltantes, cambios en la media en el tiempo, puede estar reflejando patrones de 'no-información'; no se puede comparar el punto 1 con 50 observaciones con el punto 3 con 30
- Aprenderemos la aproximación clásica "last observation carried forward" por simplicidad
- Existen otras estrategias más sofisticadas que pueden ser más apropiadas (sería un curso de missing data)

Missing data: LOCF

<u>Subject</u>	HRSD 1	HRSD 2	HRSD 3	HRSD 4
Subject 1	20	13		
Subject 2	21	21	20	19
Subject 3	19	18	10	6
Subject 4	30		25	23

Missing data: LOCF

<u>Subject</u>	HRSD 1	HRSD 2	HRSD 3	HRSD 4
Subject 1	20	13	13	13
Subject 2	21	21	20	19
Subject 3	19	18	10	6
Subject 4	30	30	25	23

Estrategia 1: End-point analysis

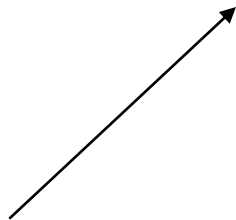
- Elimina el problema de tener medidas repetidas, considerando un único punto temporal (el último)
- Ignora datos intermedios de forma completa
- Se pregunta sobre si las medias de los dos grupos son diferentes en el tiempo final, ajustando por diferencias al principio (baseline) – utiliza ANCOVA
- Otros investigadores (biología) suelen comparar los grupos en cada punto de seguimiento, pero esta aproximación aumenta el error de tipo I considerablemente

Estrategia 1: End-point analysis

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
time1	1	3.951	3.9512	0.3355	0.6031
group	1	9.549	9.5488	0.8107	0.4343
Residuals	3	35.333	11.7778		

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.933e+01	5.548e+00	5.287	0.0132 *
time1	6.713e-17	2.253e-01	0.000	1.0000
groupB	3.000e+00	3.332e+00	0.900	0.4343



Diferencias medias entre ambos grupos ajustado
por las diferencias basales

Estrategia 1: End-point analysis

- En promedio, ¿hay diferencias significativas en cada tiempo?
 - No se puede decir nada
- En promedio, ¿hay cambios significativos desde el inicio (baseline)?
 - No se puede decir nada
- ¿Difieren los grupos en algún tiempo?
 - No hay diferencias a tiempo 4
- ¿Difieren los dos grupos en la respuesta a lo largo del tiempo?
 - No se puede decir nada

Estrategia 2: ANOVA univariante con medidas repetidas

Es como un ANOVA, pero teniendo en cuenta las diferencias entre sujetos

Aproximación Naive: ANOVA

- ANOVA con los datos en formato 'largo', ignorando la correlación intra-sujetos
- Compara medias a cada tiempo como si fueran muestras independientes (análogo a usar two-sample t test con un test apareado).
- Estrategia 'under-powered'

Aproximación Naive: ANOVA

grp	time1	time2	time3	time4	<u>MEAN</u>
A	31	29	15	26	24.75
A	24	28	20	32	
A	14	20	28	30	
MEAN:	23.00	25.67	21.00	19.33	
B	38	34	30	34	29.33
B	25	29	25	29	
B	30	28	16	34	
MEAN:	31.00	30.33	23.67	32.33	

Overall mean=27

Within time

Between groups


Aproximación Naive: ANOVA

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
time	3	224.8	74.93		2.291	0.1173
group	1	126.0	126.04		3.854	0.0673
time:group	3	26.8	8.93		0.273	0.8439
Residuals	16	523.3	32.71			

Estrategia 2: ANOVA univariante con medidas repetidas

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
time	3	224.8	74.93	2.567	0.0932	.
group	1	182.0	182.03	6.237	0.0246	*
time:group	3	26.8	8.93	0.306	0.8207	
Residuals	15	437.8	29.18			



No diferencias aparentes en la respuesta a lo largo del tiempo entre grupos

Estrategia 2: ANOVA univariante con medidas repetidas

- En promedio, ¿hay diferencias significativas en cada tiempo?
 - No, 'time' no estadísticamente sig ($p=.0932$)
- En promedio, ¿hay cambios significativos desde el inicio (baseline)?
 - No, 'time' no estadísticamente significativo
- ¿Difieren los grupos en algún tiempo?
 - Si, hay algún punto en los que difieren ($p=0.0246$)
- ¿Difieren los dos grupos en la respuesta a lo largo del tiempo?
 - No 'group:time' no estadísticamente significativo ($p=0.8207$)

Estrategia 3: MANOVA

- Análisis Multivariante: Más de una variable para una variable dependiente
- Aproximación multivariante para medidas repetidas – Trata la respuesta como un vector multivariante
- Puede ser aplicado en otras situaciones con múltiples variables dependientes

Estrategia 3: MANOVA

- Recordamos t-test

$$\bar{Y}_{diff} = \frac{\sum_{i=1}^n y_2 - y_1}{n}$$

$$\frac{\bar{Y}_{diff}}{SD(\bar{Y}_{diff})} \sim T_{n-1}$$

t de Student

- MANOVA e un t-test apareado donde la variable respuesta es un vector de diferencias en vez de una diferencia simple (T número de medidas repetidas)

$$F = \left(\frac{N-T+1}{(N-1)(T-1)} \right) H^2$$

$$H^2 = \frac{N \mathbf{y}_{diff}^T \mathbf{y}_{diff}}{\mathbf{S}_{diff}^2}$$

T de Hotelling

Estrategia 3: MANOVA

Diferencias T1

id	group	diff1	diff2	diff3
1	A	-2	-14	11
2	A	4	-8	12
1	A	6	8	2
2	B	-4	-4	4
3	B	4	-4	4
6	B	-2	-12	18

Nota: considera que todas las diferencias son iguales (pesos), por eso es difícil de interpretar si los intervalos de tiempo no están equi-espaciados

Note: asume que las diferencias siguen una distribución normal multivariante + homogeneidad de varianzas

Estrategia 3: MANOVA

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
group	1	0.49226	0.24238	4	1	0.8879
Residuals	4					

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
group	1	0.50774	0.24238	4	1	0.8879
Residuals	4					

	Df	Hotelling-Lawley	approx F	num Df	den Df	Pr(>F)
group	1	0.96951	0.24238	4	1	0.8879
Residuals	4					

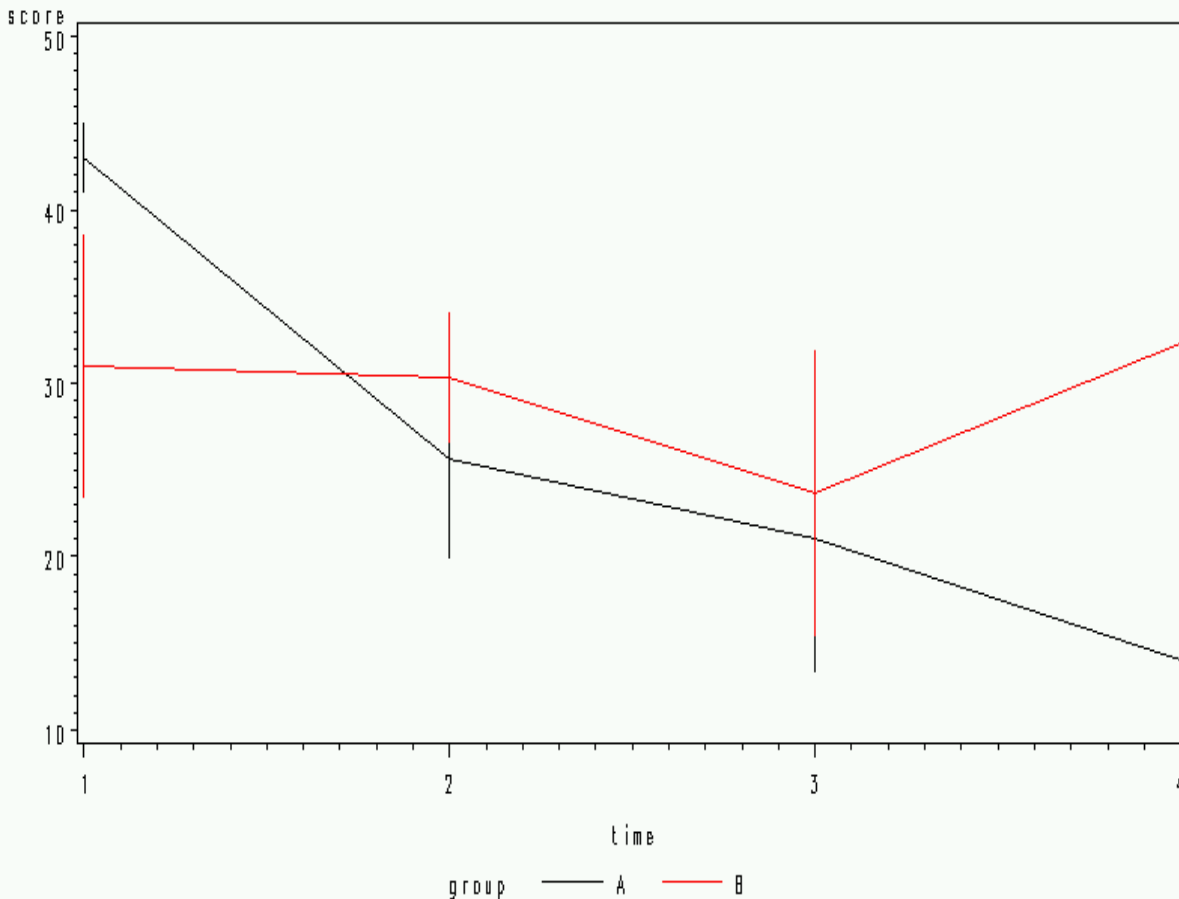
Estrategia 3: MANOVA

- En promedio, ¿hay diferencias significativas en cada tiempo?
 - No se puede decir nada
- En promedio, ¿hay cambios significativos desde el inicio (baseline)?
 - No se puede decir nada
- ¿Difieren los grupos en algún tiempo?
 - No se puede decir nada
- ¿Difieren los dos grupos en la respuesta a lo largo del tiempo?
 - No 'group' no estadísticamente significativo ($p=0.8879$)

Conclusiones

- Si la hipótesis de 'compound symmetry' se cumple, tenemos más poder con el análisis univariado que con el multivariado (más grados de libertad)
- Pero si no se cumple, estaremos aumentando el error de tipo I -> usar MANOVA
- Tenemos que imputar datos
- Los datos tienen que estar medidos en el mismo tiempo
- Simple de implementar e interpretar
- Siguiendo clase ... modelos mixtos (solucionan algunos de estos problemas)

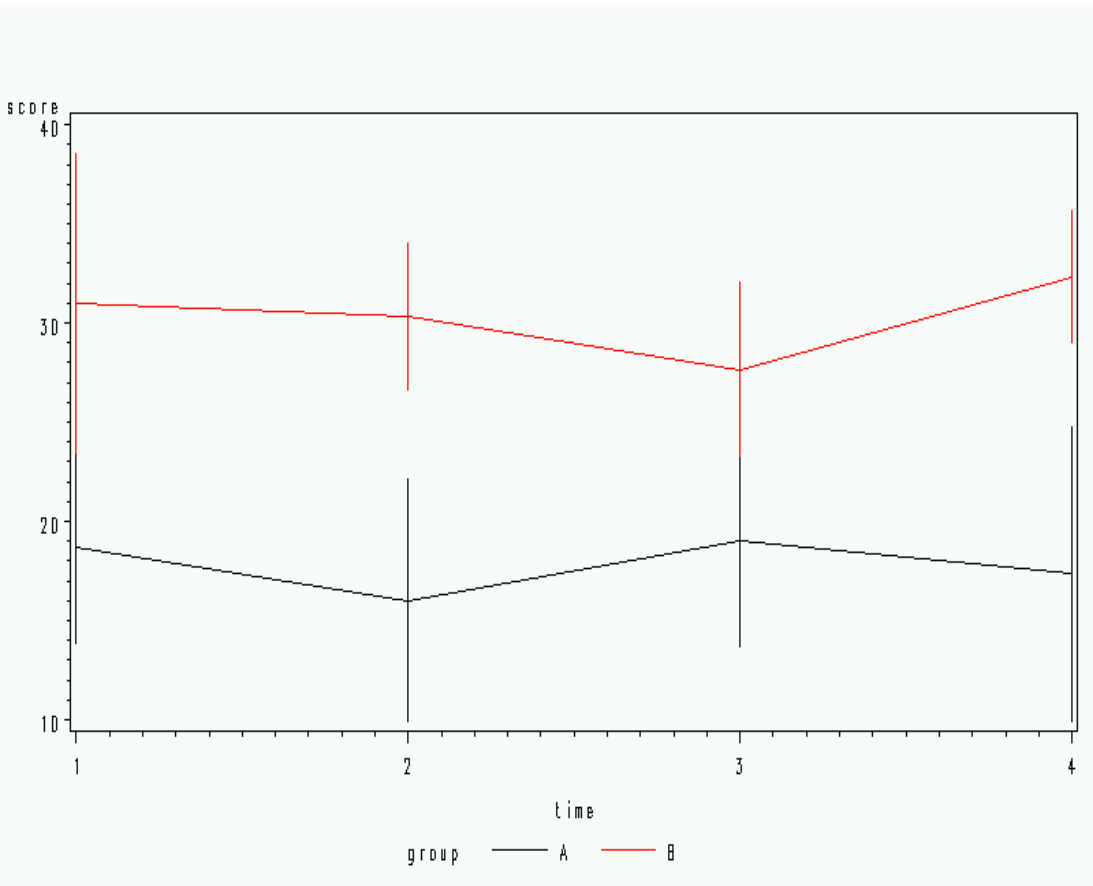
Ejercicios



¿qué efecto
esperarías encontrar
estadísticamente
significativos?

- tiempo
- grupo
- tiempo*grupo

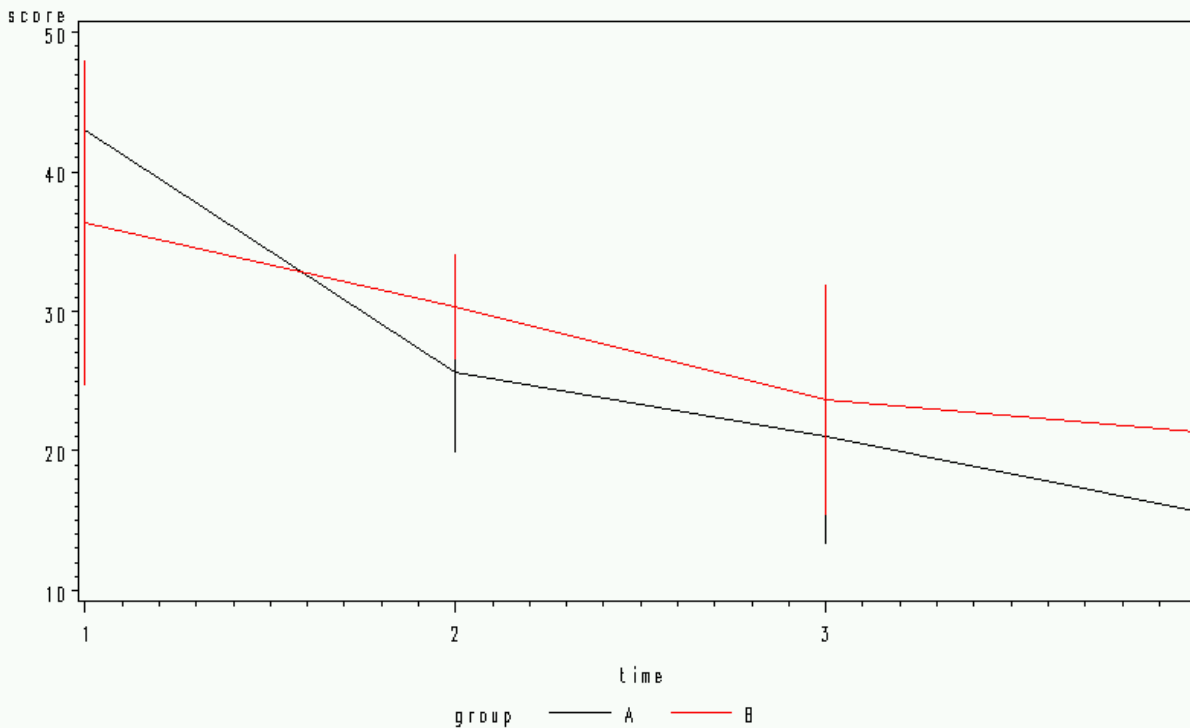
Ejercicios



¿qué efecto
esperarías encontrar
estadísticamente
significativos?

- tiempo
- grupo
- tiempo*grupo

Ejercicios



¿qué efecto
esperarías encontrar
estadísticamente
significativos?

- tiempo
- grupo
- tiempo*grupo