Multivariate statistics: Assignment 1

Team B: **Kendall Brown** *stnumber* **Raïsa Carmen** *s0204278*
Stefan Velev *stnumber* **Adhithya Unni Narayanan** *stnumber*
Audrey-Justine Towo Kamga *stnumber*

Abstract

1 Introduction and data exploration

This report assesses hearing thresholds for a sample of 546 healthy male volunteers. The subjects were 52 years old on average at the start of the study and are followed for an average of 7.57 years. The hearing threshold is measured, on average, every 1.59 years. Table 1 describes the demographics in more detail. It can be seen that the data is highly unbalanced; there is a lot of variation in the time a volunteer is followed and in the number of times a volunteer visits. Normally, each ear is measured at each visit but this only happened in 93.22% of all visits. The left (right) ear was tested in 96.72% (96.5%) of all visits.

Figures 1 and 2 show the trends in the hearing threshold for all volunteers over time. These figures shows that many volunteers' hearing threshold over time has an erratic pattern meaning there is likely high variability **within subjects**. Additionally, variability between subjects is also high, especially for older volunteers.

Table 1: Demographics for all respondents

Age at the beginning of the study	
min	17.20
max	87.00
median	54.10
mean (sd)	51.99 $\pm$ 18.70
Years of follow-up	
min	0.00
max	22.40
median	6.30
mean (sd)	7.57 $\pm$ 6.30
Number of visits	
min	1
max	15
median	3.00
mean (sd)	4.19 $\pm$ 2.88

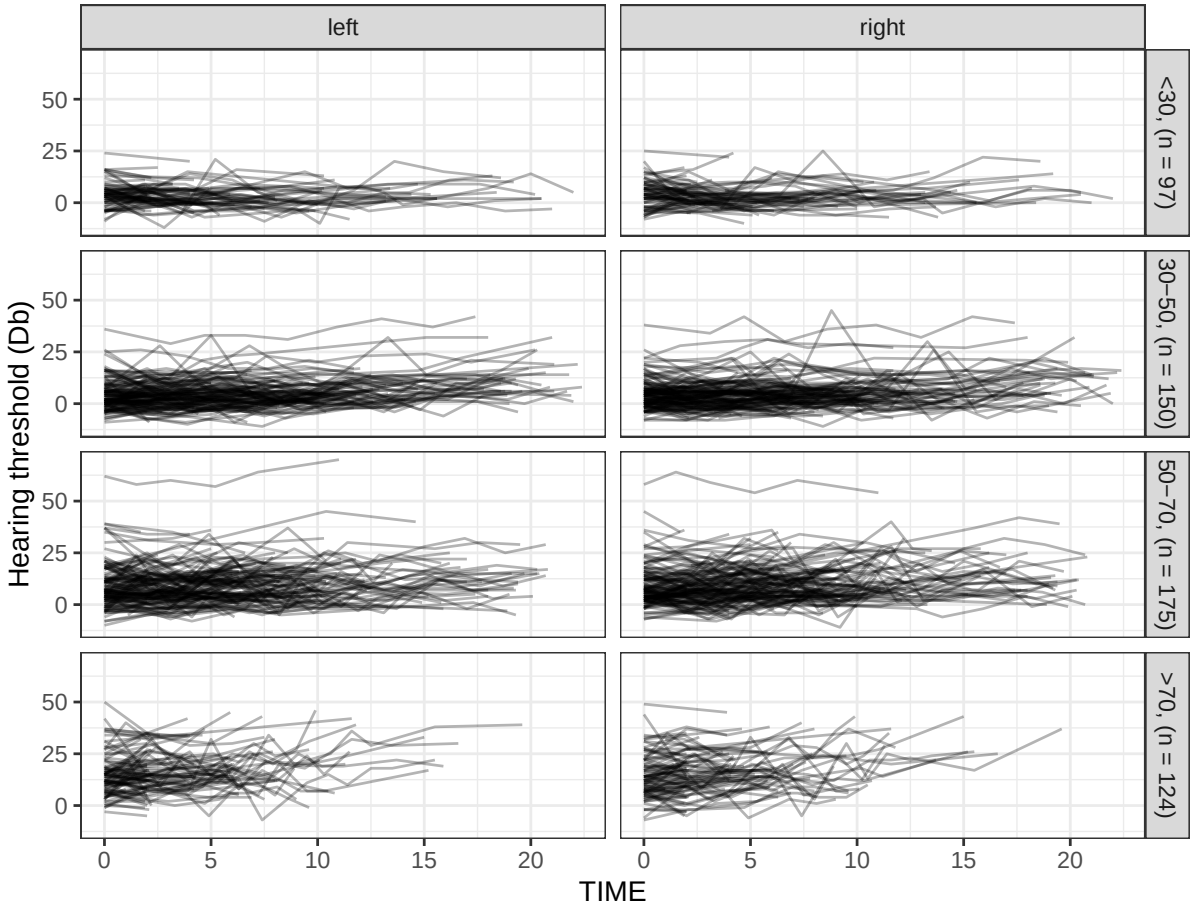

Figure 1: Hearing threshold over time, divided by left and right ear and by age group at the start of the study

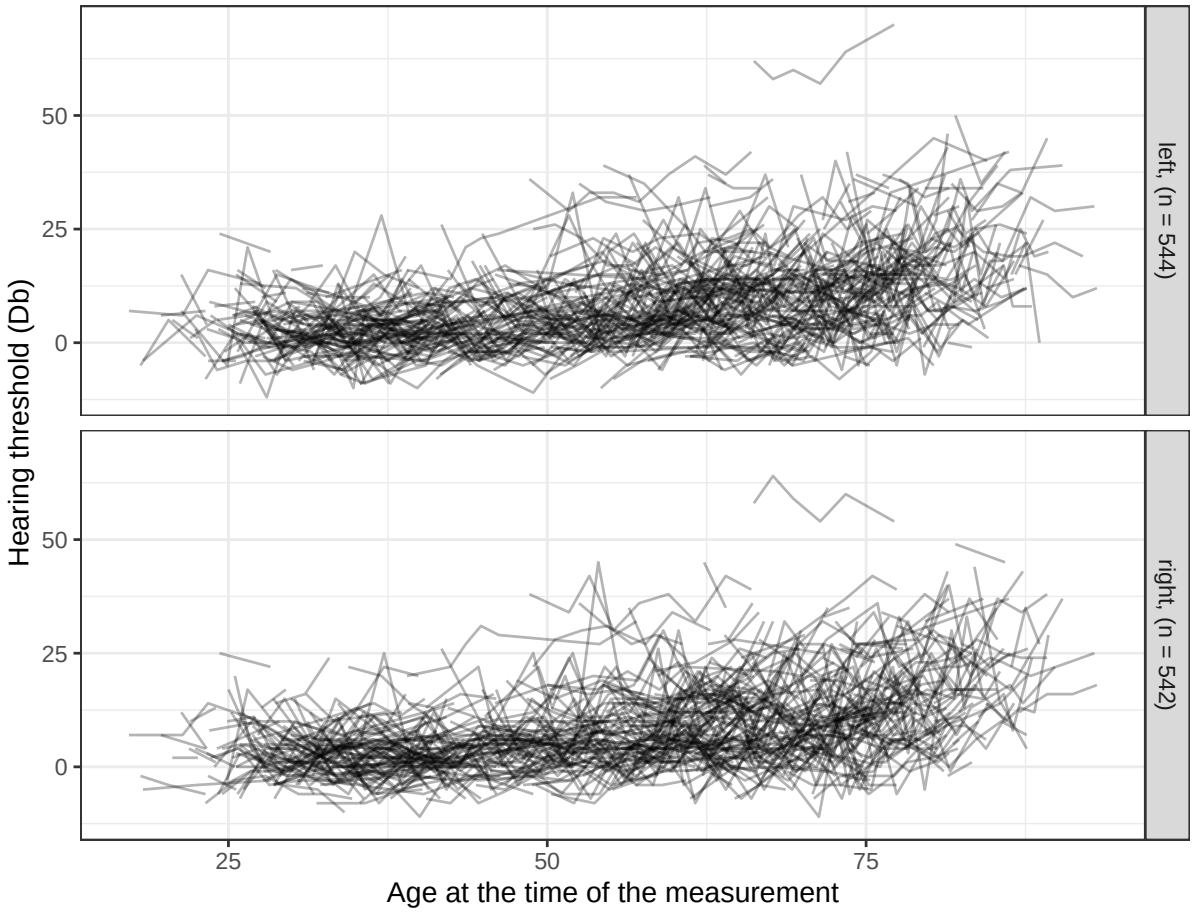

Figure 2: Hearing threshold over time, divided by left and right ear.

1.1 Mean, variance and correlation structure

Figure 3 shows the mean and 95% confidence interval for the hearing threshold (dB) for different age group. The age is the age at the time the measurement was taken. While the top graph shows 15 age groups with approximately equal sample size (between 141 and 149 measurements in each group), the bottom graph shows age groups with the same interval range of about 5 years. Both graphs show that the variance increases with age and there doesn't seem to be a significant difference, on average, between left and right ears.

It is obviuous that the data has a hierarchical structure where the highest hierarchical level is the level of the individual (id) and the second level is the side (left or right ear). Some basic descriptive statistics were derived using the *statsBy* function from the psych package. The intraclass correlation is 0.71 for the hearing threshold (y) and, as expected, very high (0.92) for age at the time of measurement. These high values reflect that the total variance in these variables that is associated with the groups is very high.

The total correlation between y and the age at the measurement, ignoring the hierarchical structure, is 0.45. Following Marzban et al. (2013) and Mongomery (2017), the correlation matrix for each subject that has at least 2 measurements for both ears is calculated, one can obtain within-group correlations and variances (where a group is a unique combination of subject and side). Figure 4 shows the histograms over all groups for the correlation between the age at the measurement and y, and the variances for both age and y (though the variance in age is not very informative). The mean or median of all within-group correlations is often used as a measure of within-group correlation. The mean (0.16) and median (0.32) of the within-group correlations are indicated in red and blue respectively on the graph. It can we seen that within-group correlation spans the entire range from -1 to 1, meaning that age has a strong positive relationship to the hearing threshold for some and a strong negative relationship for others. Age will be a good predictor for hearing threshold for some but not for others. Between-group correlation is obtained by averaging the age at measurement and the hearing threshold for each group (see scatterplot in Figure 4) and then computing the correlation across the groups, obtaining 0.52.

Describe the data, and use graphical techniques to explore the mean structure, the variance structure and the correlation structure. Summarize your conclusions. What are the implications with respect to statistical modeling ?

2 Methodology

In this section, we explore a couple of different methods to analyze the data. All analysis was carried out with the statistical software R. All scripts are freely available at this git repository.

2.1 Summary statistics

One possibility to deal with the hierarchical structure of the data is to summarize the data and reduce the number of measurements per subject to one.

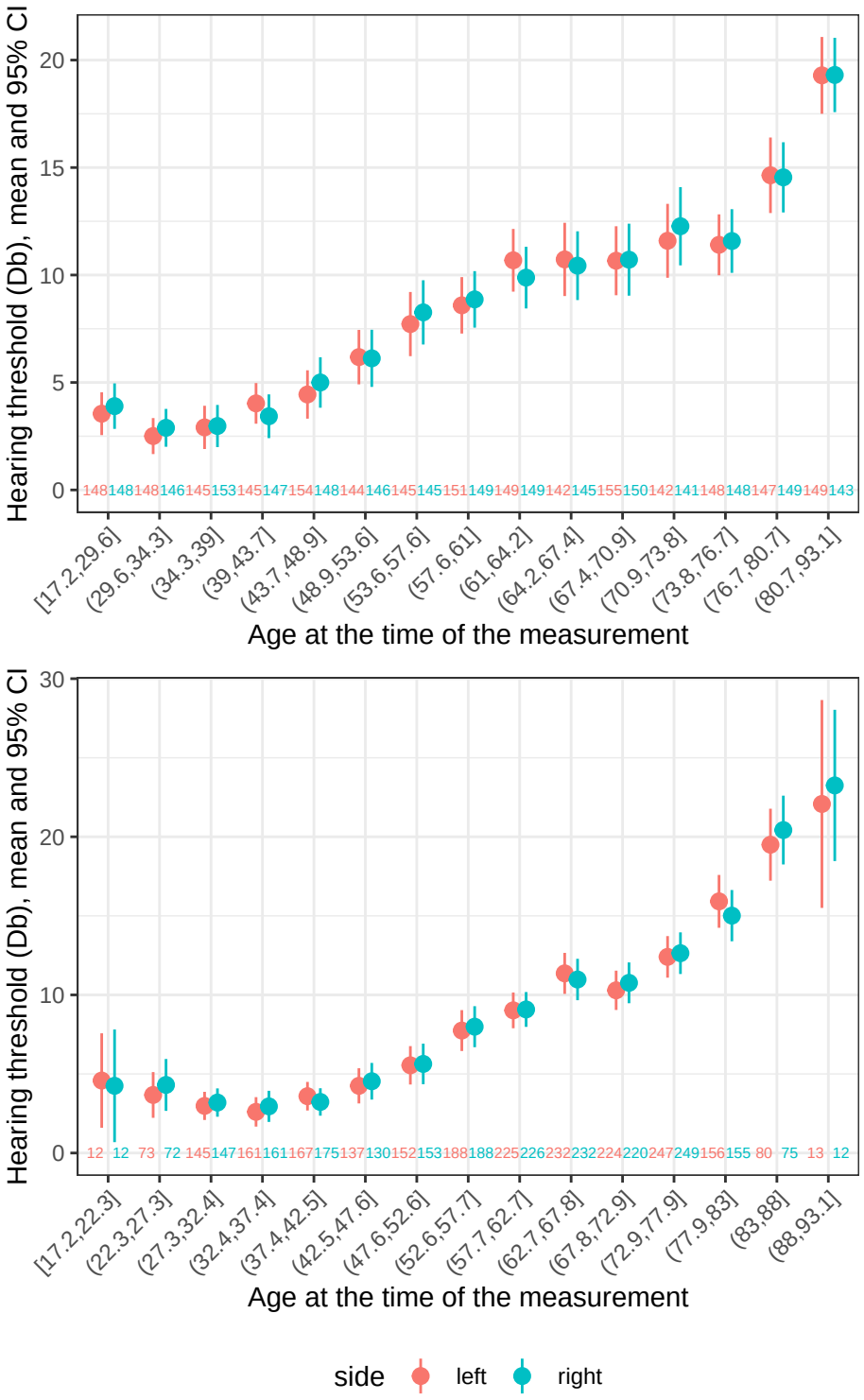

Figure 3: Hearing threshold over time, divided by left and right ear. numbers in the bottom show the number of measurements that were taken.

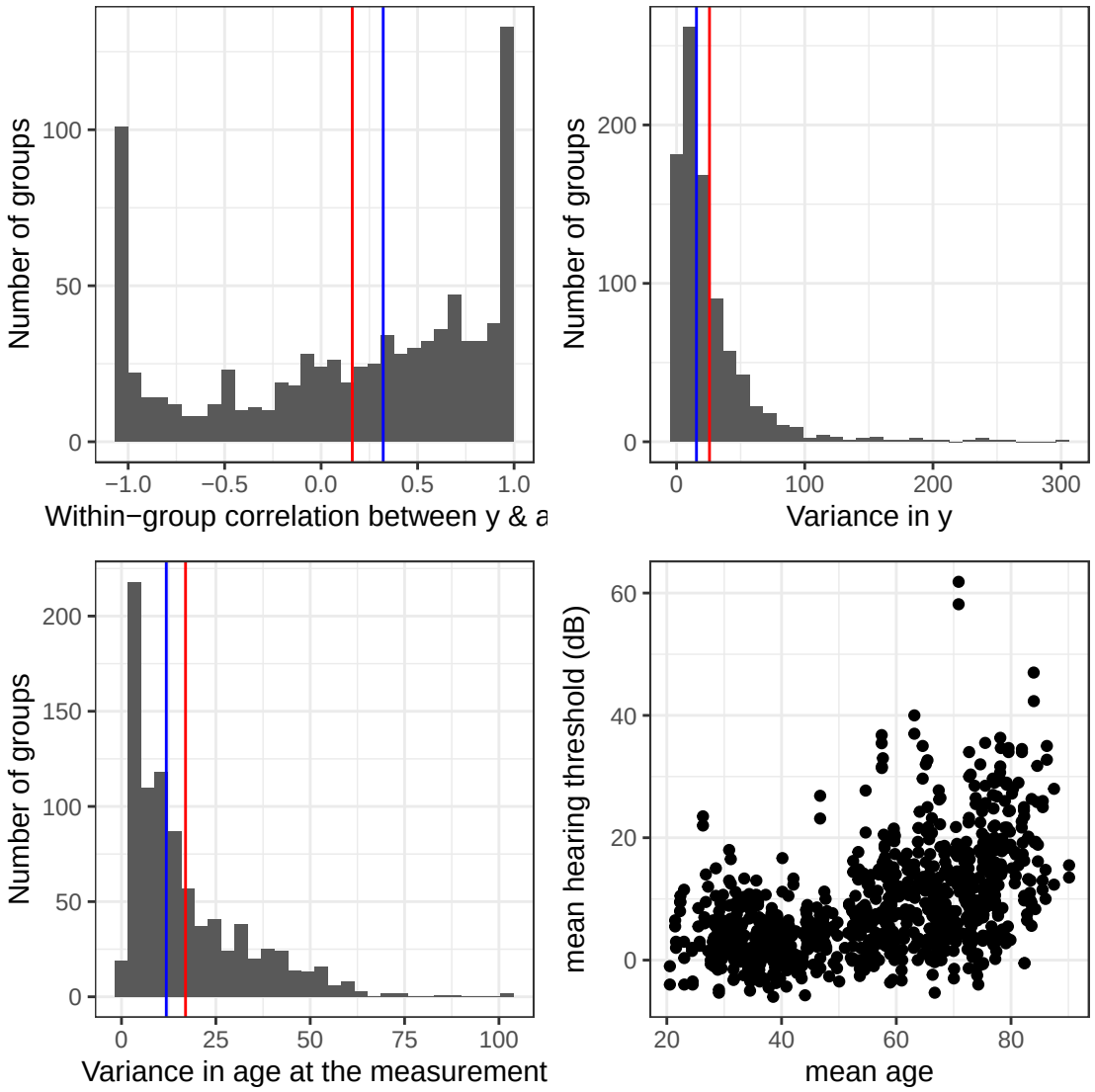

Figure 4: The mean is indicated in red, the median in blue

This can be done, for instance, by calculating the average change in the hearing threshold for each subject i, Δ_i as in equation (1) where Y_{ijk} is the k^{th} measurement for ear j of subject i and n_{ij} is the total number of measurements of ear j of subject i.

$$\Delta_i = \frac{(Y_{il1} - Y_{iln_{ij}}) + (Y_{ir1} - Y_{irn_{ij}})}{2} \tag{1}$$

An immediate problem with this method is that we have learned the the within-group variance in the hearing threshold is quite large. Focusing only on the first and last measurement is therefore risky. Additionally, if time would influence hearing threshold, we need to correct for the time between the first and last measurement since this differs a lot between subjects. This method also requires that each subjects' both ears have been measured at least twice.

An even simpler summary statistic would be to use the average hearing threshold over all measurements and both ears.

2.2 Multivariate model

2.3 Two-stage analysis

2.4 Random-effects model

3 Results

3.1 Summary statistics

3.2 Multivariate model

3.3 Two-stage analysis

3.4 Random-effects model

4 Discussion and conclusion

4.1 further research

Bibliography

Marzban, Caren, Paul R Illian, David Morison, and Pierre D Mourad. 2013. "Within-Group and Between-Group Correlation: Illustration on Non-Invasive Estimation of Intracranial Pressure." *Viewed Nd, from Http://Faculty. Washington. Edu/Marzban/Within_ Between_simple. Pdf*.

Mongomery, DC. 2017. "Design and Analysis of Experiments." *John Willy & Sons*.