# Table of Contents

[Table of Contents 2](#_Toc89346008)

[ANOVA Test 3](#_Toc89346009)

[Assumption 3](#_Toc89346010)

[How to validate the assumption 3](#_Toc89346011)

[What happens if an assumption is violated 3](#_Toc89346012)

[Descriptive Analysis – Parkinson 4](#_Toc89346013)

[Clean Data 4](#_Toc89346014)

[ANOVA Analysis on total\_UPDRS and subject 5](#_Toc89346015)

[Distribution Check 5](#_Toc89346016)

[Hypothesis 5](#_Toc89346017)

[ANOVA Test 6](#_Toc89346018)

[ANOVA Analysis on total\_UPDRS and Age Group 6](#_Toc89346019)

[Convert Age into age group 6](#_Toc89346020)

[Boxplot on Age Group and total\_UPDRS 7](#_Toc89346021)

[Hypothesis 7](#_Toc89346022)

[ANOVA Test – age group 8](#_Toc89346023)

[Homogeneity Test – Bartlett test 8](#_Toc89346024)

[Normality test 8](#_Toc89346025)

# ANOVA Test

Analysis of variance (ANOVA) is the hypothesis-based test, which allows comparisons between three or more data groups. In general, there are two types of ANOVA:

1. one-way ANOVA – Compare the variance in group means by considering one independent variable only.
2. Two-way ANOVA – like one-way ANOVA; however, it tests two factors on a dependent variable.

## Assumption

There are three assumptions on ANOVA Test

1. Normality – each observation is taken from a normal population distribution
2. No multicollinearity – all sample is independence
3. Homoscedasticity – the homogeneity of variance Y is constant across the range of X value.

Null Hypothesis H0: The means of all groups are equal

Alternative Hypothesis Ha: the mean of at least one group is different.

## How to validate the assumption

There are some ways to validate the ANOVA assumption.

1. Visualization
   1. Histogram - can give an idea of data where it is normally distributed.
   2. Boxplot – can look at the length of the boxplot to determine whether the variable is identical.
2. Q-Q Plot in R – we can plot the observations onto Q-Q Plot and add a straight diagonal line to check homoscedasticity.
3. Statistical test
   1. Shapiro-Wilk test to check on normality.
   2. Bartlett's test to check on the homogeneity of variance

## What happens if an assumption is violated

There are two options if the assumption is violated

1. perform transformation such as logarithm to turn data into normal distributed.
2. perform a non-parametric test, i.e., Kruskal-Wallis test.

# Descriptive Analysis – Parkinson

Parkinson's disease is a kind of nervous system disorder. It starts from a barely noticeable tremor but gradually becomes stiffness or slowing of movement. Tremors are common, but the disorder also commonly causes stiffness or slowing of movement.

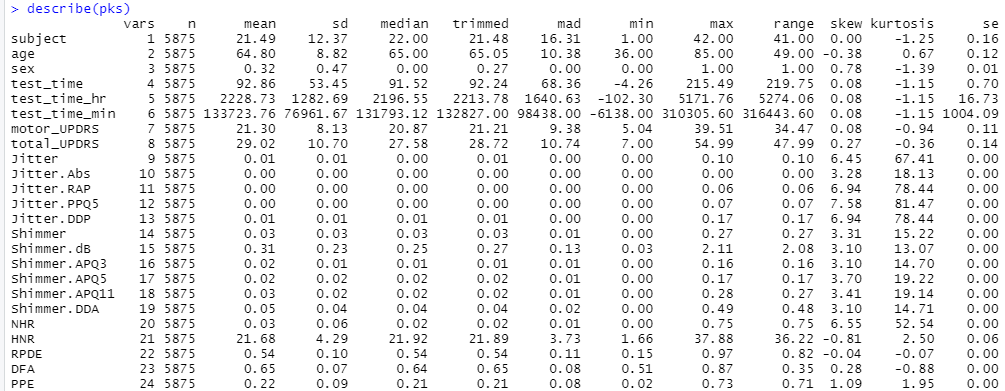
This report analyses the given data set from Parkinson's Disease and discusses the topic of ANOVA, such as whether we should remove or keep, why, and the consequences.

## Clean Data

This assignment will reuse the clean dataset from d2l. The given dataset has 5,875 observations, a collection of biomedical voice data taken over a period from 42 individuals with Parkinson's disease.



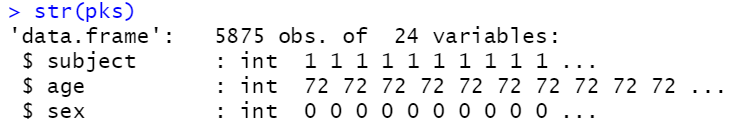
By looking into the dataset, there are no missing records. Thus, we are moving on to the ANOVA session.



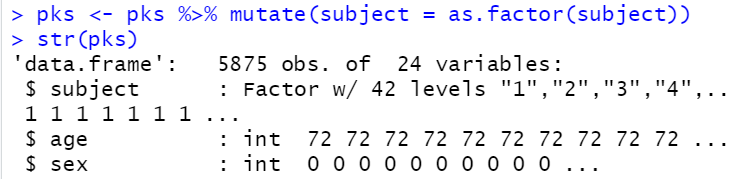
# ANOVA Analysis on total\_UPDRS and subject

Let's check on the original dataset, then apply the factor type conversion on a subject variable.

**Before conversion**



**After conversion**



## Distribution Check

The histogram shows that the population is almost normally distributed, and Q\_Q Plot can cross-check the normality.

|  |  |
| --- | --- |
| Histogram | Q-Q Plot |
|  |  |

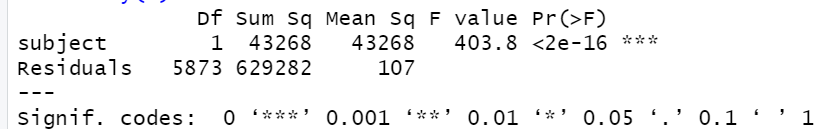
## Hypothesis

Null Hypothesis (H0): The mean values for 'total\_UPDRS' among subjects are equal

Alternative Hypothesis (Ha): Not all the mean values for 'total\_UPDRS' among subjects are equal

## ANOVA Test

The null hypothesis indicates the population means are all equal. In this case, the p-value (2e-16) of the ANVOA test is less than the 0.05 significant level. We reject the null hypothesis and state that not all the mean values for 'total\_UPDRS' among subjects are equal.



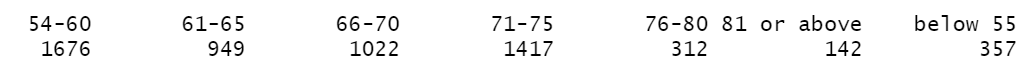
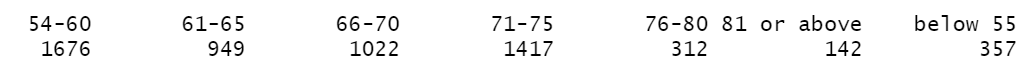
# ANOVA Analysis on total\_UPDRS and Age Group

## Convert Age into age group

We examine if there are differences in the mean value of total\_UPDRS among ages. The observations are distributed below. As you can see, the earliest group from 36 and not much appearance before age 55, therefore I group the data below 55, 55 – 60, 61 – 65, 66 – 70, 71 – 75, 76 – 80, and 80 plus group.



Data after grouping



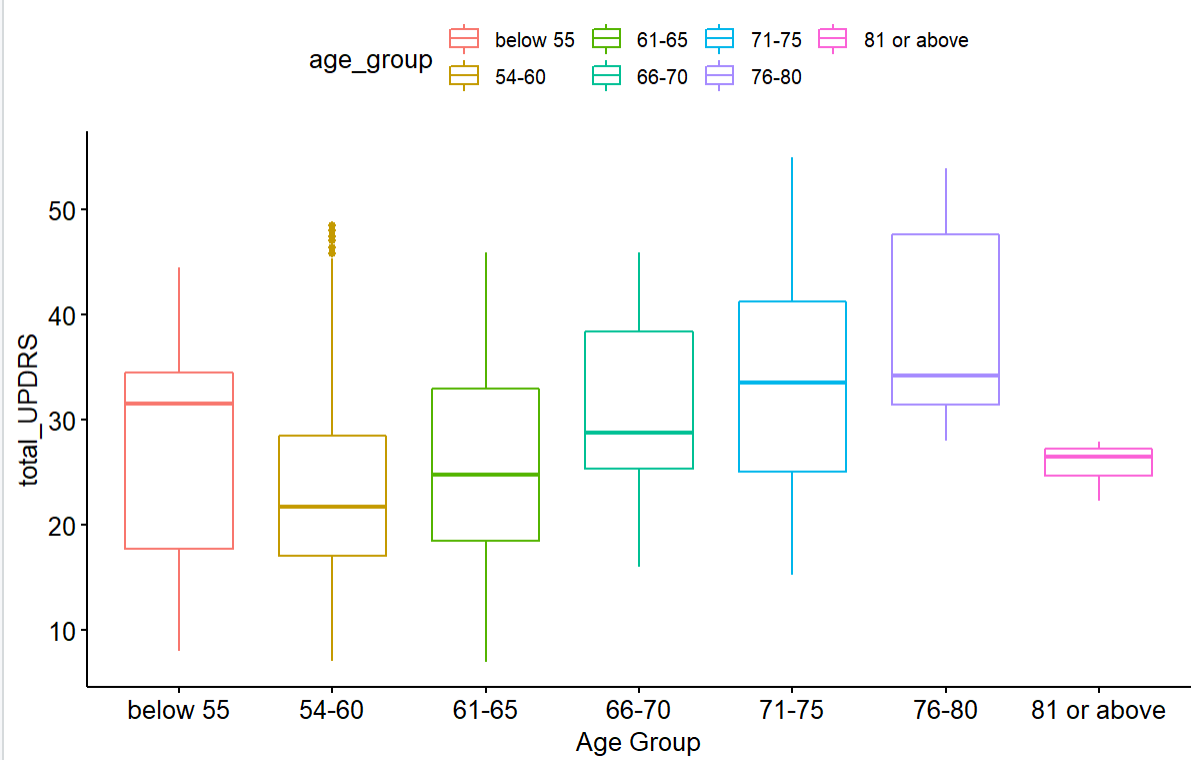
Cross-checking on age\_group variable to ensure it is a factor before running the ANOVA test.



## Boxplot on Age Group and total\_UPDRS

The boxplot shows the age group from binning the age column, and all the mean values are different age group is not the same.

* The noticeable mean difference is below 55 groups with a mean value at 31 and 55-60 groups with a mean value at 23.
* Some outliners in the age group of 54-60.
* Some age group, for instance, 61-65 and 71-75, has a more extended variance range than others, i.e., 66-70.
* The comprehensive range of all ages among on total\_UPDRS is not the same.



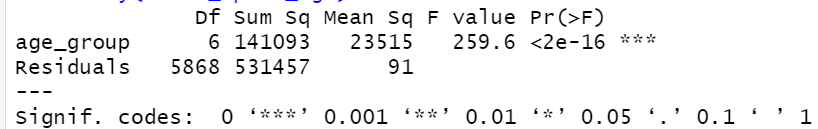
## Hypothesis

Null Hypothesis (H0): The mean values for 'total\_UPDRS' among the ages group are equal

Alternative Hypothesis (Ha): The mean values of at least one age group among 'total\_UPDRS' is different

## ANOVA Test – age group

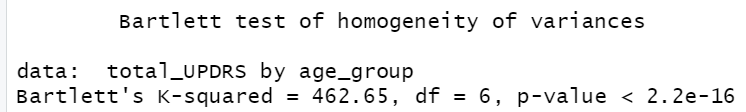
We have seven age groups, degree of free equal to six (n-1), and the null hypothesis indicates the population means are all equal. In this case, the p-value (2e-16) of the ANVOA test is less than the 0.05 significant level. We reject the null hypothesis and state that not all age\_group mean values for 'total\_UPDRS' are equal.



## Homogeneity Test – Bartlett test

Null hypothesis (H0): Variances are homogeneous

The result of Bartlett's homogeneity test below shows that the p-value (2.2e-16) is lower than the significant level of 0.05, and we have enough evidence to reject the null hypothesis. Thus, the age group variance among total\_UPDRS is not homogeneous.



## Normality test

Check the normality on Age with density plot and gg-plot, and both graphs show the population of age variable is not normally distributed.

|  |  |
| --- | --- |
| Density plot on Total\_UPDRS | Density plot on Age |
|  |  |

Appendix – R Code

rm(list=ls())

library(readr)

#-------------- Read file --------------------

pks <- read.table(file.choose(" "), header=T, sep=",")

str(pks)

#-------------- Factor on subject --------------------

library("dplyr")

pks <- pks %>% mutate(subject = as.factor(subject))

str(pks)

#-------------- Histogram --------------------

hist(pks$total\_UPDRS,main= "Histogram on total\_UPDRS")

#-------------- Check QQ Plot and line --------------------

#install.packages('qqplotr')

library('qqplotr')

#qqnorm(): produces a normal QQ plot of the variable

#qqline(): adds a reference line

qqnorm(pks$total\_UPDRS, pch = 1, frame = FALSE)

qqline(pks$total\_UPDRS, col = "steelblue", lwd = 2)

mean(pks$total\_UPDRS)

mean(pks$subject)

#-------------- anova test, total\_UPDRS and subject --------------------

anova\_updrs\_subject <- aov(total\_UPDRS~subject, data=pks)

summary(anova\_updrs\_subject)

#-------------- Age group--------------------

table(pks$age)

pks$age\_group <- pks$age

pks$age\_group <- ifelse((pks$age>=0 & pks$age<=54) , 'below 55',pks$age\_group)

pks$age\_group <- ifelse((pks$age>54 & pks$age<=60) , '55-60',pks$age\_group)

pks$age\_group <- ifelse((pks$age>60 & pks$age<=65) , '61-65',pks$age\_group)

pks$age\_group <- ifelse((pks$age>65 & pks$age<=70) , '66-70',pks$age\_group)

pks$age\_group <- ifelse((pks$age>70 & pks$age<=75) , '71-75',pks$age\_group)

pks$age\_group <- ifelse((pks$age>75 & pks$age<=80) , '76-80',pks$age\_group)

pks$age\_group <- ifelse((pks$age>80) , '81 or above',pks$age\_group)

pks$age\_group<-as.factor(pks$age\_group)

table(pks$age\_group)

table(pks$age\_group)

library("ggpubr")

library("ggplot2")

ggboxplot(pks,x="age\_group",

y="total\_UPDRS",

color="age\_group",

palette ="Okabe-Ito",

order=c("below 55","54-60","61-65","66-70", "71-75","76-80","81 or above"),

ylab="total\_UPDRS",xlab="Age Group")

str(pks$age\_group)

#-------------- ANOVA test --------------------

anova\_updrs\_age <- aov(total\_UPDRS~age\_group, data=pks)

summary(anova\_updrs\_age)

#-------------- Homogeneity test --------------------

bartlett.test(total\_UPDRS~age\_group, data=pks)

##

## Bartlett's Homogeneity Test (alpha = 0.05)

## -----------------------------------------------

## data : Sepal.Length and Species

##

## statistic : 462.65

## parameter : 6

## p.value : 2.2e-16

##

## Result : Variances are not homogeneous.

## -----------------------------------------------

######## Nomarity test ################

library(nortest)

shapiro.test(pks$total\_UPDRS ~pks$age\_group)

library("ggpubr")

ggdensity(pks$age,

main = "Density plot of Age",

xlab = "Age")

library(ggpubr)

ggqqplot(pks$age)