Assignment2\_Part1\_VoiceInSchizophrenia

Signe Kløve Kjær and Thea Rolskov Sloth

Oct, 2018

Link to code: <https://github.com/YpipY/EXPMETH3/blob/master/Assignment%203%2C%20voice%20schizo/signe_thea_A3_P1_SchizophreniaVoice.Rmd>

## Assignment 2 - Part 1 - Assessing voice in schizophrenia

Schizophrenia has been associated with "inappropriate" voice, sometimes monotone, sometimes croaky. A few studies indicate that pitch might be an index of schizophrenia. However, an ongoing meta-analysis of the literature (which you will have a go at in the last assignment) indicates that pitch mean and standard deviation are only weak indicators of diagnosis. Can we do better with our new fancy complex skills?

The corpus you are asked to analyse is a set of voice recordings from people with schizophrenia (just after first diagnosis) and 1-1 matched controls (on gender, age, education). Each participant watched 10 videos of triangles moving across the screen and had to describe them (so you have circa 10 recordings per person). We have already extracted the pitch once every 10 milliseconds and you will have to use this data to assess differences in the voice.

Can you characterize voice in schizophrenia as acoustically different? Report the methods you used to answer this question and the results from the analyses. Add a couple of lines trying to interpret the results (make sense of the difference). E.g. People with schizophrenia tend to have high-pitched voice, and present bigger swings in their prosody than controls. Add a couple of lines describing limitations of the data/analyses if any is relevant.

N.B. There are looots of files to be dealt with. Maybe too many for your computer, depending on how you load the files. This is a challenge for you. Some (complementary) possible strategies: - You can select a subset of files only (and you have to justify your choice). - You can learn how to use the apply() or map() functions. - You can coordinate with classmates.

Hint: There is some information in the filenames that you might need. Hint: It might be a good idea to first create a function which loads and parses one file, and then loop through the list of files to read them all. For example

# 1. In the course of this assignment you have to first select one datafile and figure out how to:

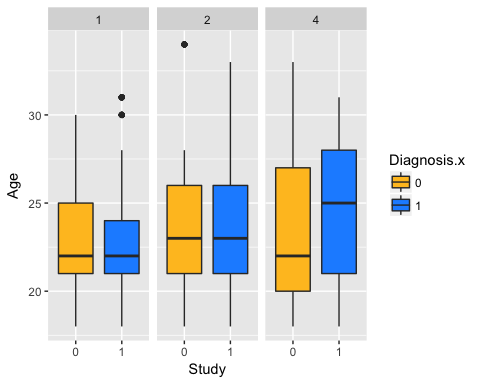
* Extract "standard" descriptors of pitch: Mean, standard deviation, range
* Extract less "standard" descriptors of pitch you can think of (e.g. median, iqr, mean absoluted deviation, coefficient of variation)

# 2. Second you will have to turn the code into a function and loop through all the files (or even better use apply/sapply/lapply)

* Remember to extract the relevant information from the file names (Participant, Diagnosis, Trial, Study)

#define pathway to pitch folder  
path = "/Users/thearolskovsloth/Studygroup/Assignment 2, part 1/cogsci-methods-3-assignments-master/Assignment3\_VoiceSchizo/Pitch/"  
  
#create function that extracts descriptors from pitch files  
read\_pitch <- function(filename) {  
 # read data  
 data <- read.delim(file = paste(path, filename, sep = ""), sep = "\t", header = TRUE)   
 # parse filename  
 meta = str\_extract\_all(filename, "\\d+")  
 study = as.numeric(meta[[1]][1])  
 Diagnosis = as.numeric(meta[[1]][2])  
 Subject = as.numeric(meta [[1]][3])  
 triangles = as.numeric(meta[[1]][4])  
   
 # extract descriptors  
 mean\_f0 = round(mean(data$f0), digits = 2)  
 sd = sd(data$f0)  
 quantiles = quantile(data$f0, probs = c(0.05, 0.25, 0.50, 0.75, 0.95))  
 q05 = quantiles[[1]]  
 q25 = quantiles[[2]]  
 median = quantiles[[3]]  
 q75 = quantiles[[4]]  
 q95 = quantiles[[5]]  
 iqr2575 = q75-q25  
 coe.var = coefficient.variation(sd(data$f0), mean(data$f0))  
 MAD = mad(data$f0) #average distance from something...  
   
 # combine all this data  
 output\_list = data.frame(Subject, Diagnosis, study, triangles, mean\_f0, sd, q05, q25, median, q75, q95, iqr2575, coe.var, MAD)  
   
 return(output\_list)  
}  
  
  
# test it on just one file while writing the function  
test\_data = read\_pitch("Study1D0S101T1\_f0.txt")  
  
# run through all pitch data files  
pitch\_data = list.files(path = path) %>%  
 purrr::map\_df(read\_pitch)  
  
#write a csv file with combined data  
write.csv(pitch\_data, file = "THEA\_pitch\_data.csv")

#lets plot the data!  
ggplot(all\_pitch, aes(x= Diagnosis.x, y = Age, fill = Diagnosis.x)) +  
 geom\_boxplot() +   
 facet\_wrap(~Study) +   
 scale\_fill\_manual(values=c("goldenrod1","dodgerblue")) +  
 labs(x= "Study", y = "Age")



# 3. Make one model per acoustic feature and test whether you can observe significant difference due to Diagnosis. Tip: Which other fixed factors should you control for (that is, include in the model)? Which random ones?

* Bonus points: cross-validate the model and report the betas and standard errors from all rounds to get an idea of how robust the estimates are.

# 3a. Is study a significant predictor in these models? What should you infer from this? Does study interact with diagnosis? What should you infer from this?

# 6. Write a paragraph reporting methods and results

We created a model for each acoustic feature. The standard model was:

These models showed that Diagnosis is a significant predictor for:

mean (β=13.36 (SE= 4.26), t = 3.14, p<.05)

q05 (β= 18.68 (SE= 4.55), t = 4.09, p<.05)

q25 (β= 16.40 (SE= 4.47), t = 3.67 , p<.05)

median (β= 11.29 (SE= 4.26), t = 2.65 , p<.05)

q75 (β= 10.98 (SE= 4.53), t = 2.42 , p<.05)

coe.var (β= -0.05 (SE= 0.01), t = -3.57 , p<.05)

MAD (β= -4.52 (SE= 1.28), t = -3.53 , p<.05)

Diagnosis was not a significant predictor for: sd, q95 and inter quantile range q75-q25.

The fixed effects was diagnosis as we wanted to test its axplanatory ability on pitch. Furthermore, Gender was included as a fixed effect. This was because we expect Gender to have a systematic influence on pitch, since males generally have a lower pitched voice than females.

The random effects was Subject as a random intercept, so we allow for subjects to have a different base level of pitch. Also, we included Triangles as a random intercept, to allow for the different videos of triangles to provoke different levels in the acoustic features.

To assess the different acoustic features we ran a 5 fold cross-validation of all the models. This was done in order to compare the predictive abilities of the different models with a root mean square error measure. The results of the cross validations were:

acoustic feature mean rmse value standardized: mean rmse value/sd in percent

|  |  |  |
| --- | --- | --- |
| acoustic feature | mean rmse value | standardized: mean rmse value/sd in percent |
| mean | 36.09 | 63.21 % |
| sd | 25.97 | 98.41 % |
| q05 | 38.17 | 72.86 % |
| q25 | 38.82 | 71.35 % |
| median | 33.92 | 63.86 % |
| q75 | 48.09 | 68.95 % |
| q95 | 70.31 | 81.49 % |
| iqr2575 | 45.05 | 93.39 % |
| coe.var | 0.13 | 107.92 % |
| MAD | 13.30 | 109.12 % |

The third column in the table is the calculated standardized measure of the models' predictive abilities. The calculations were done by taking the mean rmse value for the acoustic features and deviding them by the standard deviation of for that particular acoustic feature. This is done becasue we want to compare the error across different types of data with different units. Thereby, we get a standardized measure: the percentage from the standard deviation. Thus, the lower value, the better the model is at predicting.

If we look at the rmse value for the median\_model, then the rmse value is 33.92. In order to get the standardized measure we divide by the standard deviation of that acoustic feature (the variance around the mean). If the rmse value is lower than the sd value, we will get a standardized value below 1. If the sd value and the rmse value is the same, it would indicate that the error of the model is identical to the variance just present in the data. A good model should therefore have a lower error than the variance around the mean (sd).

From this we can evaluate that the error is smallest in the models predicting median and mean.

When testing if study is a significant predictor, we included study as a fixed effect in the models. Then the models looked like this:

None of these were significant.

Neither was the interaction of study:

The fact that Study is not a significant predictor is a good thing. If it had in fact been signficant would have indicated that there had been some systematic difference in how the experiment was conducted across the studies. If there had been an interaction of study and diagnosis, it would have indicated that the study had a different effect on controls and schizophrenia patients.

Maybe we should have checked assumptions?? Probably.........

## 