Seminar 1

R Seminar 1: Discussion of the analysis

The first part of the seminar will be devoted to discussing the analysis of the Anxiety data set carried out using R. The students should carefully think about the points emphasized below and being proactive during the discussion. The students should write a report discussing the results of the analysis in detail and provide a printed and electronic version (via email) of the report and the R code (one per tutorial group).

Problem: The researchers want to study the levels of social anxiety across four different groups:

1. Healthy controls;
2. Social anxiety without depression;
3. Social anxiety with depression and
4. Depression

Points for discussion

1. Based on a graphical exploration of the data
   - What are your preliminary conclusions regarding the level of social anxiety in each group?
   - What can you say regarding the variability of social anxiety across different groups?
   - Are there any outliers?
2. Based on the mean and the standard deviation of social anxiety in each group
   - What preliminary conclusions can you draw from these values?
   - Do these results qualitatively agree with the ones you got from the graphical exploration of the data?
3. Which model was used for the analysis ?

4. Which hypotheses are tested with the model?

5. Based on the p-value associated with the hypotheses considered in (4)

   - What are your conclusions?

   - Are the groups equal?

6. Which groups were actually different?

7. What are your final conclusions regarding the relationship between Social anxiety and group?

8. Discuss the adequacy of the model based on model checking techniques. Discuss the significance of any model assumption violation. If necessary carry out an alternative analysis to confirm your results.

Additional exercises

The students **should write** a report with the answers of the following questions before the seminar (one per tutorial group). The report **should be handed in during the seminar**. Justify your answers!

1. If an F statistic for a one-way ANOVA table is significant, then

   a. the effect being tested is significant.

   b. the means being tested are not all equal.

   c. the associated p-value is smaller than α.

   d. all of above are true.

2. Which of the following statements is correct based in the following output?

   ANOVA table for $\mathrm{H}_0 : \mu_1 = \ldots = \mu_r$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	1	0.0625	0.06250	0.00	0.9578
Error	14	302.0675	21.57625		
Total	15	302.1300			

Levene tests: $\mathrm{H}_0 : \sigma_1^2 = \ldots = \sigma_r^2$ versus $\mathrm{H}_A : \sigma_i^2 \neq \sigma_j^2$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groups	1	8.3377	8.3377	0.00690	0.9650
Residuals	14	16928.6	1209.2		

a. Levene's test is only 0.0069; therefore, the F test is invalid.

b. Levene's test is ok and the F is 0.00 (less than α); therefore we conclude that the means are not all the same.

c. we can assume that the variances are equal and it appears that the means are also equal.

d. neither the variances nor the means are equal since both p-values are large.

e. the variances are more likely equal since Levene's test p-value is larger than F p-value.

Consider the following ANOVA output table for the next four items.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	?	?	?	?
Error	?	9.205	?		
Total	23	51.380			

3. In the above table, how many treatments or groups k are there?

a. 3

b. 4

c. 1

d. 2

e. none of the above

4. In the above table, what is N, where $N = n_1 + \ldots + n_k$?

   a. 13

   b. 20

   c. 23

   d. 24

   e. none of the above

5. In the above table, what is F Value?

   a. 30.545

   b. 42.175

   c. 14.058

   d. 0.46025

   e. none of the above

6. If for the previous table $\alpha = 0.05$ and we were to test the hypothesis:

$$\mathrm{H}_0 : \mu_1 = \ldots = \mu_k \quad \text{versus} \quad \mathrm{H}_1 : \text{at least two means differ,}$$

   then we would have to look in a table to find F_{α, df_1, df_2} and compare the F value with it. What would df_1 and df_2 be for this problem?

   a. $df_1 = 3$ and $df_2 = 21$

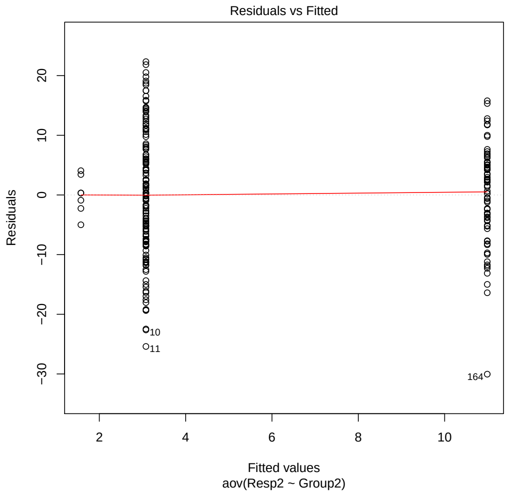
   b. $df_1 = 3$ and $df_2 = 20$

   c. $df_1 = 4$ and $df_2 = 21$

   d. $df_1 = 4$ and $df_2 = 20$

   e. none of the above

7. A manager carried out a study to evaluate the impact of three training systems on productivity (measured as the number of items produced per hour) in a company producing computer components. The data were analyzed with a one-way analysis of variance model and the following plot displays the standardized residuals for each of the three training groups.



From the previous residual plot, we can see that

a. the normality assumption seems to be violated.

b. the homogeneity of variance assumption seems to be violated.

c. the independence assumption seems to be violated.

d. of the assumption(s) this plot is used to detect, none seem to be violated.

e. none of above.

8. Which of the following are true?

a. the F statistics can never take a negative value.

b. a one-way ANOVA F-test tests whether the means are equal or not.

c. the F statistics always compares variability like given by the sum of squares.

d. all of the above are true.

9. We used an ANOVA model to compare the performance of three treatments. The sample sizes per treatment were 523, 125 and 127 respectively. When evaluating the assumptions of the model a qq-plot graph gives evidence of certain lack of normality. Further we want to check the homoscedasticity assumption as well. Which of the following statements is true?

a. the normality assumption was violated so we should not check homoscedasticity because one assumption was already violated.

b. our sample sizes are large enough so our model is robust against departures from the homoscedasticity assumption and we do not need to test for that.

c. we should check homoscedasticity using the Levene test.

10. A researcher wants to study people's intentions to get a flu-vaccine shot in an area threatened by an epidemic. To that end 90 persons were classified into three groups according to their risk of getting sick and were asked to quantify their likelihood of getting the flu-vaccine shot on a probability scale ranging from 0 to 1. All persons were together in the same room when they were asked about their likelihood of getting the shots. Therefore, it is possible that some persons overheard the answers of nearby respondents. The researcher wishes to test whether the means of the likelihood of getting the shots are the same for the three risk groups. Are the ANOVA model assumptions likely to hold in the present situation? Justify your answer.

11. A study was designed to investigate the efficacy of four different treatments, A, B, C and D, conceived to control the main symptoms of schizophrenia. In total 400 patients were randomly assigned to each of the treatments, each group receiving the same number of patients. After 8 weeks of treatment the status of the patients was evaluated using the Positive and Negative Syndrome Scale (PANSS).

    a. write down the model you would use to analyze these data.

    b. write down the hypothesis of interest for this problem.

    c. write down the assumptions of your model and discuss their plausibility in this example.

12. In the previous study, after collecting the data, the investigators observe that the biggest difference appears to be between treatments D and B. They want to know if this difference is statistically significant. What should they do in this case?

    a. they should apply a t-test to compare both groups and use 5% as the level of significance.

    b. this is an effect suggested by the data and they should never use a test to confirm it.

    c. they should apply first an ANOVA F test to compare the four groups and if this test is significant they should apply a 5% significant t-test to compare D and B.

    d. they should apply a Bonferroni procedure to account for multiple comparisons with $k = 2$.

    e. they should apply Tukey procedure to compare the treatments.

    Hint: Take into account that the hypothesis and its test should not be based on data inspection. So if you would like to test the difference between D and B, it should be part of explorative tests on all pairwise differences and you should somehow correct for multiple testing.