

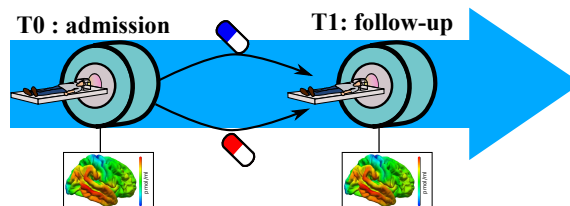
Exercises day 8

Basic Statistics for health researchers

13 May 2024

Exercise A: what to adjust on?

In the lecture it was mentioned that using the change between baseline and follow-up provides a natural adjustment for certain but not all covariates (we assume that all covariates have a linear effect). Consider the following study:



The study aims at assessing the impact of an antidepressive treatment (SSRI) on the brain serotonergic system. Patients were recruited, underwent baseline measurements, and were either given placebo or SSRI. A follow-up measurement was performed a week later. At each timepoint, a PET scan is performed to quantify the availability of serotonin receptors in the brain, which involves the injection of a radioactive contrast agent to the patient. A difference in change in PET signal between the two groups would be indicative of a treatment effect. However other factors may influence the PET signal:

- genetic polymorphisms (e.g. 5-HTTLPR)
- age (decline of 10% per decade)
- scanner type (binary variable, only 2 scanner types)
- radioactive dose (scan and patient dependent)

1. Which factors are "naturally" adjusted for when computed the change score?
Denote these factors \mathbf{X} and the remaining factors \mathbf{Z} .
How would you test the treatment effect if there was no \mathbf{Z} -factors?
2. How would you control for the \mathbf{Z} -factors?
What would be the benefit(s) of this adjustment?
(consider the case of a randomized study and an observational study)
3. In randomized experiment, adjusting for post-randomization variables is generally not recommended. Why? Is that problematic in this example?

Exercise B: analyzing a longitudinal study

In this exercise, we will reproduce the graphics and results presented during the lecture¹. A few extra-analyses will also be suggested. The exercise is divided in 3 independent parts:

- Part 1: descriptive statistics
- Part 2: comparing the change using t-tests
- Part 3: comparing the change using a mixed model

We recommend that you spend approximately 30 min to prepare each part.

Handling repeated measurement require substantial data management and involve new R functions. To save time, this document and the R demo file contain most of the R code needed to perform the analysis. This should help focus on the concepts seen during the lecture and the interpretation of the software output. But that should not prevent you to ask questions about the code.

To load the data in  you should have installed the package nlme and run:

```
data(armd.wide, package = "nlmeU")
```

The following code converts the data from the wide to the long format:

```
armd.long <- reshape(armd.wide, direction = "long",  
                    varying = paste0("visual", c(0,4,12,24,52)),  
                    times = c(0,4,12,24,52),  
                    timevar = "week.num",  
                    v.names = "visual")  
  
armd.long$week <- as.factor(armd.long$week)
```

You will also need to load the following packages:

¹If you would like to practice on another dataset you can have a look to the vitamin study (`data(vitaminW)`, 10 animals, 6 timepoints) or to the abeta study (`data(abetaW)`, 131 individuals, 2 timepoints).

Part 1: descriptive statistics

In this first part we will replicate the descriptive statistics presented during the lecture (slides 14-18).

1. We can display the dataset in the wide format using `str`. What is the meaning of the values in the columns `treat.f` and `miss.pat`?

```
str(armd.wide)
```

```
'data.frame':      240 obs. of  10 variables:
 $ subject : Factor w/ 240 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ lesion  : int   3 1 4 2 1 3 1 3 2 1 ...
 $ line0   : int  12 13 8 13 14 12 13 8 12 10 ...
 $ visual0 : int  59 65 40 67 70 59 64 39 59 49 ...
 $ visual4 : int  55 70 40 64 NA 53 68 37 58 51 ...
 $ visual12: int  45 65 37 64 NA 52 74 43 49 71 ...
 $ visual24: int  NA 65 17 64 NA 53 72 37 54 71 ...
 $ visual52: int  NA 55 NA 68 NA 42 65 37 58 NA ...
 $ treat.f : Factor w/ 2 levels "Placebo","Active": 2 2 1 1 2 2 1 1 2 1 ...
 $ miss.pat: Factor w/ 9 levels "----", "---X",...: 4 1 2 1 9 1 1 1 1 2 ...
```

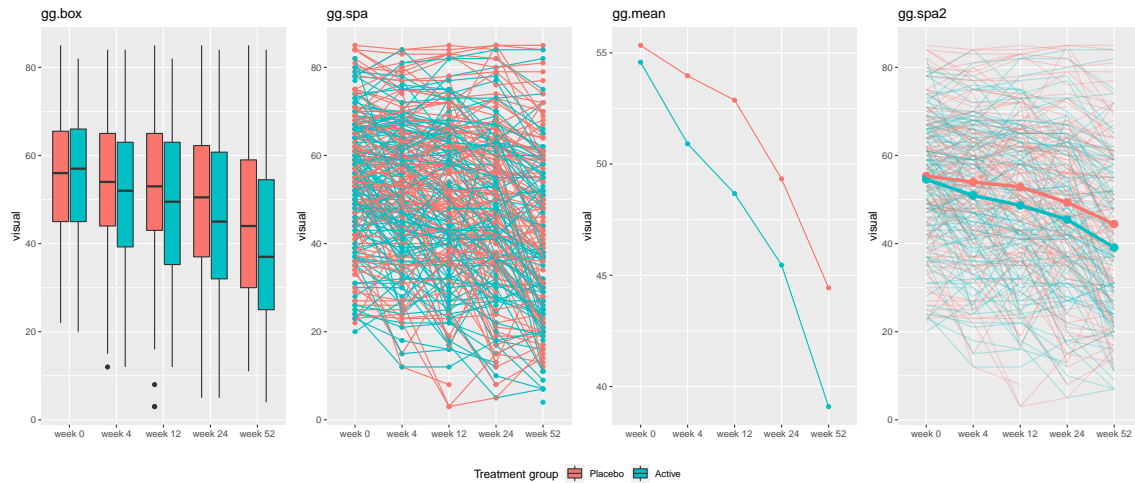
The `summarize` function can be used to compute summary statistics per group. Its first argument is a formula where the outcome is on the left hand side and the grouping variable(s) on the right-hand side, separated with `+`.

2. What information does the following software output provides?
How would you do proceed to compute the mean and variance per time, regardless to the treatment group?

```
armd.s <- summarize(visual ~ week + treat.f, na.rm = TRUE,
                    data = armd.long)
armd.s
```

	week	treat.f	observed	missing	mean	sd	min	q1	median	q3	max
1	0	Placebo	119	0	55.33613	15.00129	22	45.00	56.0	65.50	85
2	4		117	2	53.96581	15.90973	12	44.00	54.0	65.00	84
3	12		117	2	52.87179	17.20091	3	43.00	53.0	65.00	85
4	24		112	7	49.33036	18.51242	5	37.00	50.5	62.25	85
5	52		105	14	44.43810	18.53683	11	30.00	44.0	59.00	85
6	0	Active	121	0	54.57851	14.82270	20	45.00	57.0	66.00	82
7	4		114	7	50.91228	15.81114	12	39.25	52.0	63.00	84
8	12		110	11	48.67273	17.47665	12	35.25	49.5	63.00	82
9	24		102	19	45.46078	18.08050	5	32.00	45.0	60.75	84
10	52		90	31	39.10000	18.40069	4	25.00	37.0	54.50	84

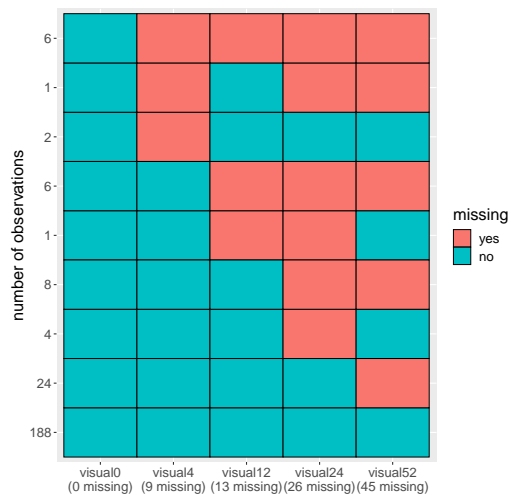
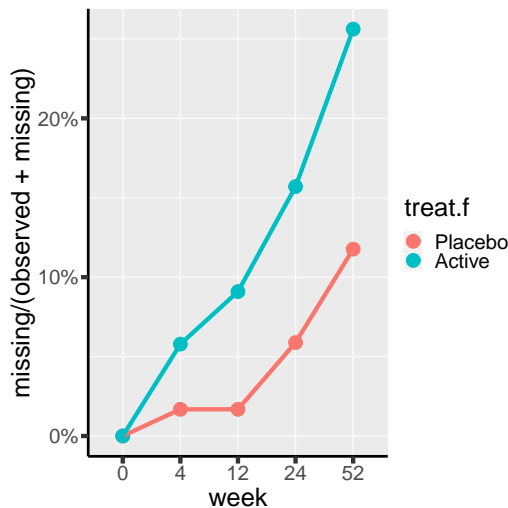
3. Discuss which of the following graphical representation (line 45-89 of the R demo file) you find the most useful to summarize the data? What information is missing?



4. What type of information is provided by the following figures?

```
## left panel
gg.NA <- ggplot(armd.s , aes(x = week, y = missing/(observed+missing),
  color = treat.f, group = treat.f))
gg.NA <- gg.NA + geom_point(size = 6) + geom_line(linewidth = 2)
gg.NA <- gg.NA + scale_y_continuous(labels = scales::percent)
gg.NA

## right panel
armd.visual <- armd.wide[,paste0("visual",c(0,4,12,24,52))]
plot(summarizeNA(armd.visual))
```



Part 2: Univariate approach

5. What are the following lines of code achieving?

```
test <- is.na(armd.wide$visual0)+is.na(armd.wide$visual52)
armd.wideCC <- armd.wide[test==0,]
armd.wideCC$change52 <- armd.wideCC$visual52 - armd.wideCC$visual0
```

Tip: use a subset of the data, e.g. `armd.wide2 <- armd.wide[c(1,2,5,50),]` to run the previous code and inspect each intermediate result.

6. Assess the treatment effect by comparing the change between the two groups using a t-test. Extract the estimated effect, its confidence interval, and p-value.

How does this analysis compares with the summary statistics computed in question 2?

7. Why do we get a (slightly) different p.value when using the `lm` function compared to the `t.test`?

```
e.lm <- lm(change ~ treat.f, data = armd.wideCC)
summary(e.lm)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11.180952	1.557168	-7.180312	1.466539e-11
treat.fActive	-4.296825	2.292089	-1.874633	6.235402e-02

8. [optional] Repeat this analysis considering another timepoint (e.g. 24 weeks). What are the limitations of this approach?

Part 3: Multivariate approach

To start with we restrict the analysis to the first and last endpoint:

```
armd.long52 <- armd.long[armd.long$week %in% c("0","52"),]  
armd.long52$week <- droplevels(armd.long52$week)
```

9. What is the interpretation of coefficients from the following mixed model (e052.lmm)? Can you deduce from the coefficients the estimated average vision at each timepoint?

Do you retrieve the estimated treatment effect by `lm` / t-test on the change?

```
dfCC <- armd.long52[armd.long52$subject %in% armd.wideCC$subject,]  
e052.lmm <- lmm(visual ~ treat.f*week,  
               repetition = ~week|subject,  
               data = dfCC)  
model.tables(e052.lmm)
```

	estimate	se	df	lower	upper	p.value
(Intercept)	55.619048	1.452203	193.0400	52.754826	58.4832695	0.000000e+00
treat.fActive	-1.041270	2.137585	193.0400	-5.257290	3.1747506	6.267228e-01
week52	-11.180952	1.557168	192.9844	-14.252206	-8.1096988	1.466849e-11
treat.fActive:week52	-4.296825	2.292089	192.9844	-8.817588	0.2239375	6.235414e-02

10. The same mixed model can be fitted on all patients still considering only 2 timepoints (e52.lmm) or on all patients and all timepoints (e.lmm). Which one of e052.lmm, e52.lmm, e.lmm provides the most reliable estimate of the treatment effect?

```
e52.lmm <- lmm(visual ~ treat.f*week,  
              repetition = treat.f~week|subject,  
              data = armd.long52)  
model.tables(e52.lmm)
```

	estimate	se	df	lower	upper	p.value
(Intercept)	55.3361345	1.375166	118.0246	52.612938	58.05933102	0.000000e+00
treat.fActive	-0.7576221	1.925328	237.8529	-4.550494	3.03524992	6.943006e-01
week52	-11.0843836	1.591884	106.4540	-14.240293	-7.92847431	2.853697e-10
treat.fActive:week52	-4.3935823	2.265183	195.5661	-8.860905	0.07374083	5.386507e-02

```
e.lmm <- lmm(visual ~ treat.f*week,
             repetition = ~week|subject,
             data = armd.long)
model.tables(e.lmm)
```

	estimate	se	df	lower	upper	p.value
(Intercept)	55.3361345	1.366936	238.0191	52.643297	58.02897213	0.000000e+00
treat.fActive	-0.7576221	1.925135	238.0200	-4.550100	3.03485623	6.942712e-01
week4	-1.2812792	0.764694	231.3334	-2.787934	0.22537572	9.517842e-02
week12	-2.3516584	1.091400	219.6983	-4.502611	-0.20070566	3.227167e-02
week24	-6.0200224	1.318454	212.4899	-8.618947	-3.42109743	8.414486e-06
week52	-11.3109451	1.598782	192.6856	-14.464305	-8.15758503	2.701706e-11
treat.fActive:week4	-2.2042232	1.087419	231.9888	-4.346702	-0.06174429	4.380391e-02
treat.fActive:week12	-3.5079396	1.560344	222.4007	-6.582891	-0.43298809	2.554512e-02
treat.fActive:week24	-3.0695747	1.895345	216.4638	-6.805269	0.66611980	1.067885e-01
treat.fActive:week52	-4.8662683	2.317422	198.7570	-9.436157	-0.29637910	3.700270e-02

11. [optional] Fit a mixed model including in the mean structure the categorical time variable and an interaction between the continuous time variable and the treatment variable.

What is the estimated treatment effect in this new model?

	estimate	se	df	lower	upper	p.value
(Intercept)	54.954	0.9608	239.0	53.0614	56.846944	0.000e+00
week4	-2.207	0.5520	242.6	-3.2939	-1.119199	8.506e-05
week12	-3.585	0.8193	258.5	-5.1982	-1.971577	1.758e-05
week24	-6.563	1.0585	279.3	-8.6469	-4.479695	2.016e-09
week52	-11.601	1.5316	203.3	-14.6206	-8.580713	1.249e-12
week.num:treat.fActive	-0.083	0.0409	187.4	-0.1637	-0.002311	4.385e-02