Exercises day 4

Basic Statistics for health researchers 2023

November 1, 2023

Warming up

Before starting the exercise below, learn from the R-demo of Lecture 4 (available from the course webpage):

1. Read and run the code.
2. Check that the output matches the results presented on the slides.
3. Do not hesitate to add your own comments into the script.

Exercise A (one-way ANOVA)

This exercise is very closely related to **Question 3** of **Exercise A** of **course day 2**. It is based on the same dataset: the biometrics dose response data (`biom` available from the `DoseFinding` package of R). For solving the previous exercise, we have used the Bonferroni correction, to correct for multiple testing when comparing multiple means. But we have now discussed a more powerful approach, the "min-P" method, than we can now use.

Note: in this exercise, we assume that subject-matter knowledge does not support the assumption of homogeneity of variance. Hence, we will use a statistical method which does not make the assumption of equal variance in all groups (as with the Welch's t-test).

Question 1

Use the "min-P" method to compare the average dose response between dose=0 and any other dose, while controlling the risk of making at least one false positive finding.

1. Compute the p-values and confidence limits for each comparison, while aiming to control the risk of making at least one false positive finding.
2. Compare the new results to those obtained with Bonferroni, in course day 2.

Question 2 (Additional)

We suggest to skip this question and move directly to exercise B. Come back to this "additional" question after exercise B, if time allows.

We now assume that the aim of the study was to compare all doses against each other.

1. Compute the p-values and confidence limits for each comparison, while aiming to control the risk of making at least one false positive finding.
2. Compare these new results to those obtained at Question 1. What are the difference for the comparisons between dose=0 and any other dose? Does it make sense? Why?
3. For comparison and pedagogical purpose:
    - compute the unadjusted p-value for the comparison of doses 0 and 0.6 (Hint: use the `t.test()` function).
    - Compute the corresponding adjusted p-value using the Bonferroni correction.
    - In this example, would you say that it matters to use the "modern and efficient" min-P approach instead of the simpler "good old" Bonferroni correction?

Exercise B (one-way and two-way ANOVA)

For this exercise we work once more with the Sickle Cell Disease (SCD) data.

Part 1 (one-way ANOVA)

In this first part, we assume that the aim is to study whether there is an association between the Mean Arterial Pressure (MAP) and the Body Mass Index (BMI) categorized into "underweight", "normal", "overweight" and "obese".

Question 1 (same as in exercises day 1)

1. Load the SCD data into R and visualize a summary of the data.
2. Create (and add to the data) these additional variables:
    - `BMI`: Body Mass Index (BMI)
    - `BMIgroup`: categorical version of the BMI, where groups are defined as suggested by the World Health Organization (BMI below 18.5 corresponds to "underweight", "normal" if between 18.5 and 25, "overweight" if between 25 and 30 and "obese" if above 30).
    - `MAP`: Mean Arterial Pressure, defined as the sum of the diastolic pressure (`Pdias`) plus one third of the difference between the systolic pressure (`Psys`) and the diastolic pressure (`Pdias`).

Question 2

We assume that subject matter knowledge supports the assumption of equal variances of MAP in each BMI group (or at least, does not really contradict it).

1. Fit an ANOVA model accordingly and print the summary of the model fit.
2. What is the estimated mean MAP in each BMI group?
3. What is the estimated standard deviation of the "error term" (aka "residual")? What is its interpretation in terms of how spread are typical MAP values around the means?
4. Compute the (sample) mean of MAP in each BMI group and compare it to the estimates from the model. Does it seem to make sense?
5. Compute the (sample) standard deviation of MAP in each BMI group and compare it to the estimate from the model. Does it seem to make sense?

Question 3

1. Produce appropriate plots to check the main assumption(s) of the ANOVA model. Hint: you can e.g. produce a residual plot and its corresponding "Wally plot".
2. What do you conclude? Do the main assumptions seem reasonable?

Question 4

1. Use a F-test and conclude whether the data suggest an association between BMI and MAP.
2. What could have made us prefer a F-test rather than the min-P method here?

Question 5

Some people sometimes recommend to use BMI as a continuous variable instead of categorizing it, for the statistical analysis. Some go as far as saying that this is always a "better and more powerful" approach. What disadvantages do you think that this approach could unfortunately have?

Part 2 (two-way ANOVA)

In this second part, we assume that the aim is to study whether there is an association between the Mean Arterial Pressure (MAP) and sex. We further assume that subject matter knowledge supports the following assumptions:

- equal variances of MAP in all groups defined from BMI and sex (or at least, does not really contradict it).
- if an association between sex and MAP exists, it should not depend on BMI.
- an association between BMI and MAP exists and it is similar for men and women.

Question 6

Transform the sex variable into a factor variable.

Question 7

1. Fit an appropriate two-way ANOVA model and print the summary of the model fit.
2. What are the estimated mean differences in MAP between men and women?
3. Why was it a good idea to transform the sex variable into a factor variable before fitting the ANOVA model?
4. What is the estimated standard deviation of the "error term" (aka "residual")? What is its interpretation in terms of how spread typical MAP values are around the means?

Question 8

1. Create a new variable `group` that simultaneously indicates the sex and the BMI of the subject corresponding to each observation. Hint: you can use the `interaction()` function as follows

   ```
   d$group <- interaction(d$sex,d$BMIgroup) .
   ```

2. Produce boxplots summarizing the distribution of MAP in each of the groups.
3. Overall, do the boxplots match with the main results of the ANOVA model fit? Why?
4. Overall, do the boxplots seem to support or contradict the main modeling assumptions?

Question 9

1. Produce appropriate plots to check the main assumption(s) of the ANOVA model. Hint: you can e.g. produce a residual plot and its corresponding "Wally plot".
2. What do you conclude? Do the main assumption seem reasonable?

Question 10

1. According to the two-way ANOVA model, is there a significant difference between the mean MAP of men and women?
2. Are the results from the F-test and the min-P approach different? Why?

Question 11

1. Instead of using the two-way ANOVA model, use a simpler t-test to compare the mean MAP of men and women. Conclude.
2. What are the similarities and differences between these new results and those from the two-way ANOVA approach? Try to explain their origin.