Final project in Data Science at Harvard University

Exploring indicators of Parkinsons Disease

Jan Thomsen

01/08/2021

Contents

Abstract: This is the final assignment for the Harvard Data Science Professional certificate Program with Professor of Biostatistics Rafael Irizarry from Harvard University.

In this capstone project, we had to choose your own dataset and we have to analyze it and show our machine learning knowledge.

My motivation for diving into the area of Parkinsons Disease is that the last 2-3 years i have lived a process helping my father who has been diagnosed with Alzheimer Disease, which is somewhat related. The process from showing symptoms to actually being diagnosed and then degenerate into an unconscious state, has been a big challenge as the son.

My conditions to do the medical analysis is on on at third party basis and the emphasis has been to show my knowledge that i have accomplished during these courses.

This is also the final assignment for the Harvard Data Science Professional certificate Program with Professor of Biostatistics Rafael Irizarry from Harvard University.

It is the 9th and last course in the Data Science series offered by Harvard University:

- **1. R basics**
- **2. Visualization**
- **3. Probability**
- **4. Inference and modeling**
- **5. Productivity tools**
- **6. Wrangling**
- **7. Linear regression**
- **8. Machine learning**
- **9. Capstone**

In this capstone project, we given the dataset and instructions we have to clean, analyze and modelling it and show our Data Science knowledge."

Chapter 1

Nomenclature

PD Parkingsins Diasease

Matrix column entries (attributes):

- **MDVP:Fo(Hz)** - Average vocal fundamental frequency
- **MDVP:Fhi(Hz)** - Maximum vocal fundamental frequency
- **MDVP:Flo(Hz)** - Minimum vocal fundamental frequency
- **MDVP:Jitter(%)**,
- **MDVP:Jitter(Abs)**,
- **MDVP:RAP,MDVP:PPQ,Jitter:DDP** - Several measures of variation in fundamental frequency
- **MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shim** - Several measures of variation in amplitude
- **NHR,HNR** - Two measures of ratio of noise to tonal components in the voice status - Health status of the subject (one) - Parkinson's, (zero) - healthy
- **RPDE,D2** - Two nonlinear dynamical complexity measures
- **DFA** - Signal fractal scaling exponent
- **spread1,spread2, PPE** - Three nonlinear measures of fundamental frequency variation

Chapter 2

Executive Summary

Firstly the situation with the Parkinsons Disease '(PD)' has become increasingly worrying especially when you experience it entering your personal life.

During the research presented in the dataset, I plan to focus on the following:

1. Explore variables of Parkinsons through voice detection
2. Which variables are important
3. Perform a statistical test to see it is feasible method
4. Views on more advanced model for voice PD detection

For achieving the task of analyzing the dataset I have used various knowledge obtained in the 8 courses, but also my prior knowledge.

Chapter 3

Exploratory Data Analysis

3.1 The Dataset

The dataset is from the website Parkinsons Data Set - Oxford Parkinsons Disease Detection Dataset from 2008 which i find sufficiently challenging for this project. The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinsons disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to status column which is set to 0 for healthy and 1 for PD. There are around six recordings per patient, the name of the patient is identified in the first column.

As the table shows the original data set have 197 instances, 195 observations with 24 variables.

Total sample:

```
## 'data.frame':    195 obs. of  24 variables:
##  $ name            : chr  "phon_R01_S01_1" "phon_R01_S01_2" "phon_R01_S01_3" "phon_R01_S0
##  $ MDVP.Fo.Hz.     : num  120 122 117 117 116 ...
##  $ MDVP.Fhi.Hz.    : num  157 149 131 138 142 ...
##  $ MDVP.Flo.Hz.    : num  75 114 112 111 111 ...
##  $ MDVP.Jitter...  : num  0.00784 0.00968 0.0105 0.00997 0.01284 ...
##  $ MDVP.Jitter.Abs.: num  0.00007 0.00008 0.00009 0.00009 0.00011 0.00008 0.00003 0.00003
##  $ MDVP.RAP        : num  0.0037 0.00465 0.00544 0.00502 0.00655 0.00463 0.00155 0.00144
##  $ MDVP.PPQ        : num  0.00554 0.00696 0.00781 0.00698 0.00908 0.0075 0.00202 0.00182
##  $ Jitter.DDP      : num  0.0111 0.0139 0.0163 0.015 0.0197 ...
##  $ MDVP.Shimmer    : num  0.0437 0.0613 0.0523 0.0549 0.0643 ...
##  $ MDVP.Shimmer.dB.: num  0.426 0.626 0.482 0.517 0.584 0.456 0.14 0.134 0.191 0.255 ...
##  $ Shimmer.APQ3    : num  0.0218 0.0313 0.0276 0.0292 0.0349 ...
##  $ Shimmer.APQ5    : num  0.0313 0.0452 0.0386 0.0401 0.0483 ...
##  $ MDVP.APQ        : num  0.0297 0.0437 0.0359 0.0377 0.0447 ...
##  $ Shimmer.DDA     : num  0.0654 0.094 0.0827 0.0877 0.1047 ...
```

```
##  $ NHR             : num  0.0221 0.0193 0.0131 0.0135 0.0177 ...
##  $ HNR             : num  21 19.1 20.7 20.6 19.6 ...
##  $ status          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ RPDE            : num  0.415 0.458 0.43 0.435 0.417 ...
##  $ DFA             : num  0.815 0.82 0.825 0.819 0.823 ...
##  $ spread1         : num  -4.81 -4.08 -4.44 -4.12 -3.75 ...
##  $ spread2         : num  0.266 0.336 0.311 0.334 0.235 ...
##  $ D2              : num  2.3 2.49 2.34 2.41 2.33 ...
##  $ PPE             : num  0.285 0.369 0.333 0.369 0.41 ...
```

Global is for voice to text not Parkinson's disease we need to use the Coleman status to work separate the two groups into People Without Parkinson disease to underseas and individuals with with Parkinson's disease.

Furthermore we will look into third. There are different variables to see if there is some some indicators that are more important than otherwise it will be you a very advanced job to try analyze the data.

In the following summary if you could get a good overview of the data set and the differences.

Summary of total sample:

```
##      name              MDVP.Fo.Hz.      MDVP.Fhi.Hz.    MDVP.Flo.Hz.
##  Length:195         Min.   : 88.33   Min.   :102.1   Min.   : 65.48
##  Class :character   1st Qu.:117.57   1st Qu.:134.9   1st Qu.: 84.29
##  Mode  :character   Median :148.79   Median :175.8   Median :104.31
##                     Mean   :154.23   Mean   :197.1   Mean   :116.32
##                     3rd Qu.:182.77   3rd Qu.:224.2   3rd Qu.:140.02
##                     Max.   :260.11   Max.   :592.0   Max.   :239.17
##  MDVP.Jitter...     MDVP.Jitter.Abs.       MDVP.RAP           MDVP.PPQ
##  Min.   :0.001680   Min.   :7.000e-06   Min.   :0.000680   Min.   :0.000920
##  1st Qu.:0.003460   1st Qu.:2.000e-05   1st Qu.:0.001660   1st Qu.:0.001860
##  Median :0.004940   Median :3.000e-05   Median :0.002500   Median :0.002690
##  Mean   :0.006220   Mean   :4.396e-05   Mean   :0.003306   Mean   :0.003446
##  3rd Qu.:0.007365   3rd Qu.:6.000e-05   3rd Qu.:0.003835   3rd Qu.:0.003955
##  Max.   :0.033160   Max.   :2.600e-04   Max.   :0.021440   Max.   :0.019580
##    Jitter.DDP        MDVP.Shimmer     MDVP.Shimmer.dB.  Shimmer.APQ3
##  Min.   :0.002040   Min.   :0.00954   Min.   :0.0850   Min.   :0.004550
##  1st Qu.:0.004985   1st Qu.:0.01650   1st Qu.:0.1485   1st Qu.:0.008245
##  Median :0.007490   Median :0.02297   Median :0.2210   Median :0.012790
##  Mean   :0.009920   Mean   :0.02971   Mean   :0.2823   Mean   :0.015664
##  3rd Qu.:0.011505   3rd Qu.:0.03789   3rd Qu.:0.3500   3rd Qu.:0.020265
##  Max.   :0.064330   Max.   :0.11908   Max.   :1.3020   Max.   :0.056470
##   Shimmer.APQ5        MDVP.APQ         Shimmer.DDA           NHR
##  Min.   :0.00570   Min.   :0.00719   Min.   :0.01364   Min.   :0.000650
##  1st Qu.:0.00958   1st Qu.:0.01308   1st Qu.:0.02474   1st Qu.:0.005925
##  Median :0.01347   Median :0.01826   Median :0.03836   Median :0.011660
##  Mean   :0.01788   Mean   :0.02408   Mean   :0.04699   Mean   :0.024847
##  3rd Qu.:0.02238   3rd Qu.:0.02940   3rd Qu.:0.06080   3rd Qu.:0.025640
##  Max.   :0.07940   Max.   :0.13778   Max.   :0.16942   Max.   :0.314820
```

```
##       HNR             status            RPDE             DFA
##  Min.   : 8.441   Min.   :0.0000   Min.   :0.2566   Min.   :0.5743
##  1st Qu.:19.198   1st Qu.:1.0000   1st Qu.:0.4213   1st Qu.:0.6748
##  Median :22.085   Median :1.0000   Median :0.4960   Median :0.7223
##  Mean   :21.886   Mean   :0.7538   Mean   :0.4985   Mean   :0.7181
##  3rd Qu.:25.076   3rd Qu.:1.0000   3rd Qu.:0.5876   3rd Qu.:0.7619
##  Max.   :33.047   Max.   :1.0000   Max.   :0.6852   Max.   :0.8253
##     spread1          spread2               D2              PPE
##  Min.   :-7.965   Min.   :0.006274   Min.   :1.423   Min.   :0.04454
##  1st Qu.:-6.450   1st Qu.:0.174350   1st Qu.:2.099   1st Qu.:0.13745
##  Median :-5.721   Median :0.218885   Median :2.362   Median :0.19405
##  Mean   :-5.684   Mean   :0.226510   Mean   :2.382   Mean   :0.20655
##  3rd Qu.:-5.046   3rd Qu.:0.279234   3rd Qu.:2.636   3rd Qu.:0.25298
##  Max.   :-2.434   Max.   :0.450493   Max.   :3.671   Max.   :0.52737
```

#PD Group and the Control Group

It will be divided into two groups by sobsetting the "status" column.

1. Individual PG The PD Group.
2. Individuals Healthy (Control group)

3.2 PD group sample

```
## 'data.frame':    147 obs. of  24 variables:
##  $ name            : chr  "phon_R01_S01_1" "phon_R01_S01_2" "phon_R01_S01_3" "phon_R01_S0
##  $ MDVP.Fo.Hz.     : num  120 122 117 117 116 ...
##  $ MDVP.Fhi.Hz.    : num  157 149 131 138 142 ...
##  $ MDVP.Flo.Hz.    : num  75 114 112 111 111 ...
##  $ MDVP.Jitter...  : num  0.00784 0.00968 0.0105 0.00997 0.01284 ...
##  $ MDVP.Jitter.Abs.: num  0.00007 0.00008 0.00009 0.00009 0.00011 0.00008 0.00003 0.00003
##  $ MDVP.RAP        : num  0.0037 0.00465 0.00544 0.00502 0.00655 0.00463 0.00155 0.00144
##  $ MDVP.PPQ        : num  0.00554 0.00696 0.00781 0.00698 0.00908 0.0075 0.00202 0.00182
##  $ Jitter.DDP      : num  0.0111 0.0139 0.0163 0.015 0.0197 ...
##  $ MDVP.Shimmer    : num  0.0437 0.0613 0.0523 0.0549 0.0643 ...
##  $ MDVP.Shimmer.dB.: num  0.426 0.626 0.482 0.517 0.584 0.456 0.14 0.134 0.191 0.255 ...
##  $ Shimmer.APQ3    : num  0.0218 0.0313 0.0276 0.0292 0.0349 ...
##  $ Shimmer.APQ5    : num  0.0313 0.0452 0.0386 0.0401 0.0483 ...
##  $ MDVP.APQ        : num  0.0297 0.0437 0.0359 0.0377 0.0447 ...
##  $ Shimmer.DDA     : num  0.0654 0.094 0.0827 0.0877 0.1047 ...
##  $ NHR             : num  0.0221 0.0193 0.0131 0.0135 0.0177 ...
##  $ HNR             : num  21 19.1 20.7 20.6 19.6 ...
##  $ status          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ RPDE            : num  0.415 0.458 0.43 0.435 0.417 ...
##  $ DFA             : num  0.815 0.82 0.825 0.819 0.823 ...
##  $ spread1         : num  -4.81 -4.08 -4.44 -4.12 -3.75 ...
##  $ spread2         : num  0.266 0.336 0.311 0.334 0.235 ...
```

```
##  $ D2              : num  2.3 2.49 2.34 2.41 2.33 ...
##  $ PPE             : num  0.285 0.369 0.333 0.369 0.41 ...
```

```
##      name              MDVP.Fo.Hz.      MDVP.Fhi.Hz.    MDVP.Flo.Hz.
##  Length:147         Min.   : 88.33   Min.   :102.1   Min.   : 65.48
##  Class :character   1st Qu.:117.57   1st Qu.:133.8   1st Qu.: 80.88
##  Mode  :character   Median :145.17   Median :163.3   Median : 99.77
##                     Mean   :145.18   Mean   :188.4   Mean   :106.89
##                     3rd Qu.:170.07   3rd Qu.:207.2   3rd Qu.:129.24
##                     Max.   :223.36   Max.   :588.5   Max.   :199.02
##  MDVP.Jitter...     MDVP.Jitter.Abs.       MDVP.RAP           MDVP.PPQ
##  Min.   :0.001680   Min.   :1.000e-05   Min.   :0.000680   Min.   :0.00092
##  1st Qu.:0.004005   1st Qu.:3.000e-05   1st Qu.:0.002030   1st Qu.:0.00219
##  Median :0.005440   Median :4.000e-05   Median :0.002840   Median :0.00314
##  Mean   :0.006989   Mean   :5.068e-05   Mean   :0.003757   Mean   :0.00390
##  3rd Qu.:0.007670   3rd Qu.:6.000e-05   3rd Qu.:0.004100   3rd Qu.:0.00436
##  Max.   :0.033160   Max.   :2.600e-04   Max.   :0.021440   Max.   :0.01958
##    Jitter.DDP        MDVP.Shimmer     MDVP.Shimmer.dB.  Shimmer.APQ3
##  Min.   :0.002040   Min.   :0.01022   Min.   :0.0900   Min.   :0.004550
##  1st Qu.:0.006085   1st Qu.:0.01829   1st Qu.:0.1680   1st Qu.:0.009135
##  Median :0.008530   Median :0.02838   Median :0.2630   Median :0.014840
##  Mean   :0.011273   Mean   :0.03366   Mean   :0.3212   Mean   :0.017676
##  3rd Qu.:0.012300   3rd Qu.:0.04253   3rd Qu.:0.3945   3rd Qu.:0.022815
##  Max.   :0.064330   Max.   :0.11908   Max.   :1.3020   Max.   :0.056470
##   Shimmer.APQ5        MDVP.APQ        Shimmer.DDA           NHR
##  Min.   :0.00570   Min.   :0.00811   Min.   :0.01364   Min.   :0.002310
##  1st Qu.:0.01057   1st Qu.:0.01555   1st Qu.:0.02740   1st Qu.:0.008445
##  Median :0.01650   Median :0.02157   Median :0.04451   Median :0.016580
##  Mean   :0.02028   Mean   :0.02760   Mean   :0.05303   Mean   :0.029211
##  3rd Qu.:0.02493   3rd Qu.:0.03483   3rd Qu.:0.06846   3rd Qu.:0.027960
##  Max.   :0.07940   Max.   :0.13778   Max.   :0.16942   Max.   :0.314820
##       HNR              status        RPDE              DFA
##  Min.   : 8.441   Min.   :1   Min.   :0.2637   Min.   :0.5743
##  1st Qu.:18.782   1st Qu.:1   1st Qu.:0.4391   1st Qu.:0.6856
##  Median :21.414   Median :1   Median :0.5305   Median :0.7267
##  Mean   :20.974   Mean   :1   Mean   :0.5168   Mean   :0.7254
##  3rd Qu.:24.165   3rd Qu.:1   3rd Qu.:0.6046   3rd Qu.:0.7649
##  Max.   :29.928   Max.   :1   Max.   :0.6852   Max.   :0.8253
##     spread1          spread2              D2             PPE
##  Min.   :-7.121   Min.   :0.06341   Min.   :1.766   Min.   :0.09319
##  1st Qu.:-6.038   1st Qu.:0.19951   1st Qu.:2.181   1st Qu.:0.17010
##  Median :-5.440   Median :0.24088   Median :2.440   Median :0.22272
##  Mean   :-5.333   Mean   :0.24813   Mean   :2.456   Mean   :0.23383
##  3rd Qu.:-4.664   3rd Qu.:0.30366   3rd Qu.:2.668   3rd Qu.:0.27440
##  Max.   :-2.434   Max.   :0.45049   Max.   :3.671   Max.   :0.52737
```

3.3 Control group sample

With healthy individuals to control how how are healthy individuals has their attributes on the voice detection.

```
## 'data.frame':    48 obs. of  24 variables:
##  $ name            : chr  "phon_R01_S07_1" "phon_R01_S07_2" "phon_R01_S07_3" "phon_R01_S0
##  $ MDVP.Fo.Hz.     : num  197 199 198 202 203 ...
##  $ MDVP.Fhi.Hz.    : num  207 210 215 212 212 ...
##  $ MDVP.Flo.Hz.    : num  192 192 193 197 196 ...
##  $ MDVP.Jitter...  : num  0.00289 0.00241 0.00212 0.0018 0.00178 0.00198 0.00298 0.00281
##  $ MDVP.Jitter.Abs.: num  1e-05 1e-05 1e-05 9e-06 9e-06 1e-05 1e-05 1e-05 9e-06 9e-06 ...
##  $ MDVP.RAP        : num  0.00166 0.00134 0.00113 0.00093 0.00094 0.00105 0.00169 0.00157
##  $ MDVP.PPQ        : num  0.00168 0.00138 0.00135 0.00107 0.00106 0.00115 0.00182 0.00173
##  $ Jitter.DDP      : num  0.00498 0.00402 0.00339 0.00278 0.00283 0.00314 0.00507 0.0047
##  $ MDVP.Shimmer    : num  0.01098 0.01015 0.01263 0.00954 0.00958 ...
##  $ MDVP.Shimmer.dB.: num  0.097 0.089 0.111 0.085 0.085 0.107 0.164 0.154 0.126 0.134 ...
##  $ Shimmer.APQ3    : num  0.00563 0.00504 0.0064 0.00469 0.00468 ...
##  $ Shimmer.APQ5    : num  0.0068 0.00641 0.00825 0.00606 0.0061 ...
##  $ MDVP.APQ        : num  0.00802 0.00762 0.00951 0.00719 0.00726 ...
##  $ Shimmer.DDA     : num  0.0169 0.0151 0.0192 0.0141 0.014 ...
##  $ NHR             : num  0.00339 0.00167 0.00119 0.00072 0.00065 0.00135 0.0074 0.00675
##  $ HNR             : num  26.8 30.9 30.8 32.7 33 ...
##  $ status          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RPDE            : num  0.422 0.432 0.466 0.369 0.34 ...
##  $ DFA             : num  0.741 0.742 0.739 0.742 0.742 ...
##  $ spread1         : num  -7.35 -7.68 -7.07 -7.7 -7.96 ...
##  $ spread2         : num  0.178 0.173 0.175 0.179 0.164 ...
##  $ D2              : num  1.74 2.1 1.51 1.54 1.42 ...
##  $ PPE             : num  0.0856 0.0685 0.0963 0.0561 0.0445 ...
```

```
##      name            MDVP.Fo.Hz.     MDVP.Fhi.Hz.    MDVP.Flo.Hz.
##  Length:48          Min.   :110.7   Min.   :113.6   Min.   : 74.29
##  Class :character   1st Qu.:120.9   1st Qu.:139.4   1st Qu.: 98.24
##  Mode  :character   Median :199.0   Median :231.2   Median :113.94
##                     Mean   :181.9   Mean   :223.6   Mean   :145.21
##                     3rd Qu.:229.1   3rd Qu.:251.2   3rd Qu.:199.18
##                     Max.   :260.1   Max.   :592.0   Max.   :239.17
##  MDVP.Jitter...     MDVP.Jitter.Abs.       MDVP.RAP           MDVP.PPQ
##  Min.   :0.001780   Min.   :7.000e-06   Min.   :0.000920   Min.   :0.001060
##  1st Qu.:0.002655   1st Qu.:1.000e-05   1st Qu.:0.001332   1st Qu.:0.001480
##  Median :0.003355   Median :2.500e-05   Median :0.001625   Median :0.001775
##  Mean   :0.003866   Mean   :2.337e-05   Mean   :0.001925   Mean   :0.002056
##  3rd Qu.:0.004530   3rd Qu.:3.000e-05   3rd Qu.:0.001907   3rd Qu.:0.002227
##  Max.   :0.013600   Max.   :8.000e-05   Max.   :0.006240   Max.   :0.005640
##    Jitter.DDP        MDVP.Shimmer     MDVP.Shimmer.dB.  Shimmer.APQ3
##  Min.   :0.002760   Min.   :0.00954   Min.   :0.0850   Min.   :0.004680
```

```
##  1st Qu.:0.003998   1st Qu.:0.01448   1st Qu.:0.1290   1st Qu.:0.007350
##  Median :0.004875   Median :0.01671   Median :0.1540   Median :0.008775
##  Mean   :0.005776   Mean   :0.01762   Mean   :0.1630   Mean   :0.009504
##  3rd Qu.:0.005725   3rd Qu.:0.02021   3rd Qu.:0.1893   3rd Qu.:0.011513
##  Max.   :0.018730   Max.   :0.04087   Max.   :0.4050   Max.   :0.023360
##   Shimmer.APQ5         MDVP.APQ        Shimmer.DDA           NHR
##  Min.   :0.006060   Min.   :0.00719   Min.   :0.01403   Min.   :0.000650
##  1st Qu.:0.008193   1st Qu.:0.01124   1st Qu.:0.02206   1st Qu.:0.004188
##  Median :0.010225   Median :0.01302   Median :0.02633   Median :0.004825
##  Mean   :0.010509   Mean   :0.01330   Mean   :0.02851   Mean   :0.011483
##  3rd Qu.:0.011980   3rd Qu.:0.01595   3rd Qu.:0.03454   3rd Qu.:0.009213
##  Max.   :0.024980   Max.   :0.02745   Max.   :0.07008   Max.   :0.107150
##       HNR             status       RPDE             DFA            spread1
##  Min.   :17.88   Min.   :0   Min.   :0.2566   Min.   :0.6267   Min.   :-7.965
##  1st Qu.:22.99   1st Qu.:0   1st Qu.:0.3721   1st Qu.:0.6543   1st Qu.:-7.258
##  Median :25.00   Median :0   Median :0.4354   Median :0.6825   Median :-6.826
##  Mean   :24.68   Mean   :0   Mean   :0.4426   Mean   :0.6957   Mean   :-6.759
##  3rd Qu.:26.14   3rd Qu.:0   3rd Qu.:0.5077   3rd Qu.:0.7423   3rd Qu.:-6.350
##  Max.   :33.05   Max.   :0   Max.   :0.6638   Max.   :0.7857   Max.   :-5.199
##     spread2               D2              PPE
##  Min.   :0.006274   Min.   :1.423   Min.   :0.04454
##  1st Qu.:0.120623   1st Qu.:1.974   1st Qu.:0.09466
##  Median :0.167356   Median :2.130   Median :0.11512
##  Mean   :0.160292   Mean   :2.154   Mean   :0.12302
##  3rd Qu.:0.193766   3rd Qu.:2.339   3rd Qu.:0.14776
##  Max.   :0.291954   Max.   :2.882   Max.   :0.25240
```

First firstly you can see that if you compare the two samples just by looking at the 6 digit summary. It's obvious that there is a difference but later we will diving to the t-test that will show all we with statistical certainty can conclude that there is a difference.

Chapter 4

Sample - comments on population and samples

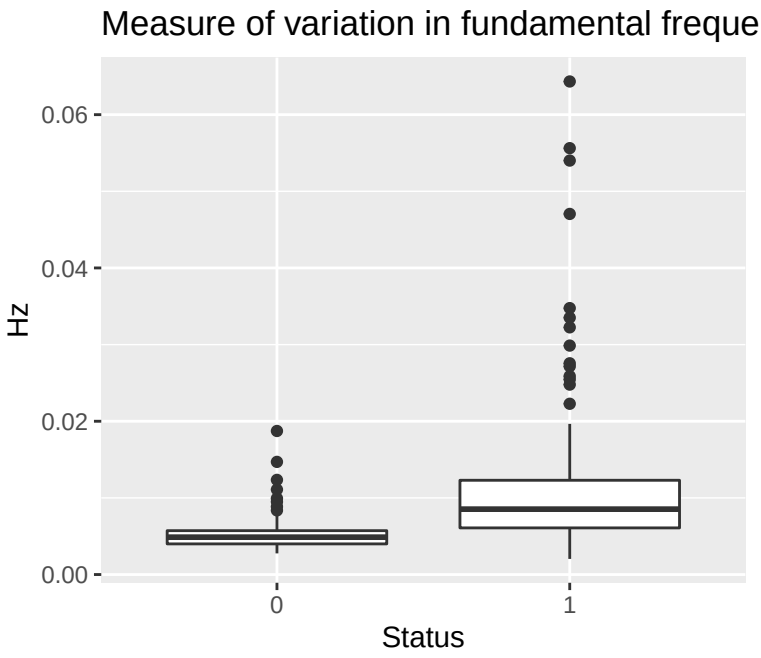
The two samples first taken are Representing the two groups with 23 8 8 individual with appr. 6 measures, **48/8 and 147/23**.

Will it say something about the population? No, it will not say anything about the population, because we do not knowvthe proportions of the two groups in the population and the sampling method. as we are informed the measurements was taking with “replacement”, because there are 6 per individual.

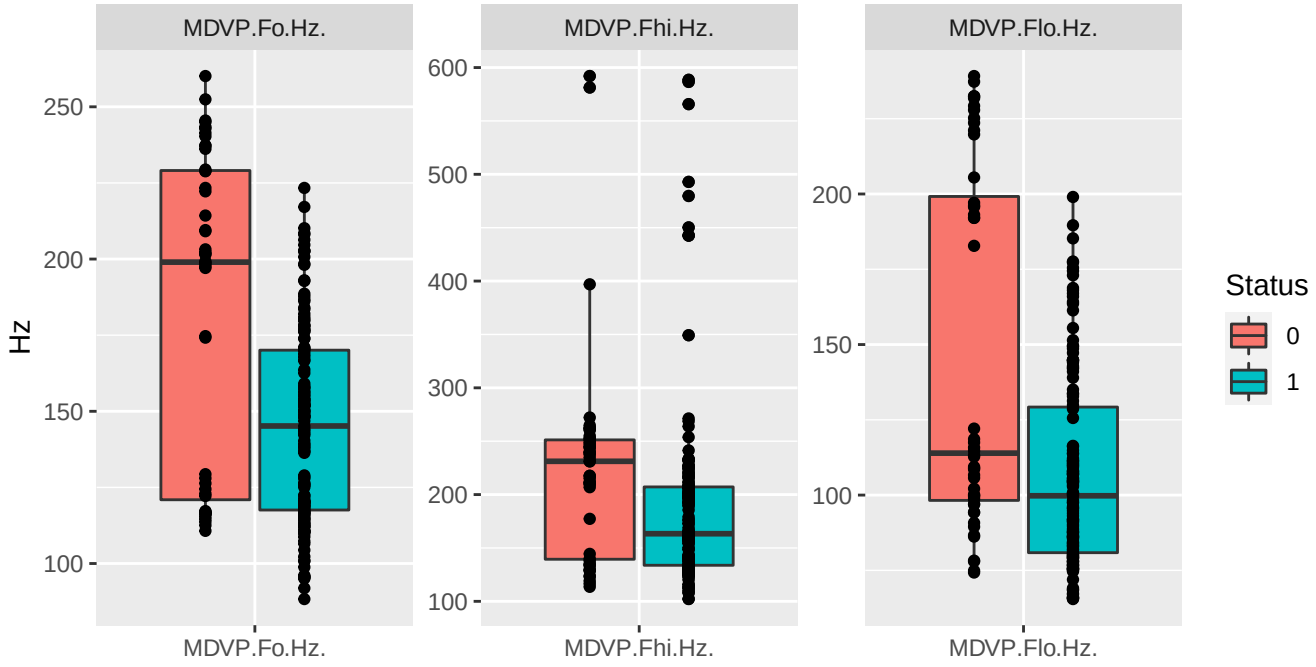
Chapter 5

Visualization of differences in the two samples

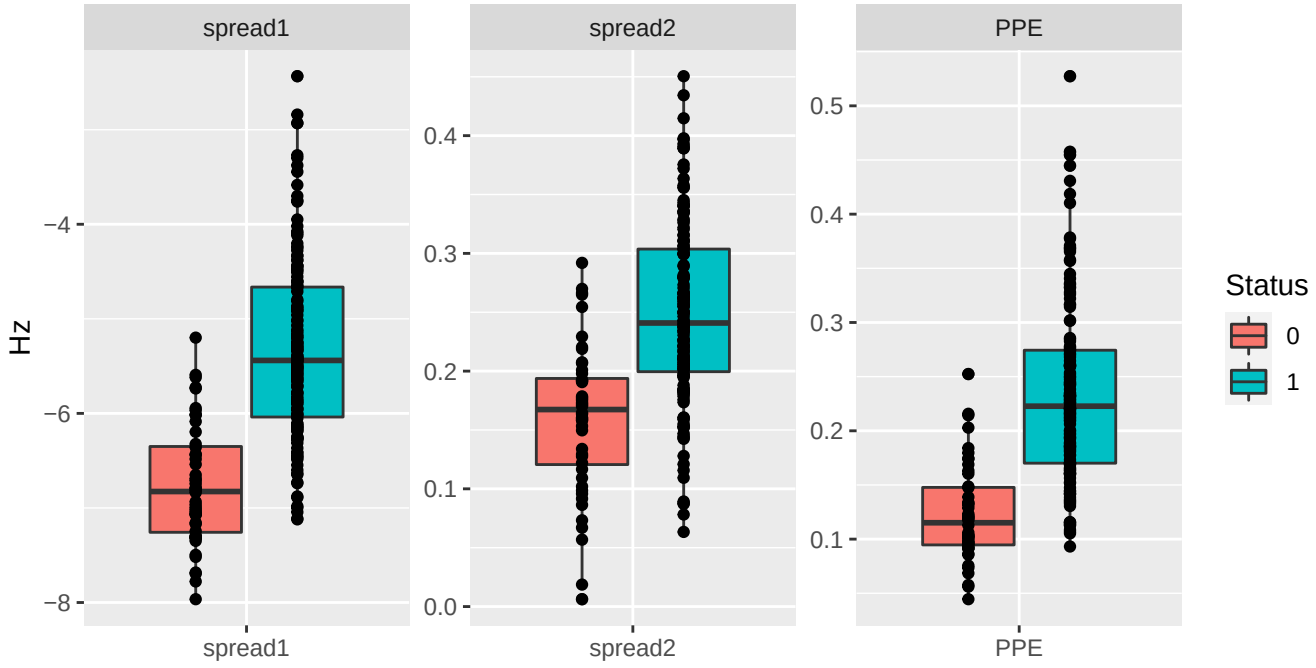
5.1 Boxplots



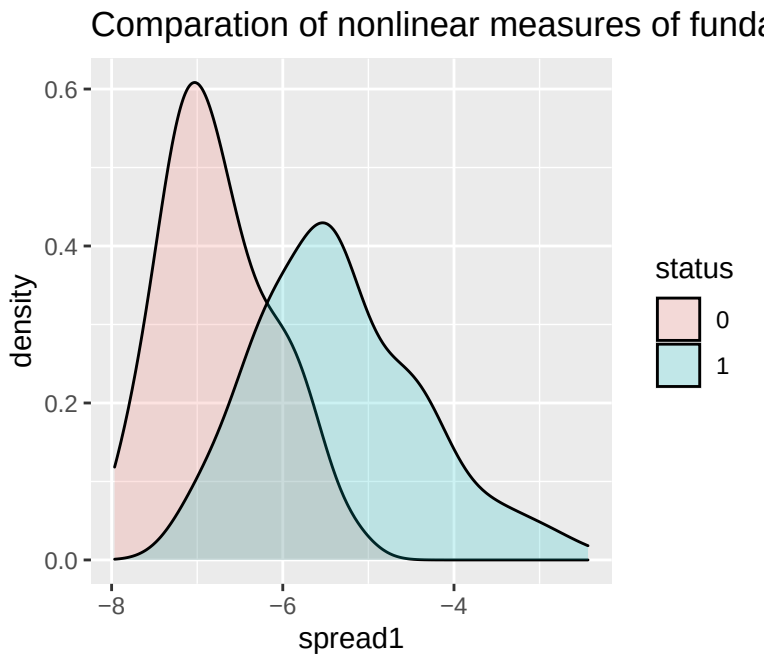
Vocal fundamental frequencys



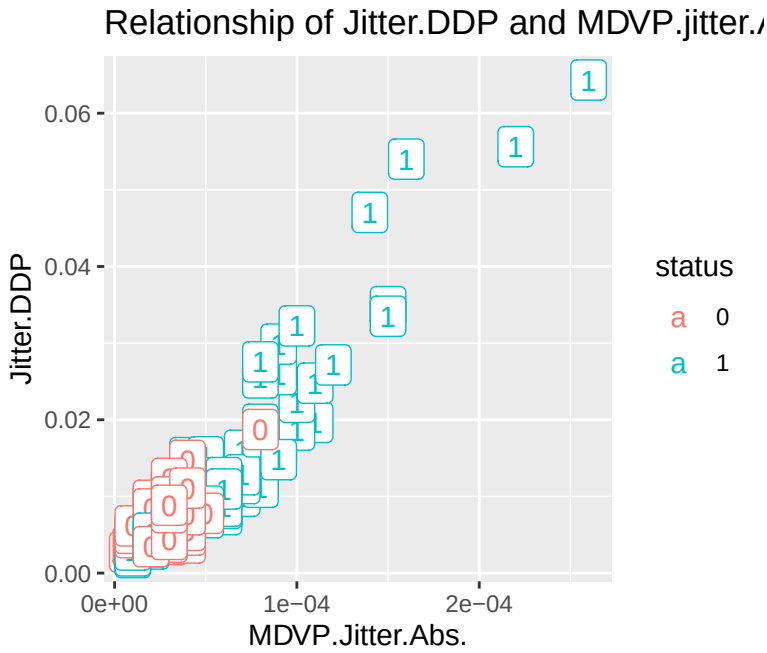
Overview important variables



5.2 Density plots



5.3 Correlations



Plots of pairs of features after pre-processing by range normalization, showing examples of high correlation

Correlation on total sample:

```
## [1] 0.922913
```

Sample PD

```
## [1] 0.9300233
```

Sample Control group

```
## [1] 0.6621335
```

You should see from the future served the redWhich is the single the PD this clustering towards the end PDD jitter This clustering on the lower side of the left sideThe chartAnd who is closely at the higher end of the scaleIt seems to haveA Good fitNow we will calculateBuilding a modelAnd are closely with the circles chat.

To see If there is a statistical significance are went toSee if there is a difference in differences in the means. The two samples we need to majorThe t-tests are off to independentSamplesThat

Appropriate topic for them the next section.

Chapter 6

Unpaired Two-Samples T-test

6.1 Assumptions

Assumption 1: Are the two samples independents? Yes, since the samples from the PD group and Control Group are not related.

Assumption 2: Are the data from each of the 2 groups follow a normal distribution? I will use Shapiro-Wilk normality test.

Null hypothesis: the data are normally distributed

Alternative hypothesis: the data are not normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  pd$spread1
## W = 0.984, p-value = 0.02568
```

From the output, the p-value < 0.05 implying that the distribution of the data **are** significantly different from normal distribution. In other words, we **cannot** assume the normality.

However I choose to perform the t-test on spread1.

6.2 Statistical hypotheses

$$H_0 : m_A = m_B$$
$$H_a : m_A \neq m_B$$

6.3 Citations on unpaired two-samples t-test

If the variance of the two groups PD and control group are equivalent **homoscedasticity**, the t-test value, comparing the two samples **A and B**, can be calculated as follows.

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

where, * m_A and m_B represent the mean value of the group A and B, respectively. * n_A and n_B and nBnB represent the sizes of the group A and B, respectively. * S^2 is an estimator of the pooled variance of the two groups. It can be calculated as follows:

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

Welch t-statistic is calculated as follows:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

The degrees of freedom of Welch t-test is estimated as follows:

$$df = (\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B^2})/(\frac{S_A^4}{n_A^2(n_B - 1)} + \frac{S_B^4}{n_B^2(n_B - 1)})$$

Compute t-test

```
##
##  Two Sample t-test
##
## data:  spread1 by status
## t = -9.5092, df = 193, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.721584 -1.130104
## sample estimates:
## mean in group 0 mean in group 1
##       -6.759264       -5.333420
```

If the p-value is inferior or equal to the significance level 0.05, we can reject the null hypothesis and accept the alternative hypothesis. In other words, we can conclude that the mean values of group PD and Control Group are **significantly different**.

Chapter 7

Ensembles

The idea of an ensemble is similar to the idea of combining data from different pollsters to obtain a better estimate of the true support for each candidate.

In machine learning, one can usually greatly improve the final results by combining the results of different algorithms.

This case is clearly a subjet and dataset for improving with ensembles.

Chapter 8

Conclusion/results

The research i have made on this project analyzing 195 instances with 24 variables i reached the following conclusions:

1. The voice levels of individuals with Parkinsons disease have a different voice level than the control group.
2. I have visualized an important part of the data set with a few charts
3. The future work will be on emsebling the data variables.

My take-out from the course The knowledge that i have gained before this course - was started in 1990 with statistics with pen and paper, so I have refreshed and updated my statistics, but most importantly - R - statistics calculations, R Markdown and the power of the new technology, i can certainly relate to now.

It has been a very interesting journey from my perspective. The core for me is that it is very important to improve the communication of knowledge, the visual part. That's a science. Its far more important than the quantity of charts that you may have in one report.

The idea must be to make complex relations into simple keystrokes for the audience or target group.

January 2021

Jan Thomsen

Chapter 9

Appendix

9.1 1.1 - The total datasets

The PD dataset

```
## 'data.frame':    147 obs. of  24 variables:
##  $ name            : chr  "phon_R01_S01_1" "phon_R01_S01_2" "phon_R01_S01_3" "phon_R01_S0
##  $ MDVP.Fo.Hz.     : num  120 122 117 117 116 ...
##  $ MDVP.Fhi.Hz.    : num  157 149 131 138 142 ...
##  $ MDVP.Flo.Hz.    : num  75 114 112 111 111 ...
##  $ MDVP.Jitter...  : num  0.00784 0.00968 0.0105 0.00997 0.01284 ...
##  $ MDVP.Jitter.Abs.: num  0.00007 0.00008 0.00009 0.00009 0.00011 0.00008 0.00003 0.00003
##  $ MDVP.RAP        : num  0.0037 0.00465 0.00544 0.00502 0.00655 0.00463 0.00155 0.00144
##  $ MDVP.PPQ        : num  0.00554 0.00696 0.00781 0.00698 0.00908 0.0075 0.00202 0.00182
##  $ Jitter.DDP      : num  0.0111 0.0139 0.0163 0.015 0.0197 ...
##  $ MDVP.Shimmer    : num  0.0437 0.0613 0.0523 0.0549 0.0643 ...
##  $ MDVP.Shimmer.dB.: num  0.426 0.626 0.482 0.517 0.584 0.456 0.14 0.134 0.191 0.255 ...
##  $ Shimmer.APQ3    : num  0.0218 0.0313 0.0276 0.0292 0.0349 ...
##  $ Shimmer.APQ5    : num  0.0313 0.0452 0.0386 0.0401 0.0483 ...
##  $ MDVP.APQ        : num  0.0297 0.0437 0.0359 0.0377 0.0447 ...
##  $ Shimmer.DDA     : num  0.0654 0.094 0.0827 0.0877 0.1047 ...
##  $ NHR             : num  0.0221 0.0193 0.0131 0.0135 0.0177 ...
##  $ HNR             : num  21 19.1 20.7 20.6 19.6 ...
##  $ status          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ RPDE            : num  0.415 0.458 0.43 0.435 0.417 ...
##  $ DFA             : num  0.815 0.82 0.825 0.819 0.823 ...
##  $ spread1         : num  -4.81 -4.08 -4.44 -4.12 -3.75 ...
##  $ spread2         : num  0.266 0.336 0.311 0.334 0.235 ...
##  $ D2              : num  2.3 2.49 2.34 2.41 2.33 ...
##  $ PPE             : num  0.285 0.369 0.333 0.369 0.41 ...
```

The healthy dataset

```
## 'data.frame':    48 obs. of  24 variables:
```

```
##  $ name            : chr  "phon_R01_S07_1" "phon_R01_S07_2" "phon_R01_S07_3" "phon_R01_S0
##  $ MDVP.Fo.Hz.     : num  197 199 198 202 203 ...
##  $ MDVP.Fhi.Hz.    : num  207 210 215 212 212 ...
##  $ MDVP.Flo.Hz.    : num  192 192 193 197 196 ...
##  $ MDVP.Jitter...  : num  0.00289 0.00241 0.00212 0.0018 0.00178 0.00198 0.00298 0.00281
##  $ MDVP.Jitter.Abs.: num  1e-05 1e-05 1e-05 9e-06 9e-06 1e-05 1e-05 1e-05 9e-06 9e-06 ...
##  $ MDVP.RAP        : num  0.00166 0.00134 0.00113 0.00093 0.00094 0.00105 0.00169 0.00157
##  $ MDVP.PPQ        : num  0.00168 0.00138 0.00135 0.00107 0.00106 0.00115 0.00182 0.00173
##  $ Jitter.DDP      : num  0.00498 0.00402 0.00339 0.00278 0.00283 0.00314 0.00507 0.0047
##  $ MDVP.Shimmer    : num  0.01098 0.01015 0.01263 0.00954 0.00958 ...
##  $ MDVP.Shimmer.dB.: num  0.097 0.089 0.111 0.085 0.085 0.107 0.164 0.154 0.126 0.134 ...
##  $ Shimmer.APQ3    : num  0.00563 0.00504 0.0064 0.00469 0.00468 ...
##  $ Shimmer.APQ5    : num  0.0068 0.00641 0.00825 0.00606 0.0061 ...
##  $ MDVP.APQ        : num  0.00802 0.00762 0.00951 0.00719 0.00726 ...
##  $ Shimmer.DDA     : num  0.0169 0.0151 0.0192 0.0141 0.014 ...
##  $ NHR             : num  0.00339 0.00167 0.00119 0.00072 0.00065 0.00135 0.0074 0.00675
##  $ HNR             : num  26.8 30.9 30.8 32.7 33 ...
##  $ status          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RPDE            : num  0.422 0.432 0.466 0.369 0.34 ...
##  $ DFA             : num  0.741 0.742 0.739 0.742 0.742 ...
##  $ spread1         : num  -7.35 -7.68 -7.07 -7.7 -7.96 ...
##  $ spread2         : num  0.178 0.173 0.175 0.179 0.164 ...
##  $ D2              : num  1.74 2.1 1.51 1.54 1.42 ...
##  $ PPE             : num  0.0856 0.0685 0.0963 0.0561 0.0445 ...
```

9.2 2 - Acknowledgements

‘Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection’ - Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007) Suitability of Dysphonia Measurements for Telemonitoring of Parkinson’s Disease - Patrick Mcsharry University of Oxford, Eric James Hunter Michigan State University, Jennifer Spielman University of Colorado Boulder