Introduction to R

Jaime Undurraga

31 March 2016

What is R?

- R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (much code written for S runs unaltered under R).

R Markdown

This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

why using R?

- It is free and open source (Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form)
- Time and processing efficient
- Many libraries (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc)
- Continuosly improved and developed by the R community
- Well documented (your answer will be usually found in stckoverflow)
- It makes statistical analysis reproducible to others
- It is a multiplatform system (FreeBSD, Linux, Windows and MacOS)
- well-designed publication-quality plots can be produced
- It is nice

R IDE

- R Studio is a nice Interface development enviorment to use R

Libraries

-R libraries can be installed and loaded

```
library(xlsx)
```

```
## Loading required package: rJava
```

```
## Loading required package: xlsxjars
```

Data Frames

- Data Frames contain and allow to manipulate data in many ways. Imported data from Excel, SPPS, cvc, txt and other formats can be loaded into R data frames

```
raw <- read.xlsx('/home/jundurraga/Dropbox/Documents/Macquarie/Introduction_to_R/pta_data.xlsx', 1, enc
head(raw, n=2)
```

```
##   subject Age_years Sex right_dB_HL_250_Hz right_dB_HL_500_Hz
## 1     1WR  48.11781   M                 20                 20
## 2     2PP  22.35616   F                 -5                 -5
##   right_dB_HL_1000_Hz right_dB_HL_2000_Hz right_dB_HL_4000_Hz
## 1                  20                  15                  10
## 2                 -10                  -5                  10
##   right_dB_HL_8000_Hz left_dB_HL_250_Hz left_dB_HL_500_Hz
## 1                  35                25                50
## 2                  10                -5                -5
##   left_dB_HL_1000_Hz left_dB_HL_2000_Hz left_dB_HL_4000_Hz
## 1                 25                 15                 20
## 2                 -5                 -5                 -5
##   left_dB_HL_8000_Hz right_ave left_ave left_right_ave better_ear_ave
## 1                 25        17       27             22             17
## 2                  5        -3       -5             -4             -5
##   left_right_ave_dB_HL_4000_Hz better_ear_dB_HL_4000_Hz
## 1                         15.0                        10
## 2                          2.5                        -5
##   left_right_ave_dB_HL_500_Hz better_ear_dB_HL_500_Hz
## 1                          35                      20
## 2                          -5                      -5
```

Descriptive statistics

```
summary(raw)
```

```
##     subject     Age_years     Sex    right_dB_HL_250_Hz right_dB_HL_500_Hz
##  10AA   : 1   Min.   :20.09   F:10   Min.   :-5.000     Min.   :-5.0
##  11RR   : 1   1st Qu.:24.45   M: 9   1st Qu.: 0.000     1st Qu.: 0.0
##  12SR   : 1   Median :35.19          Median : 0.000     Median : 5.0
##  13BL   : 1   Mean   :34.43          Mean   : 4.211     Mean   : 5.0
##  14PT   : 1   3rd Qu.:39.67          3rd Qu.: 5.000     3rd Qu.: 7.5
##  15SE   : 1   Max.   :54.00          Max.   :30.000     Max.   :25.0
##  (Other):13
##  right_dB_HL_1000_Hz right_dB_HL_2000_Hz right_dB_HL_4000_Hz
##  Min.   :-10.000     Min.   :-5.000      Min.   :-5.000
##  1st Qu.:  0.000     1st Qu.: 0.000      1st Qu.: 0.000
##  Median :  5.000     Median : 0.000      Median : 5.000
##  Mean   :  6.053     Mean   : 3.158      Mean   : 6.316
##  3rd Qu.: 10.000     3rd Qu.: 5.000      3rd Qu.:10.000
##  Max.   : 20.000     Max.   :15.000      Max.   :30.000
##
##  right_dB_HL_8000_Hz left_dB_HL_250_Hz left_dB_HL_500_Hz
##  Min.   :-5.00       Min.   :-5.000    Min.   :-10.000
##  1st Qu.: 2.50       1st Qu.:-5.000    1st Qu.:  0.000
##  Median :10.00       Median : 0.000    Median :  0.000
##  Mean   :10.53       Mean   : 2.105    Mean   :  3.421
##  3rd Qu.:15.00       3rd Qu.: 5.000    3rd Qu.:  5.000
##  Max.   :40.00       Max.   :25.000    Max.   : 50.000
##
```

```
##  left_dB_HL_1000_Hz left_dB_HL_2000_Hz left_dB_HL_4000_Hz
##  Min.   :-5.000     Min.   :-5.000     Min.   :-10.000
##  1st Qu.: 0.000     1st Qu.: 0.000     1st Qu.:  0.000
##  Median : 0.000     Median : 0.000     Median :  5.000
##  Mean   : 2.632     Mean   : 2.895     Mean   :  6.842
##  3rd Qu.: 5.000     3rd Qu.: 5.000     3rd Qu.: 12.500
##  Max.   :25.000     Max.   :15.000     Max.   : 30.000
##
##  left_dB_HL_8000_Hz   right_ave         left_ave      left_right_ave
##  Min.   :-10.0      Min.   :-3.000   Min.   :-5.000   Min.   :-4.000
##  1st Qu.:  2.5      1st Qu.: 1.500   1st Qu.: 0.500   1st Qu.: 1.500
##  Median : 10.0      Median : 3.000   Median : 2.000   Median : 3.000
##  Mean   : 10.0      Mean   : 4.947   Mean   : 3.579   Mean   : 4.263
##  3rd Qu.: 15.0      3rd Qu.: 9.000   3rd Qu.: 5.000   3rd Qu.: 5.750
##  Max.   : 40.0      Max.   :17.000   Max.   :27.000   Max.   :22.000
##
##  better_ear_ave   left_right_ave_dB_HL_4000_Hz better_ear_dB_HL_4000_Hz
##  Min.   :-5.000   Min.   :-2.500               Min.   :-10.000
##  1st Qu.: 0.500   1st Qu.: 0.000               1st Qu.: -5.000
##  Median : 2.000   Median : 5.000               Median :  5.000
##  Mean   : 2.895   Mean   : 6.579               Mean   :  3.684
##  3rd Qu.: 4.500   3rd Qu.:11.250               3rd Qu.: 10.000
##  Max.   :17.000   Max.   :30.000               Max.   : 30.000
##
##  left_right_ave_dB_HL_500_Hz better_ear_dB_HL_500_Hz
##  Min.   :-5.000              Min.   :-10.000
##  1st Qu.: 0.000              1st Qu.: -2.500
##  Median : 2.500              Median :  0.000
##  Mean   : 4.211              Mean   :  1.053
##  3rd Qu.: 5.000              3rd Qu.:  5.000
##  Max.   :35.000              Max.   : 20.000
##
```

Data frame from wide to long format

```
library(reshape2)
raw$subject <- factor(raw$subject)
PTA <- melt(raw, id.vars=c(1:3,16:23), measure.name=4:16, variable.name = "condition",  value.name = "dl
cnd = read.table(text = as.character(PTA$condition), sep = "_", colClasses = "character")
PTA <- cbind(PTA, cnd[c(1,4)])
names(PTA)[names(PTA) == 'V1'] <- 'EAR'
names(PTA)[names(PTA) == 'V4'] <- 'Frequency'
PTA$EAR <- factor(PTA$EAR)
PTA$Frequency = factor(as.numeric(PTA$Frequency))
PTA <- PTA[, -4:-12]
head(PTA, n = 5)
```

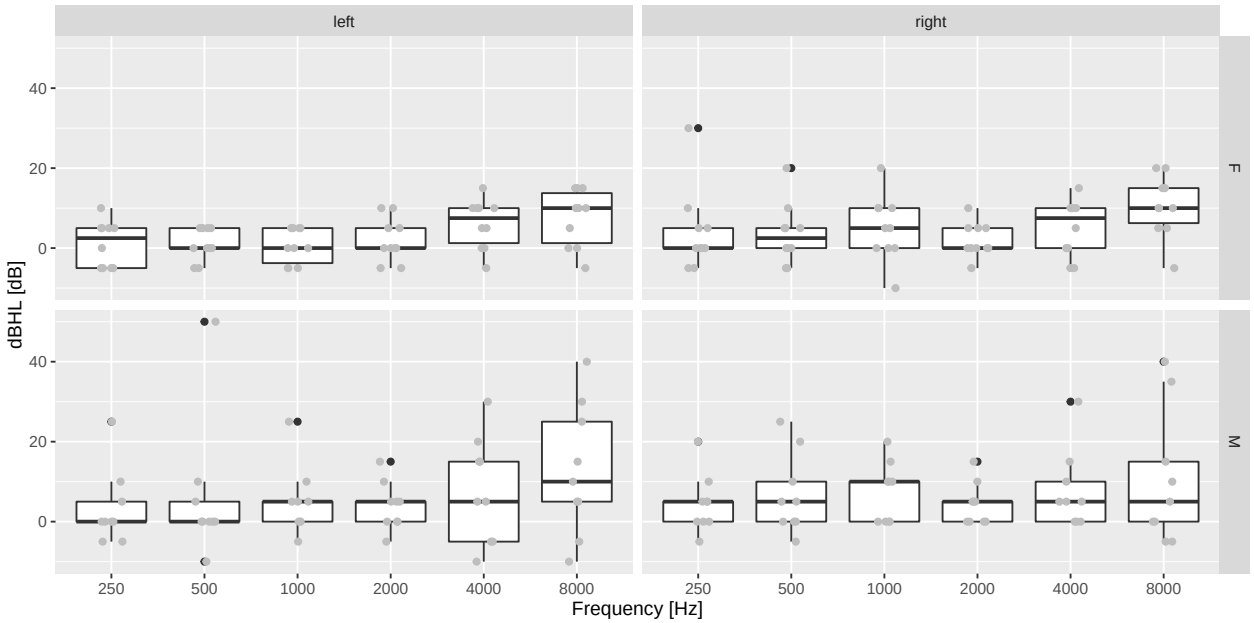
```
##   subject Age_years Sex dBHL   EAR Frequency
## 1     1WR  48.11781   M   20 right       250
## 2     2PP  22.35616   F   -5 right       250
## 3     3LS  25.58356   M   -5 right       250
## 4     4BB  39.73973   F    5 right       250
## 5     5MD  47.80000   M    0 right       250
```

Plotting with ggplot

```
library(ggplot2)
gp <- (ggplot(data = PTA,
        aes(x = Frequency,
            y = dBHL
        ))
+ geom_boxplot(notch=F)
+ geom_jitter(mapping=aes(x=Frequency, y=dBHL), width=0.3, height=0, color='gray')
+ facet_grid(Sex~EAR)
+ xlab("Frequency [Hz]")
+ ylab ("dBHL [dB]"))
```

Plotting with ggplot

```
gp
```



Adding new Factors

```
median_age <- round(median(raw$Age_years))
PTA$AGE_GROUP <- ifelse(PTA$Age_years < median_age, paste("below_", as.character(median_age),sep=""),
                        paste("above_", as.character(median_age),sep=""))
PTA$AGE_GROUP <- factor(PTA$AGE_GROUP)
head(PTA[ ,c(1:5, ncol(PTA))], n=10)
```

```
##    subject Age_years Sex dBHL   EAR AGE_GROUP
## 1      1WR  48.11781   M   20 right  above_35
## 2      2PP  22.35616   F   -5 right  below_35
## 3      3LS  25.58356   M   -5 right  below_35
## 4      4BB  39.73973   F    5 right  above_35
## 5      5MD  47.80000   M    0 right  above_35
```

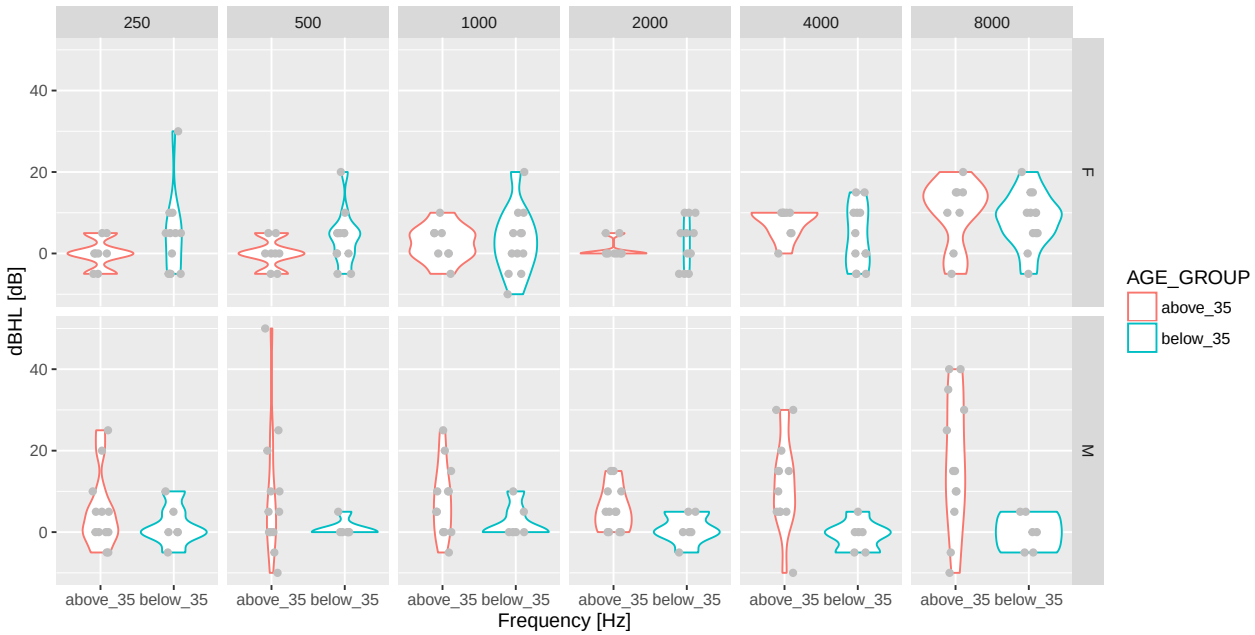
```
## 6      6DD  49.87397   F    0 right  above_35
## 7      8BV  39.60822   F    0 right  above_35
## 8      9VF  38.01918   M    5 right  above_35
## 9     10AA  20.09041   F   10 right  below_35
## 10    11RR  23.01096   F   30 right  below_35
```

Plotting with ggplot

```
gp <- (ggplot(data = PTA,
        aes(x = AGE_GROUP,
            y = dBHL,
            color=AGE_GROUP
        ))
+ geom_violin()
+ geom_jitter(mapping=aes(x=AGE_GROUP, y=dBHL), width=0.3, height=0, color='gray')
+ facet_grid(Sex~Frequency)
+ xlab("Frequency [Hz]")
+ ylab ("dBHL [dB]"))
```

Plotting with ggplot

```
gp
```



ANOVA analysis

```
library(ez)
library(pander)
anv = ezANOVA(data = PTA
              , dv = .(dBHL)
```

```
              , wid = .(subject)
              , within = .(Frequency, EAR)
              , between = .(Sex, AGE_GROUP)
              , detailed = T
              , type = 2
)
```

ANOVA analysis

```
##
##
## | Effect                        | DFn   | DFd   | SSn   | SSd   | F   | p      | p<.05   | ges   |
## |:------------------------------|:------|:------|:------|:------|:----|:-------|:--------|:------|
## | (Intercept)                   | 1     | 15    | 6316  | 5280  | 18  | 7e-04  | *       | 0.3   |
## | Sex                           | 1     | 15    | 156   | 5280  | 0.4 | 0.5    |         | 0.01  |
## | AGE_GROUP                     | 1     | 15    | 669   | 5280  | 2   | 0.2    |         | 0.04  |
## | Frequency                     | 5     | 75    | 1449  | 5449  | 4   | 0.003  | *       | 0.09  |
## | EAR                           | 1     | 15    | 86    | 1245  | 1   | 0.3    |         | 0.006 |
## | Sex:AGE_GROUP                 | 1     | 15    | 1185  | 5280  | 3   | 0.09   |         | 0.08  |
## | Sex:Frequency                 | 5     | 75    | 81    | 5449  | 0.2 | 1      |         | 0.006 |
## | AGE_GROUP:Frequency           | 5     | 75    | 640   | 5449  | 2   | 0.1    |         | 0.04  |
## | Sex:EAR                       | 1     | 15    | 23    | 1245  | 0.3 | 0.6    |         | 0.002 |
## | AGE_GROUP:EAR                 | 1     | 15    | 19    | 1245  | 0.2 | 0.6    |         | 0.001 |
## | Frequency:EAR                 | 5     | 75    | 97    | 2492  | 0.6 | 0.7    |         | 0.007 |
## | Sex:AGE_GROUP:Frequency       | 5     | 75    | 163   | 5449  | 0.4 | 0.8    |         | 0.01  |
## | Sex:AGE_GROUP:EAR             | 1     | 15    | 33    | 1245  | 0.4 | 0.5    |         | 0.002 |
## | Sex:Frequency:EAR             | 5     | 75    | 58    | 2492  | 0.4 | 0.9    |         | 0.004 |
## | AGE_GROUP:Frequency:EAR       | 5     | 75    | 44    | 2492  | 0.3 | 0.9    |         | 0.003 |
## | Sex:AGE_GROUP:Frequency:EAR   | 5     | 75    | 36    | 2492  | 0.2 | 1      |         | 0.002 |
```

ANOVA Mauchly's Test for Sphericity

```
##
##
## | Effect                        | W    | p      | p<.05   |
## |:------------------------------|:-----|:-------|:--------|
## | Frequency                     | 0.04 | 9e-05  | *       |
## | Sex:Frequency                 | 0.04 | 9e-05  | *       |
## | AGE_GROUP:Frequency           | 0.04 | 9e-05  | *       |
## | Sex:AGE_GROUP:Frequency       | 0.04 | 9e-05  | *       |
## | Frequency:EAR                 | 0.1  | 0.01   | *       |
## | Sex:Frequency:EAR             | 0.1  | 0.01   | *       |
## | AGE_GROUP:Frequency:EAR       | 0.1  | 0.01   | *       |
## | Sex:AGE_GROUP:Frequency:EAR   | 0.1  | 0.01   | *       |
```

ANOVA Sphericity Corrections

```
##
##
## | Effect                        | GGe   | p[GG]    | p[GG]<.05    | HFe   | p[HF]    | p[HF]<.05    |
## |:------------------------------|:------|:---------|:-------------|:------|:---------|:-------------|
```

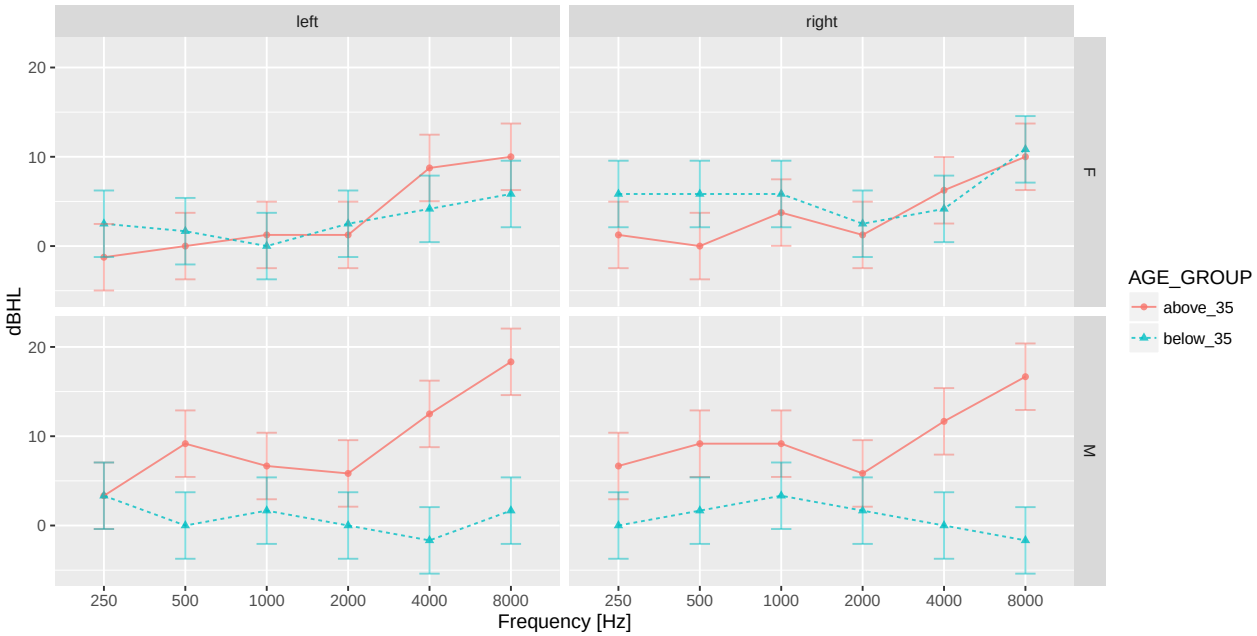
```
## | Frequency                    | 0.4   | 0.03    | *          | 0.5   | 0.02    | *          |
## | Sex:Frequency                | 0.4   | 0.8     |            | 0.5   | 0.8     |            |
## | AGE_GROUP:Frequency          | 0.4   | 0.2     |            | 0.5   | 0.2     |            |
## | Sex:AGE_GROUP:Frequency      | 0.4   | 0.7     |            | 0.5   | 0.7     |            |
## | Frequency:EAR                | 0.6   | 0.6     |            | 0.7   | 0.7     |            |
## | Sex:Frequency:EAR            | 0.6   | 0.8     |            | 0.7   | 0.8     |            |
## | AGE_GROUP:Frequency:EAR      | 0.6   | 0.8     |            | 0.7   | 0.9     |            |
## | Sex:AGE_GROUP:Frequency:EAR  | 0.6   | 0.9     |            | 0.7   | 0.9     |            |
```

ANOVA plots

```
ez1 <- ezPlot(data = PTA
              , x = Frequency
              , dv = .(dBHL)
              , wid = .(subject)
              , within = .(Frequency, EAR)
              , between = .(Sex, AGE_GROUP)
              , type = 2
              , x_lab = "Frequency [Hz]"
              , y_lab = "dBHL"
              , split = AGE_GROUP
              , col = EAR
              , row = Sex
              , print_code = F)
```

ANOVA plots

```
ez1
```

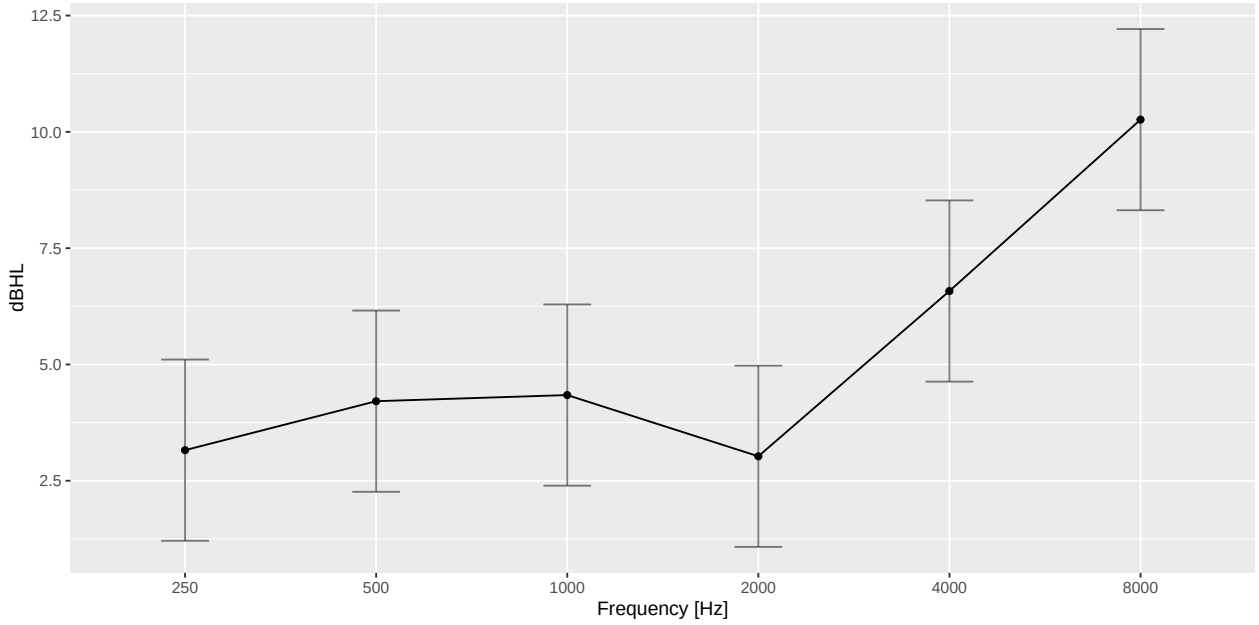


ANOVA plots

```
ez2 <- ezPlot(data = PTA
              , x = Frequency
              , dv = .(dBHL)
              , wid = .(subject)
              , within = .(Frequency)
              , within_full = .(Frequency, EAR)
              , between_full = .(Sex, AGE_GROUP)
              , type = 2
              , x_lab = "Frequency [Hz]"
              , y_lab = "dBHL"
              , print_code = F)
```

ANOVA plots

```
ez2
```



Correlations

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(plyr)
my_corr_f <- function(x,y)
{
```

```
  r_corr <- cor.test(x, y)
  b <- lm(x ~ y)
  tab_cor <- data.frame(Value = b$coefficients[2],
                        Std.Error=summary(b)$sigma,
                        t = r_corr$statistic,
                        df = r_corr$parameter,
                        p = r_corr$p.value,
                        r = r_corr$estimate)
  row.names(tab_cor) <- NULL
  return(tab_cor)
}
```

Correlations

```
corrs <- ddply(PTA, .(Sex, EAR, Frequency), function(df) my_corr_f(df$Age_years, df$dBHL))
pandoc.table(corrs, split.table = Inf, digits = 1, style="rmarkdown", justify = 'left')
```

```
##
##
## | Sex   | EAR   | Frequency   | Value   | Std.Error   | t       | df   | p     | r       |
## |:------|:------|:------------|:--------|:------------|:--------|:-----|:------|:--------|
## | F     | left  | 250         | -0.8    | 12          | -1      | 8    | 0.3   | -0.4    |
## | F     | left  | 500         | -0.3    | 12          | -0.3    | 8    | 0.8   | -0.1    |
## | F     | left  | 1000        | 1       | 11          | 1       | 8    | 0.2   | 0.4     |
## | F     | left  | 2000        | -0.09   | 13          | -0.1    | 8    | 0.9   | -0.04   |
## | F     | left  | 4000        | 0.6     | 12          | 0.9     | 8    | 0.4   | 0.3     |
## | F     | left  | 8000        | 0.6     | 12          | 1       | 8    | 0.3   | 0.4     |
## | F     | right | 250         | -0.4    | 12          | -1      | 8    | 0.3   | -0.3    |
## | F     | right | 500         | -0.6    | 11          | -1      | 8    | 0.2   | -0.4    |
## | F     | right | 1000        | -0.003  | 13          | -0.007  | 8    | 1     | -0.002  |
## | F     | right | 2000        | -0.04   | 13          | -0.05   | 8    | 1     | -0.02   |
## | F     | right | 4000        | 0.1     | 12          | 0.2     | 8    | 0.8   | 0.07    |
## | F     | right | 8000        | 0.2     | 12          | 0.4     | 8    | 0.7   | 0.1     |
## | M     | left  | 250         | 0.2     | 10          | 0.5     | 7    | 0.7   | 0.2     |
## | M     | left  | 500         | 0.3     | 9           | 2       | 7    | 0.2   | 0.5     |
## | M     | left  | 1000        | 0.6     | 9           | 2       | 7    | 0.1   | 0.6     |
## | M     | left  | 2000        | 1       | 8           | 2       | 7    | 0.09  | 0.6     |
## | M     | left  | 4000        | 0.6     | 7           | 3       | 7    | 0.02  | 0.8     |
## | M     | left  | 8000        | 0.4     | 7           | 3       | 7    | 0.03  | 0.7     |
## | M     | right | 250         | 0.6     | 9           | 1       | 7    | 0.2   | 0.5     |
## | M     | right | 500         | 0.3     | 10          | 0.8     | 7    | 0.5   | 0.3     |
## | M     | right | 1000        | 0.6     | 9           | 1       | 7    | 0.2   | 0.5     |
## | M     | right | 2000        | 1       | 9           | 2       | 7    | 0.1   | 0.6     |
## | M     | right | 4000        | 0.8     | 7           | 3       | 7    | 0.02  | 0.8     |
## | M     | right | 8000        | 0.5     | 6           | 4       | 7    | 0.008 | 0.8     |
```

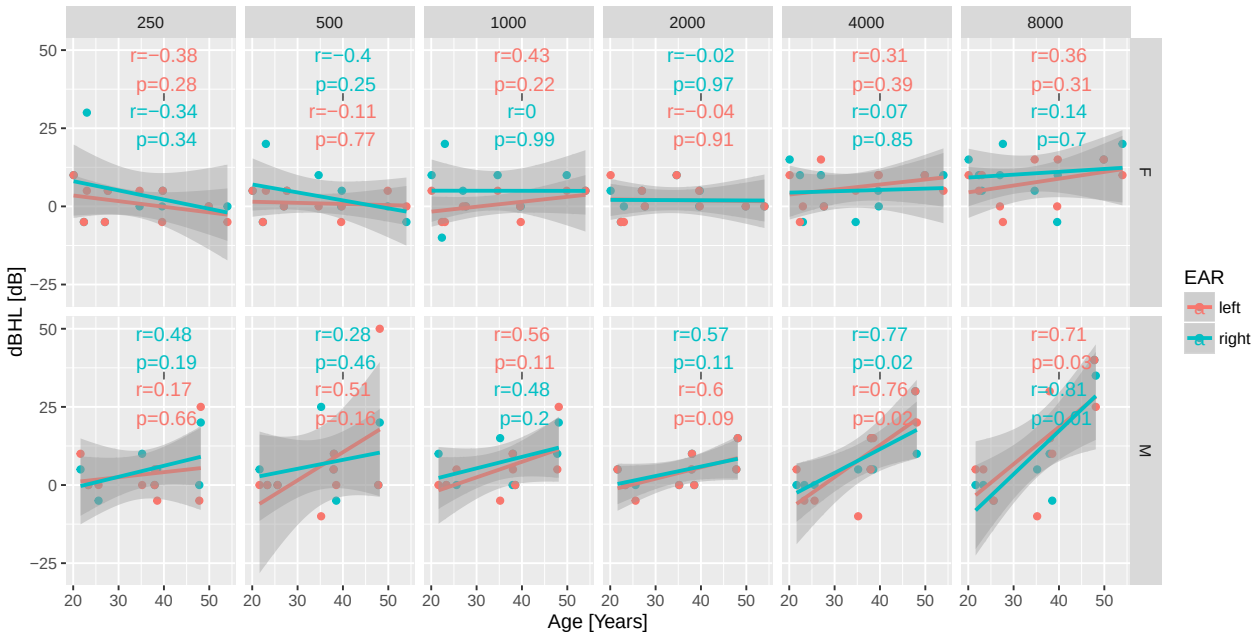
Plotting correlations

```
library(ggrepel)
gp <- (ggplot(data = PTA,
        aes(x = Age_years,
```

```
            y = dBHL,
            color=EAR
        ))
+ geom_point()
+ geom_smooth(method="lm", se=T)
+ geom_text_repel(data=corrs, aes(x=40
                                  , y=35
                                  , color=EAR
                                  , label=paste("r=", round(r, digits = 2), "\n p=", round(p, digits = 2
# + geom_text_repel(aes(label = subject))
+ facet_grid(Sex~Frequency)
+ xlab("Age [Years]")
+ ylab ("dBHL [dB]"))
```

Plotting correlations

```
gp
```



Nonparametric tests

- Independent 2-group Mann-Whitney U Test

```
mean_freq <- ddply(PTA, .(subject, Frequency)
                   , summarise
                   , dBHL = mean(dBHL)
                   , AGE_GROUP = unique(AGE_GROUP))

wt <- wilcox.test(dBHL ~ AGE_GROUP, data=mean_freq)
tab <- data.frame(W=wt$statistic, p.value=wt$p.value)
pandoc.table(tab, split.table = Inf, digits = 1, style="rmarkdown", justify = 'left')
```

```
##
##
## |    | W    | p.value   |
## |:--------|:-----|:----------|
## | **W**   | 1953 | 0.06      |
```

Nonparametric tests

- Kruskal Wallis Test One Way Anova by Ranks

```
kt <- kruskal.test(dBHL ~ AGE_GROUP, data=mean_freq)
tab <- data.frame(chi.squared=kt$statistic, df=kt$parameter ,p.value=kt$p.value)
pandoc.table(tab, split.table = Inf, digits = 1, style="rmarkdown", justify = 'left')
```

```
##
##
## |                            | chi.squared   | df   | p.value   |
## |:--------------------------------|:--------------|:-----|:----------|
## | **Kruskal-Wallis chi-squared**  | 4             | 1    | 0.06      |
```

```
kt <- kruskal.test(dBHL ~ Frequency, data=mean_freq)
tab <- data.frame(chi.squared=kt$statistic,df=kt$parameter, p.value=kt$p.value)
pandoc.table(tab, split.table = Inf, digits = 1, style="rmarkdown", justify = 'left')
```

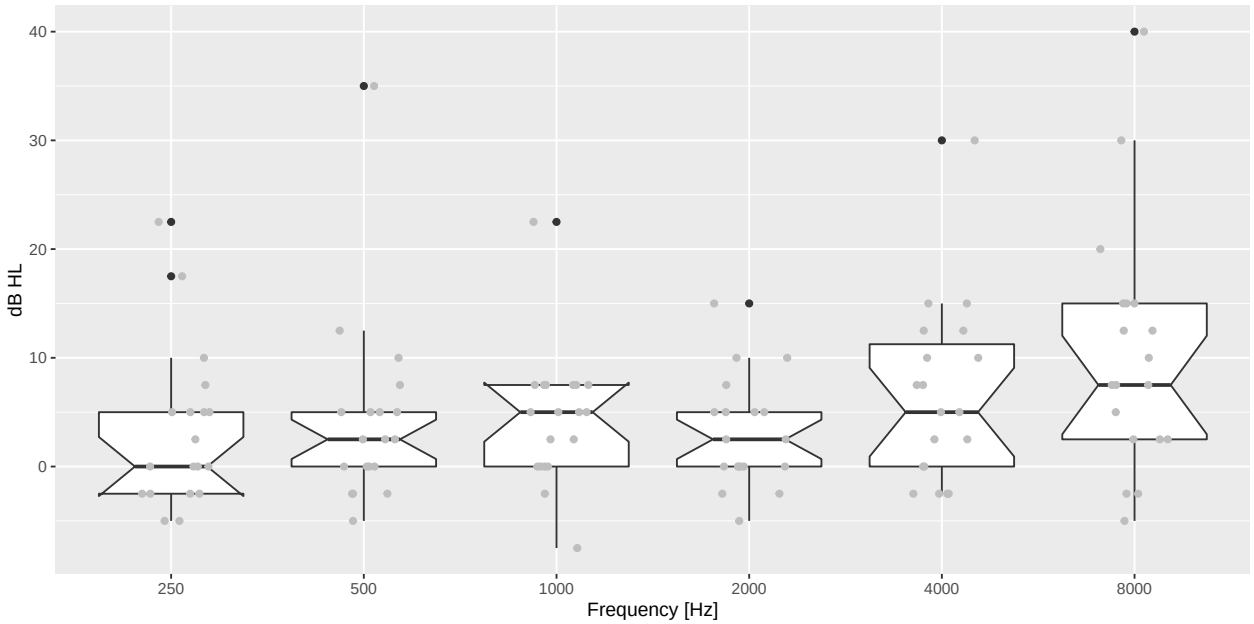
```
##
##
## |                            | chi.squared   | df   | p.value   |
## |:--------------------------------|:--------------|:-----|:----------|
## | **Kruskal-Wallis chi-squared**  | 10            | 5    | 0.09      |
```

```
gp <- (ggplot(data = mean_freq,
        aes(x = Frequency,
            y = dBHL
        ))
+ geom_boxplot(notch=T)
+ geom_jitter(mapping=aes(x=Frequency, y=dBHL), width=0.5, height=0, color='gray')
+ xlab("Frequency [Hz]")
+ ylab ("dB HL"))
```

Nonparametric tests

```
gp
```

```
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
```



Nonparametric tests

- Friedman Test

```
ft <- friedman.test(dBHL ~ Frequency | subject, data=mean_freq)
tab <- data.frame(chi.squared=ft$statistic,df=ft$parameter, p.value=ft$p.value)
pandoc.table(tab, split.table = Inf, digits = 1, style="rmarkdown", justify = 'left')
```

```
##
##
## |                       | chi.squared   | df   | p.value   |
## |:---------------------------|:--------------|:-----|:----------|
## | **Friedman chi-squared**   | 12            | 5    | 0.04      |
```