23-24-SIM-Midterm Template

Lídia Montero

November, 3rd 2023

Contents

1 Class attendance Data

Load attendraw.RData file in your current R or RStudio session. Selected numeric target is stndfnl (standardized final exam score) and let freshman status be the qualitative target. Prepare your dataset to represent factors in a suitable way.

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
## Loading required package: rpart
```

```
## Loading required package: sgeostat
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following object is masked from 'package:car':
##
##     qqPlot
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## null device
##           1
```

2 Load dataset and define factors

```
##      attend         termgpa          priGPA           ACT       
##  Min.   : 2.00   Min.   :0.000   Min.   :0.857   Min.   :13.00  
##  1st Qu.:24.00   1st Qu.:2.138   1st Qu.:2.190   1st Qu.:20.00  
##  Median :28.00   Median :2.670   Median :2.560   Median :22.00  
##  Mean   :26.15   Mean   :2.601   Mean   :2.587   Mean   :22.51  
##  3rd Qu.:30.00   3rd Qu.:3.120   3rd Qu.:2.942   3rd Qu.:25.00  
##  Max.   :32.00   Max.   :4.000   Max.   :3.930   Max.   :32.00  
##                                                                 
##      final          atndrte           hwrte            frosh       
##  Min.   :10.00   Min.   :  6.25   Min.   : 12.50   Min.   :0.0000  
##  1st Qu.:22.00   1st Qu.: 75.00   1st Qu.: 87.50   1st Qu.:0.0000  
##  Median :26.00   Median : 87.50   Median :100.00   Median :0.0000  
##  Mean   :25.89   Mean   : 81.71   Mean   : 87.91   Mean   :0.2324  
##  3rd Qu.:29.00   3rd Qu.: 93.75   3rd Qu.:100.00   3rd Qu.:0.0000  
##  Max.   :39.00   Max.   :100.00   Max.   :100.00   Max.   :1.0000  
##                                   NA's   :6                        
##       soph           skipped          stndfnl        
##  Min.   :0.0000   Min.   : 0.000   Min.   :-3.30882  
##  1st Qu.:0.0000   1st Qu.: 2.000   1st Qu.:-0.78782  
##  Median :1.0000   Median : 4.000   Median : 0.05252  
##  Mean   :0.5765   Mean   : 5.853   Mean   : 0.02966  
##  3rd Qu.:1.0000   3rd Qu.: 8.000   3rd Qu.: 0.68277  
##  Max.   :1.0000   Max.   :30.000   Max.   : 2.78361  
## 
```

All questions account for 1 point (you have to answer all of them)

1. Determine thresholds for mild and severe outliers in the target. Are there any outliers? Indicate observation id's and atypical values. *Do not take any action.*

Target is stndfnl variable. Using summary we obtain Q1, Q3 and then IQR. Lower and upper bounds for mild and severe outliers are figured out. Two mild lower outliers are seen, but no severe outlier are found.

```
varsumm <- summary( df$stndfnl ); varsumm
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -3.30882 -0.78782  0.05252  0.02966  0.68277  2.78361 
```

```
iqr <- varsumm[5]-varsumm[2]; iqr
```

```
##  3rd Qu. 
## 1.470588 
```

```
lmout <- varsumm[2] - 1.5*iqr
umout <- varsumm[5] + 1.5*iqr
lsout <- varsumm[2] - 3*iqr
usout <- varsumm[5] + 3*iqr

lmout;umout
```

```
##   1st Qu. 
## -2.993698 
##  3rd Qu. 
## 2.888656
```

```
lsout;usout
```

```
##  1st Qu. 
## -5.19958
```

```
##  3rd Qu. 
## 5.094538
```

```
llmild <- which( (df$stndfnl<lmout)|(df$stndfnl>umout) );llmild
```

```
## [1]   7 502
```

```
llsev <- which( (df$stndfnl<lsout)|(df$stndfnl>usout) );llsev
```

```
## integer(0)
```

2. Use an imputation method to address missing data in the dataset. Validate imputation results.

Missing data is found in hwrte variable. You can use either imputePCA() in missMDA package or mice method. Both methods return valid imputed values. Deciles are not affected, thus we would be allowed to retain imputed values.

```
summary(df)
```

```
##      attend         termgpa          priGPA           ACT       
##  Min.   : 2.00   Min.   :0.000   Min.   :0.857   Min.   :13.00  
##  1st Qu.:24.00   1st Qu.:2.138   1st Qu.:2.190   1st Qu.:20.00  
##  Median :28.00   Median :2.670   Median :2.560   Median :22.00  
##  Mean   :26.15   Mean   :2.601   Mean   :2.587   Mean   :22.51  
##  3rd Qu.:30.00   3rd Qu.:3.120   3rd Qu.:2.942   3rd Qu.:25.00  
##  Max.   :32.00   Max.   :4.000   Max.   :3.930   Max.   :32.00  
##                                                                 
##      final          atndrte           hwrte                frosh    
##  Min.   :10.00   Min.   :  6.25   Min.   : 12.50   Freshman-No :522  
##  1st Qu.:22.00   1st Qu.: 75.00   1st Qu.: 87.50   Freshman-Yes:158  
##  Median :26.00   Median : 87.50   Median :100.00                     
##  Mean   :25.89   Mean   : 81.71   Mean   : 87.91                     
##  3rd Qu.:29.00   3rd Qu.: 93.75   3rd Qu.:100.00                     
##  Max.   :39.00   Max.   :100.00   Max.   :100.00                     
##                                   NA's   :6                          
##        soph        skipped          stndfnl               f.type   
##  Soph-No :288   Min.   : 0.000   Min.   :-3.30882   none     :130  
##  Soph-Yes:392   1st Qu.: 2.000   1st Qu.:-0.78782   freshman :158  
##                 Median : 4.000   Median : 0.05252   sophomore:392  
##                 Mean   : 5.853   Mean   : 0.02966                  
##                 3rd Qu.: 8.000   3rd Qu.: 0.68277                  
##                 Max.   :30.000   Max.   : 2.78361                  
## 
```

```
llmiss <- which( is.na(df$hwrte) )
df[ llmiss, ]
```

```
##     attend termgpa priGPA ACT final atndrte hwrte        frosh     soph skipped
## 50      10   1.210   2.05  30    29  31.250    NA Freshman-Yes  Soph-No      22
## 185     16   1.810   1.98  23    25  50.000    NA Freshman-Yes  Soph-No      16
## 326     17   1.800   1.96  24    25  53.125    NA  Freshman-No Soph-Yes      15
## 474     12   0.333   2.47  26    23  37.500    NA  Freshman-No  Soph-No      20
## 511      9   1.350   1.52  27    30  28.125    NA Freshman-Yes  Soph-No      23
## 513      4   0.450   2.14  27    25  12.500    NA  Freshman-No  Soph-No      28
##        stndfnl    f.type
## 50   0.6827731  freshman
## 185 -0.1575630  freshman
## 326 -0.1575630 sophomore
## 474 -0.5777311      none
## 511  0.8928571  freshman
## 513 -0.1575630      none
```

```
library(missMDA)
res.imppca <- imputePCA( df[,c(1:7,10:11)])
summary(res.imppca$completeObs[,"hwrte"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.50   87.50  100.00   87.57  100.00  100.00
```

```
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
res.mice <- mice(df)
```

```
##
##  iter imp variable
##   1   1  hwrte*
##   1   2  hwrte*
##   1   3  hwrte*
##   1   4  hwrte*
##   1   5  hwrte*
##   2   1  hwrte*
##   2   2  hwrte*
##   2   3  hwrte*
##   2   4  hwrte*
##   2   5  hwrte*
##   3   1  hwrte*
##   3   2  hwrte*
##   3   3  hwrte*
##   3   4  hwrte*
##   3   5  hwrte*
##   4   1  hwrte*
##   4   2  hwrte*
##   4   3  hwrte*
##   4   4  hwrte*
##   4   5  hwrte*
##   5   1  hwrte*
##   5   2  hwrte*
##   5   3  hwrte*
##   5   4  hwrte*
##   5   5  hwrte*
```

```
## Warning: Number of logged events: 53
```

```
dfimp <-complete(res.mice)
summary(dfimp)
```

```
##      attend         termgpa          priGPA            ACT       
##  Min.   : 2.00   Min.   :0.000   Min.   :0.857   Min.   :13.00  
##  1st Qu.:24.00   1st Qu.:2.138   1st Qu.:2.190   1st Qu.:20.00  
##  Median :28.00   Median :2.670   Median :2.560   Median :22.00  
##  Mean   :26.15   Mean   :2.601   Mean   :2.587   Mean   :22.51  
##  3rd Qu.:30.00   3rd Qu.:3.120   3rd Qu.:2.942   3rd Qu.:25.00  
##  Max.   :32.00   Max.   :4.000   Max.   :3.930   Max.   :32.00  
##      final          atndrte            hwrte                  frosh    
##  Min.   :10.00   Min.   :  6.25   Min.   : 12.50   Freshman-No :522  
##  1st Qu.:22.00   1st Qu.: 75.00   1st Qu.: 87.50   Freshman-Yes:158  
##  Median :26.00   Median : 87.50   Median :100.00                     
##  Mean   :25.89   Mean   : 81.71   Mean   : 87.44                     
##  3rd Qu.:29.00   3rd Qu.: 93.75   3rd Qu.:100.00                     
##  Max.   :39.00   Max.   :100.00   Max.   :100.00                     
##        soph         skipped           stndfnl                f.type   
##  Soph-No :288   Min.   : 0.000   Min.   :-3.30882   none     :130  
##  Soph-Yes:392   1st Qu.: 2.000   1st Qu.:-0.78782   freshman :158  
##                 Median : 4.000   Median : 0.05252   sophomore:392  
##                 Mean   : 5.853   Mean   : 0.02966                  
##                 3rd Qu.: 8.000   3rd Qu.: 0.68277                  
##                 Max.   :30.000   Max.   : 2.78361                  
```

```
quantile( df$hwrte, probs=seq(0,1,by=0.1), na.rm=T)
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
##  12.5  62.5  75.0  87.5  87.5 100.0 100.0 100.0 100.0 100.0 100.0 
```

```
quantile( dfimp$hwrte, probs=seq(0,1,by=0.1), na.rm=F)
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
##  12.5  62.5  75.0  87.5  87.5 100.0 100.0 100.0 100.0 100.0 100.0
```

```
quantile( res.imppca$completeObs[,"hwrte"], probs=seq(0,1,by=0.1), na.rm=F)
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100% 
##  12.5  62.5  75.0  87.5  87.5 100.0 100.0 100.0 100.0 100.0 100.0
```

```
dfimp[ llmiss, ]
```

```
##     attend termgpa priGPA ACT final atndrte hwrte        frosh     soph skipped
## 50      10   1.210   2.05  30    29  31.250  37.5 Freshman-Yes  Soph-No      22
## 185     16   1.810   1.98  23    25  50.000  50.0 Freshman-Yes  Soph-No      16
## 326     17   1.800   1.96  24    25  53.125  12.5  Freshman-No Soph-Yes      15
## 474     12   0.333   2.47  26    23  37.500  50.0  Freshman-No  Soph-No      20
## 511      9   1.350   1.52  27    30  28.125  50.0 Freshman-Yes  Soph-No      23
## 513      4   0.450   2.14  27    25  12.500  12.5  Freshman-No  Soph-No      28
##        stndfnl    f.type
## 50   0.6827731  freshman
## 185 -0.1575630  freshman
## 326 -0.1575630 sophomore
## 474 -0.5777311      none
## 511  0.8928571  freshman
## 513 -0.1575630      none
```

```
df$hwrte <- dfimp$hwrte
```

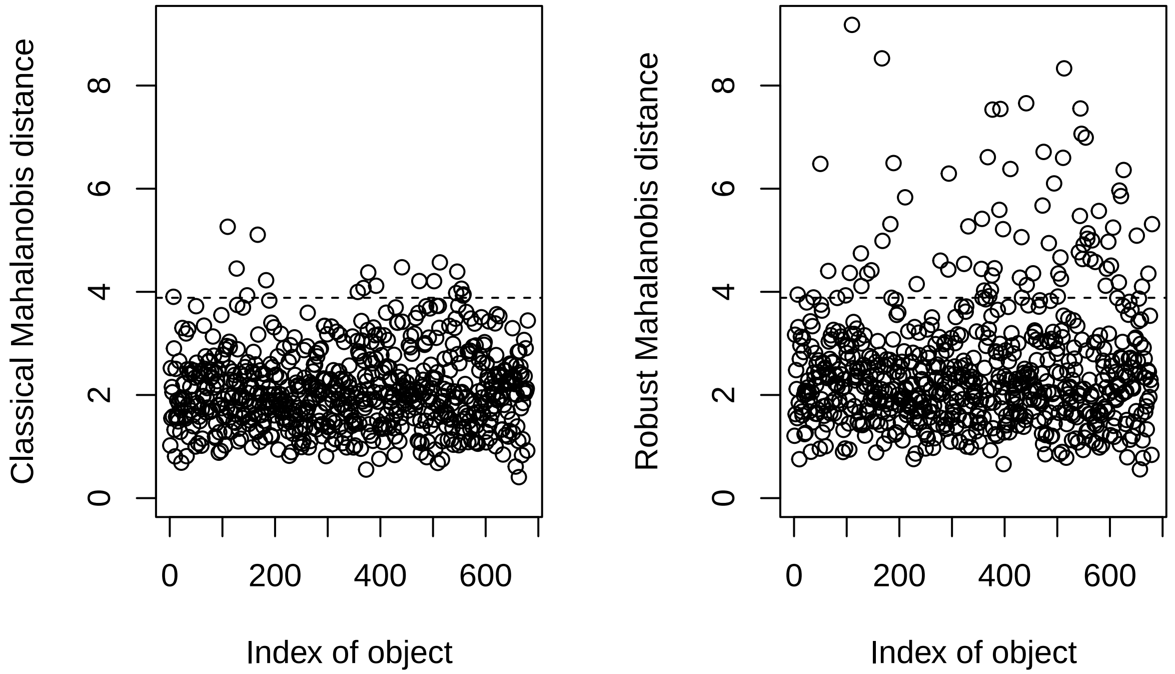
3. Are there multivariate outliers in the dataset? Indicate how many and which at 99% confidence. Explain what they seem to share in common. Do not take any action.

Library mvoutlier has to be upload in the workspace and method Moutlier executed using quantile parameter set at 0.99. A subset of the numeric variables has to be define to allow proper completion of the method. There are 19 multivariant outliers. Profiling of the binary factor indicating the status of multivariant outliers is addressed. Variables skipped, attend and atndrte are the most globally associated to multivariant outliers (results of $quanti.var list). In particular, multivariant outliers show a significant skipped and ACT mean over the overall mean and significant mean less than the overall mean for termgpa, hwrte, atndrte and attend.

```
library(mvoutlier)
names(df)
```

```
##  [1] "attend"  "termgpa" "priGPA"  "ACT"     "final"   "atndrte" "hwrte"  
##  [8] "frosh"   "soph"    "skipped" "stndfnl" "f.type"
```

```
res.mvout <- Moutlier(df[, c(2,3,4,5,6)], quantile = 0.99)
```



```
plot(res.mvout$md, res.mvout$rd)
abline(h=res.mvout$cutoff, col="red", lty = 2)
abline(v=res.mvout$cutoff, col="red", lty = 2)

llmout <- which( ( res.mvout$md > res.mvout$cutoff) & ( res.mvout$rd > res.mvout$cutoff));length(llmout
```

```
## [1] 19
```

```
df$mvout <- 0
df$mvout[ llmout ] <-1
df$mvout <- factor( df$mvout, labels=c("Mvout-No","Mvout-Yes") )

library(FactoMineR)
res.cat <- catdes(df,13)
res.cat$test.chi2
```

```
##      p.value df
```

```
res.cat$quanti.var
```

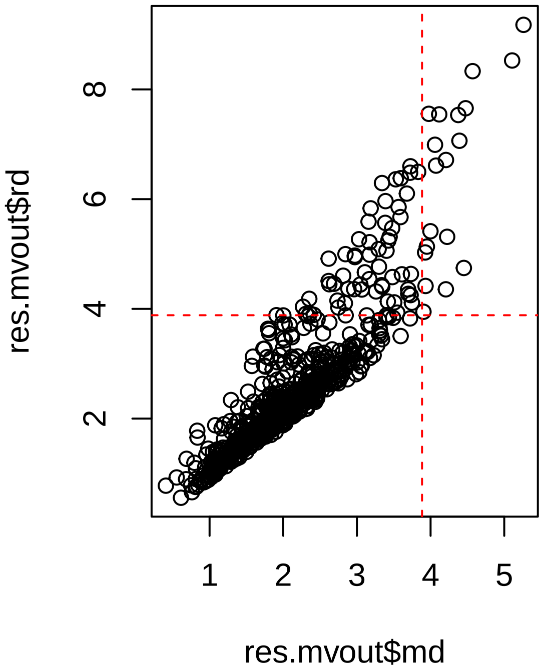
```
##               Eta2      P-value
## skipped 0.14898914 1.393199e-25
## attend  0.14898914 1.393199e-25
## atndrte 0.14898914 1.393199e-25
## hwrte   0.10227383 1.234878e-17
## termgpa 0.06272110 3.482764e-11
## priGPA  0.03518276 8.391618e-07
## ACT     0.01010938 8.697220e-03
```

```
res.cat$quanti
```

```
## $‘Mvout-No‘
##            v.test Mean in category Overall mean sd in category Overall sd
## atndrte 10.058013        82.824319    81.709559     15.1381414 17.0344515
## attend  10.058013        26.503782    26.147059      4.8442053  5.4510245
## hwrte    8.333302        88.521180    87.444853     18.4855180 19.8511624
## termgpa  6.525920         2.632253     2.601000      0.7026740  0.7360442
## priGPA   4.887647         2.604085     2.586775      0.5351007  0.5443134
## ACT     -2.619976        22.450832    22.510294      3.4587808  3.4882003
## skipped -10.058013        5.496218     5.852941      4.8442053  5.4510245
##              p.value
## atndrte 8.469038e-24
## attend  8.469038e-24
## hwrte   7.861799e-17
## termgpa 6.758547e-11
## priGPA  1.020484e-06
## ACT     8.793606e-03
## skipped 8.469038e-24
##
## $‘Mvout-Yes‘
##            v.test Mean in category Overall mean sd in category Overall sd
## skipped 10.058013        18.263158     5.852941      9.4135519  5.4510245
## ACT      2.619976        24.578947    22.510294      3.8568570  3.4882003
## priGPA  -4.887647         1.984579     2.586775      0.5188160  0.5443134
## termgpa -6.525920         1.513737     2.601000      0.9979461  0.7360442
## hwrte   -8.333302        50.000000    87.444853     27.8033508 19.8511624
## atndrte -10.058013       42.927632    81.709559     29.4173495 17.0344515
## attend  -10.058013       13.736842    26.147059      9.4135519  5.4510245
##              p.value
## skipped 8.469038e-24
## ACT     8.793606e-03
## priGPA  1.020484e-06
## termgpa 6.758547e-11
## hwrte   7.861799e-17
## atndrte 8.469038e-24
## attend  8.469038e-24
```

```
tapply(df$stndfnl,df$mvout,summary)
```

```
## $‘Mvout-No‘
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.67857 -0.78782  0.05252  0.03599  0.68277  2.78361
##
## $‘Mvout-Yes‘
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.3088 -0.5777 -0.1576 -0.1907  0.5777  1.9433
```



4. Analyze the profile of the numeric target (stndfnl) using a suitable profiling method. A detailed explanation of the procedure outcome is requested.

Condes method in FactoMineR package has to be applied. Perfect correlation is found for final (stndfn is just a linear transformation of final) and intense direct correlation is seen for termgpa, priGPA and ACT. A weak inverse correlation is found for skipped variable.

Freshman factor is globally related to the standard final mark (stndfnl). Freshman student (new students) stndfnl mark mean is less than the grand mean by 0.155 units.

```
#library(FactoMineR)

res.con <- condes(df, which(names(df)=="stndfnl"))
res.con$quanti
```

```
##         correlation      p.value
## final     1.0000000 0.000000e+00
## termgpa   0.5106093 2.005168e-46
## priGPA    0.3659273 5.673134e-23
## ACT       0.3612486 2.170214e-22
## hwrte     0.1408134 2.299832e-04
## attend    0.1400327 2.493270e-04
## atndrte   0.1400327 2.493270e-04
## skipped  -0.1400327 2.493270e-04
```

```
res.con$quali
```

```
##                  R2    p.value
## frosh  0.007856995 0.02079203
## f.type 0.009315527 0.04208410
```

```
res.con$category
```

```
##                       Estimate    p.value
## frosh=Freshman-No    0.1037582 0.02079203
## f.type=freshman     -0.1550176 0.02079203
## frosh=Freshman-Yes -0.1037582 0.02079203
```

5. Analyze the profile of the binary target (frosh) using a suitable method. A detailed explanation of the procedure outcome is requested.

Catdes method in FactoMineR package has to be applied. Frosh factor is globally related to soph factor according to $test.chi2 output list. Frosh factor is globally related to numeric variables priGPA, ACT, termgpa according to $quanti.var output list. Analyzing $category output list we see what it is obvious taking into account the definition of soph factor: a freshman (first year) can not be a sophomore student (second year). Accounting for $quanti output list, positive freshman observations show a significant lower mean on stndfnl, final, hwrte, termgpa, ACT and priGPA variables. Freshman students are attending to classes as the rest of students (attendance is not higher).

```
res.cat <- catdes(df, which(names(df)=="frosh"))
res.cat$quanti.var
```

```
##                Eta2      P-value
## priGPA  0.095156148 1.860098e-16
## ACT     0.022309326 9.241141e-05
## termgpa 0.017101529 6.295462e-04
## hwrte   0.008603233 1.554246e-02
## final   0.007856995 2.079203e-02
## stndfnl 0.007856995 2.079203e-02
```

```
res.cat$quanti
```

```
## $`Freshman-No`
##           v.test Mean in category Overall mean sd in category Overall sd
## priGPA  8.038098       2.67915134   2.58677500      0.5303443  0.5443134
## ACT     3.892047      22.79693487  22.51029412      3.4681018  3.4882003
## termgpa 3.407629       2.65395594   2.60100001      0.7435499  0.7360442
## hwrte   2.416939      88.45785441  87.44485294     19.1483457 19.8511624
## final   2.309740      26.12068966  25.89117647      4.8016992  4.7063704
## stndfnl 2.309740       0.07787598   0.02965892      1.0087603  0.9887333
##              p.value
## priGPA  9.124341e-16
## ACT     9.940185e-05
## termgpa 6.552982e-04
## hwrte   1.565163e-02
## final   2.090254e-02
## stndfnl 2.090254e-02
##
## $‘Freshman-Yes‘
##            v.test Mean in category Overall mean sd in category Overall sd
## stndfnl -2.309740       -0.1296405   0.02965892      0.9013165  0.9887333
## final   -2.309740       25.1329114  25.89117647      4.2902663  4.7063704
## hwrte   -2.416939       84.0981013  87.44485294     21.6802467 19.8511624
## termgpa -3.407629        2.4260443   2.60100001      0.6820532  0.7360442
## ACT     -3.892047       21.5632911  22.51029412      3.3854371  3.4882003
## priGPA  -8.038098        2.2815823   2.58677500      0.4738551  0.5443134
##              p.value
## stndfnl 2.090254e-02
## final   2.090254e-02
## hwrte   1.565163e-02
## termgpa 6.552982e-04
## ACT     9.940185e-05
## priGPA  9.124341e-16
```

```
res.cat$test.chi2
```

```
##              p.value df
## f.type 2.187138e-148  2
## soph    6.968296e-63  1
```

```
res.cat$category
```

```
## $‘Freshman-No‘
##                   Cla/Mod  Mod/Cla   Global       p.value     v.test
## f.type=sophomore 100.00000 75.09579 57.64706  1.335092e-74  18.273919
## soph=Soph-Yes    100.00000 75.09579 57.64706  1.335092e-74  18.273919
## f.type=none      100.00000 24.90421 19.11765  1.463496e-17   8.529999
## soph=Soph-No      45.13889 24.90421 42.35294  1.335092e-74 -18.273919
## f.type=freshman    0.00000  0.00000 23.23529 2.223341e-159 -26.899544
##
## $‘Freshman-Yes‘
##                   Cla/Mod Mod/Cla   Global       p.value     v.test
## f.type=freshman  100.00000     100 23.23529 2.223341e-159  26.899544
## soph=Soph-No      54.86111     100 42.35294  1.335092e-74  18.273919
## f.type=none        0.00000       0 19.11765  1.463496e-17  -8.529999
## f.type=sophomore   0.00000       0 57.64706  1.335092e-74 -18.273919
## soph=Soph-Yes      0.00000       0 57.64706  1.335092e-74 -18.273919
```

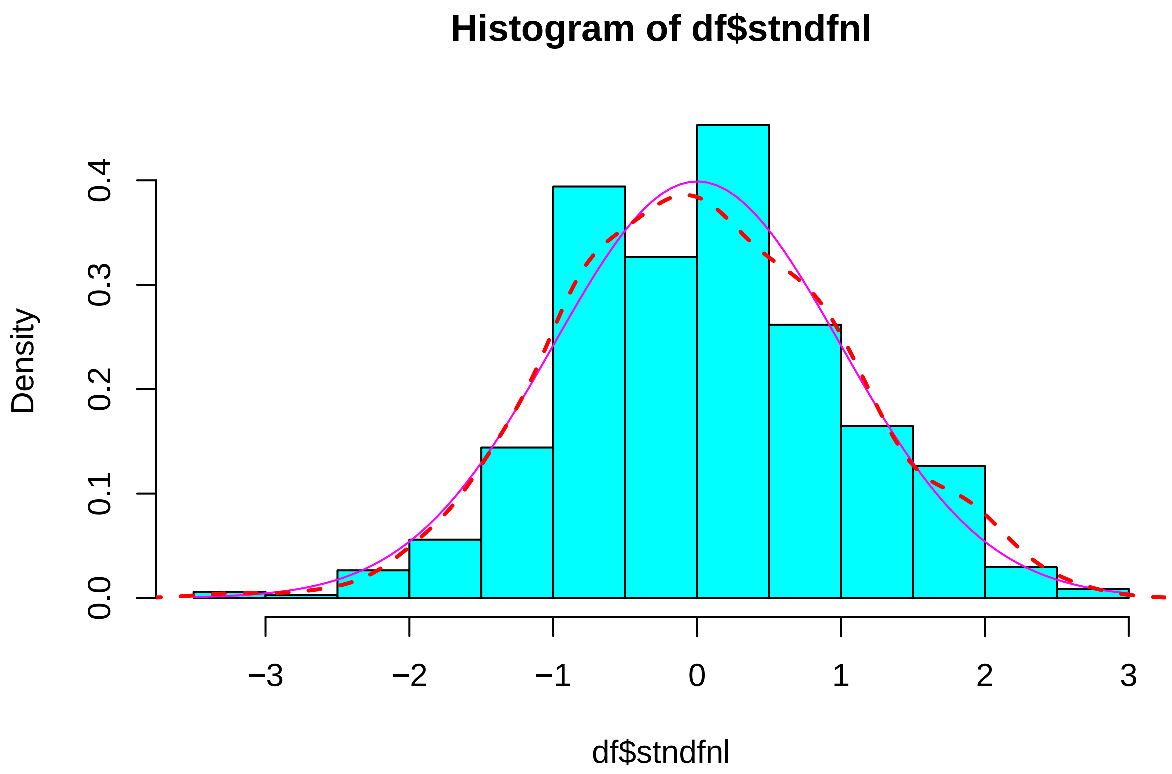
6. Is there variance homogeneity in the stndfnl target groups defined by frosh and soph classes, one by one and when applied simultaneously? Hint: You have to define a new factor.

Now you have to define a new factor that combines freshman and sophomore status in a polytomous factor. Students are in their first or second or neither first nor second year. A contingency table helps to undertand how to define the new factor.

Normal distribution of stndfnl target is rejected by Shapiro-Wilk test (or by graphic assessment using histogram and overlapping a normal curve). Non-parametric test for variance homogeneity has to be used.

Null hypothesis of homogeneity of variances in group defined by frosh, or soph or polytomous f.type can not be rejected according to Fligner-Killeen tests.

```r
hist(df$stndfnl,20, col="cyan", freq=F)
curve(dnorm(x),add=T, col="magenta")
lines(density(df$stndfnl), col = "red", lwd=2, lty=2)
```



```r
shapiro.test( df$ stndfnl )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$stndfnl
## W = 0.99354, p-value = 0.005118
```

```r
fligner.test( df$stndfnl ~ df$frosh )
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$stndfnl by df$frosh
## Fligner-Killeen:med chi-squared = 2.4918, df = 1, p-value = 0.1144
```

```r
fligner.test( df$stndfnl ~ df$soph )
```

```
##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$stndfnl by df$soph
## Fligner-Killeen:med chi-squared = 0.95257, df = 1, p-value = 0.3291
```

```r
table(df$frosh, df$soph)  # You have to define f.type with 3 levels
```

```
##
##               Soph-No Soph-Yes
##   Freshman-No     130      392
##   Freshman-Yes    158        0
```

```r
fligner.test( df$stndfnl ~ df$f.type )
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$stndfnl by df$f.type
## Fligner-Killeen:med chi-squared = 2.5024, df = 2, p-value = 0.2862
```

7. Mean stndfnl target can be considered to be the equal across groups defined by frosh target? Use a two.sided test at 1% significance level and indicate the confidence interval for freshman target population mean. Indicate whether equal variances and normal distribution of stndfnl hypothesis hold in the population.

A non-parametric test for mean homogeneity of stdnfnl in groups defined frosh binary factor has to be used. Null hypothesis can not be rejected at any significance level. It is worth to note that the parametric t.test two sided hypothesis can not be rejected at 1% significance level, but it would be rejected at 5%. The same can be said of Wilcoxon test output, the best option in this case, given non normal distribution of the target. 99% two-sided confidence interval of stndfnl for fresman group lies between -0.317 to 0.058, it is obtained using t.test, the only possibility.

```r
fligner.test( df$stndfnl ~ df$frosh )
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$stndfnl by df$frosh
## Fligner-Killeen:med chi-squared = 2.4918, df = 1, p-value = 0.1144
```

```r
t.test( df$stndfnl ~ df$frosh, conf.level=0.99, mu=0, paired=F, equal.var=T)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$stndfnl by df$frosh
## t = 2.458, df = 285.63, p-value = 0.01456
## alternative hypothesis: true difference in means between group Freshman-No and group Freshman-Yes is
## 99 percent confidence interval:
##  -0.01140912  0.42644200
## sample estimates:
##  mean in group Freshman-No mean in group Freshman-Yes
##                 0.07787598                -0.12964046
```

```
wilcox.test( df$stndfnl ~ df$frosh, conf.level=0.99, mu=0, paired=F, equal.var=T) # Non-parametric shou
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df$stndfnl by df$frosh
## W = 46162, p-value = 0.02258
## alternative hypothesis: true location shift is not equal to 0
```

```
llf <- which( df$frosh == "Freshman-Yes"); length( llf )
```

```
## [1] 158
```

```
stnfrmean <- mean(df$stndfnl[llf])
t.test( df$stndfnl[llf] , mu=stnfrmean, conf.level=0.99)
```

```
##
##  One Sample t-test
##
## data:  df$stndfnl[llf]
## t = 0, df = 157, p-value = 1
## alternative hypothesis: true mean is not equal to -0.1296405
## 99 percent confidence interval:
##  -0.3172059  0.0579250
## sample estimates:
##  mean of x
## -0.1296405
```

```
t.test( df$stndfnl[llf] , mu=0, conf.level=0.99) # Same result in terms of CI
```

```
##
##  One Sample t-test
##
## data:  df$stndfnl[llf]
## t = -1.8022, df = 157, p-value = 0.07343
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -0.3172059  0.0579250
## sample estimates:
##  mean of x
## -0.1296405
```

8. State and test one.sided hypothesis to assess whether stndfnl is less for freshman than the rest at 1% significance level. Indicate and justify a 95% confidence interval for freshman target population mean.

Null hypothesis has to be stated as H0 mu_No = mu_Yes and H1: mu_No < mu_Yes, so alternative is set to "greater" according to the order of the levels in frosh factor. Using a parametric test H0 is rejected thus H1 is confirmed. In the non-parametric test H0 can not be rejected at the 1% significance level. A 95% CI for freshman stndfnl has a lower bound of -0.249 (and infinite upper bound) under t.test or -0.263 under the non-parametric Wilcoxon test.

You can figure out CI using formulas included in the theory slides.

```
#t.test( df$stndfnl ~ df$frosh, conf.level=0.99, mu=0, paired=F, equal.var=T, alternative = "greater")
wilcox.test( df$stndfnl ~ df$frosh, conf.level=0.99, mu=0, paired=F, equal.var=T, alternative = "greate
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df$stndfnl by df$frosh
## W = 46162, p-value = 0.01129
## alternative hypothesis: true location shift is greater than 0
## 99 percent confidence interval:
##  -5.501999e-05           Inf
## sample estimates:
## difference in location
##              0.2100477
```

```
llf <- which( df$frosh == "Freshman-Yes"); length( llf )
```

```
## [1] 158
```

```
stnfrmean <- mean(df$stndfnl[llf])
wilcox.test( df$stndfnl[llf] , mu=0, conf.level=0.95, alternative="greater",equal.var=T,conf.int=T) # p
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df$stndfnl[llf]
## V = 5260, p-value = 0.962
## alternative hypothesis: true location is greater than 0
## 95 percent confidence interval:
##  -0.262567       Inf
## sample estimates:
## (pseudo)median
##     -0.1575428
```

```
t.test( df$stndfnl[llf] , mu=0, conf.level=0.95, alternative="greater",equal.var=T)
```

```
##
##  One Sample t-test
##
## data:  df$stndfnl[llf]
## t = -1.8022, df = 157, p-value = 0.9633
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2486618        Inf
## sample estimates:
##  mean of x
## -0.1296405
```

9. The standard deviation of stndfnl in the freshman group should not exceed 1. For the sample in the freshman group in your dataset, calculate the deviation of stndfnl assuming that normal

assumption holds. State and include any assumptions needed to test at the 1% significance level whether population standard deviation is larger than 1 in the freshman group. Figure out the 99% upper threshold for stndfnl in the freshman population standard deviation.

This question is about variance in the freshman group. We select the subset of observations belonging to the freshman group and state the null hypothesis of true variance equal 1 vs the alternative hypothesis of being greater than one at 1% significance level. P value indicates that H0 can not be rejected and lower bound to variance is 0.638, thus lower bound to standard deviation in freshman group should be 0.799. Using alternative hypothesis variance less than 1, H0 can not be rejected at 1% significance level and upper bound to standard deviation can be seen at 1.04.

```r
ll <- which(df$frosh=="Freshman-Yes")
var(df$stndfnl[ll])
```

```
## [1] 0.8175457
```

```r
var(df$stndfnl[-ll])
```

```
## [1] 1.019551
```

```r
sd(df$stndfnl[ll])
```

```
## [1] 0.9041823
```

```r
sd(df$stndfnl[-ll])
```

```
## [1] 1.009728
```

```r
varTest(df$stndfnl[ll], sigma.squared=1,alternative="less",conf.level = 0.99)
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 1
##
## Alternative Hypothesis:          True variance is less than 1
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 0.8175457
##
## Data:                            df$stndfnl[ll]
##
## Test Statistic:                  Chi-Squared = 128.3547
##
## Test Statistic Parameter:        df = 157
##
## P-value:                         0.04564503
##
## 99% Confidence Interval:         LCL = 0.000000
##                                  UCL = 1.080993
```

```
varTest(df$stndfnl[ll], sigma.squared=1,alternative="greater",conf.level = 0.99)
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 1
##
## Alternative Hypothesis:          True variance is greater than 1
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 0.8175457
##
## Data:                            df$stndfnl[ll]
##
## Test Statistic:                  Chi-Squared = 128.3547
##
## Test Statistic Parameter:        df = 157
##
## P-value:                         0.954355
##
## 99% Confidence Interval:         LCL = 0.6381429
##                                  UCL =       Inf
```

```
sqrt(0.6381429) # LCL refers to variance
```

```
## [1] 0.7988385
```

```
sqrt(1.080993)
```

```
## [1] 1.039708
```

10. Determine a 99% confidence interval for the population proportion of neither freshman, nor sophomore. Test the null hypothesis that selecting a unit neither freshman, nor sophomore has a population probability greater than 0.25.

We have the group not freshmand and not sophomore. Thus the test has to be done based on the proportion for this group (None level in f.type definition). The population proportion of None students is clearly less than 25%, according to the one-sided less test.

```
prop.table(table(df$f.type))
```

```
##
##      none  freshman sophomore
## 0.1911765 0.2323529 0.5764706
```

```
ll <- which( df$f.type =="none"); length( ll )
```

```
## [1] 130
```

```
prop.test(x=130, n=680, conf.level=0.99, correct=F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  130 out of 680, null probability 0.5
## X-squared = 259.41, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.1553913 0.2329300
## sample estimates:
##         p
## 0.1911765
```

```
prop.test(x=130, n=680, p=0.25,conf.level=0.99, correct=F, alternative ="greater")
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  130 out of 680, null probability 0.25
## X-squared = 12.549, df = 1, p-value = 0.9998
## alternative hypothesis: true p is greater than 0.25
## 99 percent confidence interval:
##  0.1585883 1.0000000
## sample estimates:
##         p
## 0.1911765
```

```
prop.test(x=130, n=680, p=0.25,conf.level=0.99, correct=F, alternative ="less")  # In fact is less than
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  130 out of 680, null probability 0.25
## X-squared = 12.549, df = 1, p-value = 0.0001982
## alternative hypothesis: true p is less than 0.25
## 99 percent confidence interval:
##  0.0000000 0.2286415
## sample estimates:
##         p
## 0.1911765
```

11. Test the null hypothesis that the proportion of freshman group and the population proportion of neither freshman, nor sophomore is the same at 1% significance level.

This is a two.sided test of equal proportions in the two populations (freshman and none) at 1% significance level. The null hypothesis of identical proportions can not be rejected at 1% significance level.

```
prop.table(table(df$f.type))
```

```
##
##      none  freshman sophomore
## 0.1911765 0.2323529 0.5764706
```

```
table(df$f.type)
```

```
##
##      none  freshman sophomore
##       130       158       392
```

```
prop.test(x=c(130, 158), n=c(680, 680), conf.level=0.99, correct=F)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(130, 158) out of c(680, 680)
## X-squared = 3.4536, df = 1, p-value = 0.06312
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.09817718  0.01582423
## sample estimates:
##    prop 1    prop 2
## 0.1911765 0.2323529
```

Do not forget to knit your .Rmd file to .pdf (or to word and afterwards to pdf) before posting it on the ATENEA platform task (only for pdf). Markdown should be also posted in the corresponding task.