Statistical project

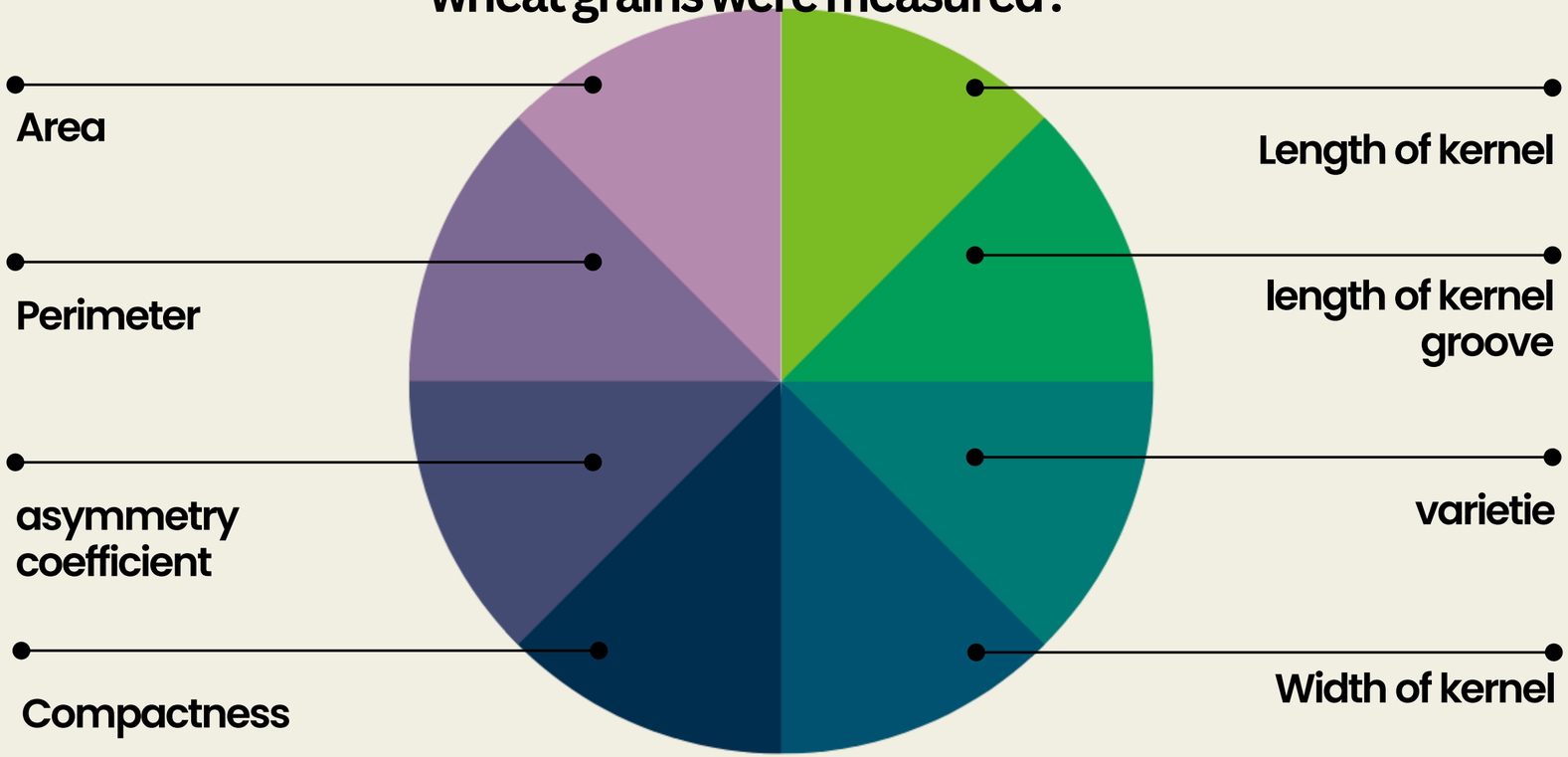
PROJECT: WHEAT GRAIN

Language: R

Plan

01 Introduction

To build our database, 8 geometric parameters wheat grains were measured :



02 Importing data

```
> seed = read.table(file=file.choose(),header=TRUE,sep=",",dec=".")
> str(seed)
'data.frame':   185 obs. of  8 variables:
 $ A       : num  15.3 14.9 14.3 13.8 16.1 ...
 $ P       : num  14.8 14.6 14.1 13.9 15 ...
 $ C       : num  0.871 0.881 0.905 0.895 0.903 ...
 $ Lk      : num  5.76 5.55 5.29 5.32 5.66 ...
 $ Wk      : num  3.31 3.33 3.34 3.38 3.56 ...
 $ Ac      : num  2.22 1.02 2.7 2.26 1.35 ...
 $ Lkg     : num  5.22 4.96 4.83 4.8 5.17 ...
 $ varietie: chr  "Canadian wheat" "Canadian wheat" "Canadian wheat"
```

```
> summary(seed)
       A               P               C                Lk
 Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
 1st Qu.:12.15   1st Qu.:13.38   1st Qu.:0.8559   1st Qu.:5.236
 Median :13.99   Median :14.09   Median :0.8723   Median :5.452
 Mean   :14.59   Mean   :14.44   Mean   :0.8701   Mean   :5.593
 3rd Qu.:16.63   3rd Qu.:15.38   3rd Qu.:0.8879   3rd Qu.:5.886
 Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675
                                 NA's   :5        NA's   :1
       Wk              Ac              Lkg          varietie
 Min.   :2.630   Min.   :0.7651   Min.   :4.605   Length:185
 1st Qu.:2.893   1st Qu.:2.5530   1st Qu.:5.012   Class :character
 Median :3.186   Median :3.5970   Median :5.194   Mode  :character
 Mean   :3.227   Mean   :3.7155   Mean   :5.371
 3rd Qu.:3.514   3rd Qu.:4.7730   3rd Qu.:5.795
 Max.   :4.033   Max.   :8.4560   Max.   :6.550
```

185 Observations

8 Variables

6 Variables

QUANTITATIVE EXPLANATORY

A, P,C Lk,Wk,Ac

2 VARIABLES

TO EXPLAIN

1 QUANTITATIVE

Lkg

1 QUALITATIVE

variety

03 Data preprocessing

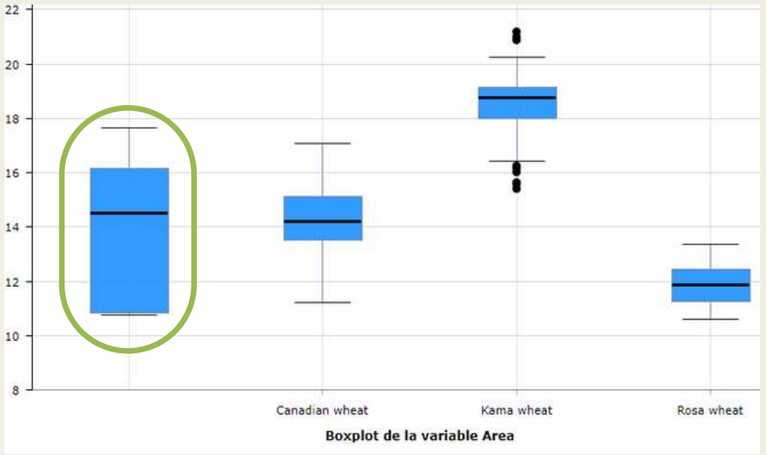
```
> unique(seed$varietie)
[1] "Canadian wheat" ""                "Kama wheat"
[4] "Rosa wheat"
> by(seed,seed$varietie,nrow)
seed$varietie:
[1] 6
------------------------------------------------
seed$varietie: Canadian wheat
[1] 58
------------------------------------------------
seed$varietie: Kama wheat
[1] 53
------------------------------------------------
seed$varietie: Rosa wheat
[1] 68
```

```
> Percent_MV_varieties = (nrow(subset(seed,varietie == "" ))/nrow(seed))*100
> Percent_MV_varieties
[1] 3.243243
```

185 OBS
100%

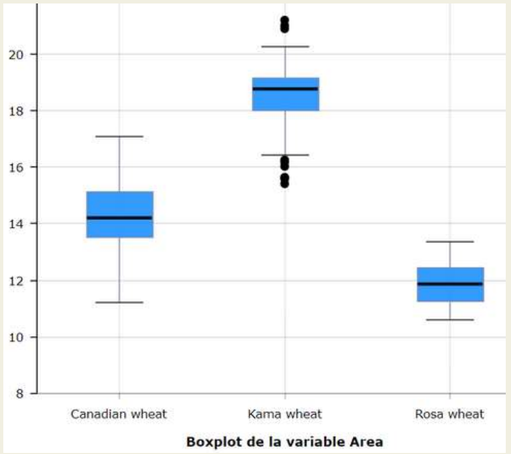
6 MQ
3%

Visualization of outlier data by variety



4 modalities

Delete the lines where there are missing values in varietie.



3 modalities

03

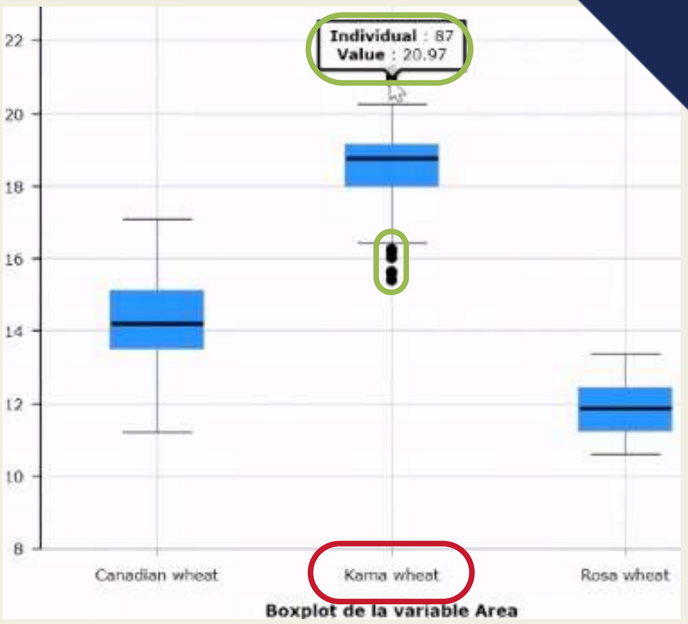
Data preprocessing

1. Visualization of Outlier values:

Example Area:

```
summary(seed$A[seed$varietie=="Kama wheat"])
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
15.38   17.99   18.76   18.43   19.14   21.18
```

```
Q1=17.99
Q3=19.14
Vmin=Q1-1.5*(Q3-Q1)
Vmax=Q3+1.5*(Q3-Q1)
IC=(Vmax-Vmin)*0.05
min_outliers=which(seed$A[seed$varietie=="Kama wheat"]<Vmin)
max_outliers=which(seed$A[seed$varietie=="Kama wheat"]>Vmax)
Area_outliers=c(min_outliers,max_outliers)
Area_outliers
for (a in Area_outliers) {
  seed$A[seed$varietie == "Kama wheat"][a]=NA
```



2. We replace the outliers with Nan:

```
> grubbs.test(seed$A[seed$varietie=="Kama wheat"], type = 11,two.sided = FALSE)

        Grubbs test for two opposite outliers

data:  seed$A[seed$varietie == "Kama wheat"]
G = 4.76987, U = 0.70795, p-value = 0.2918
alternative hypothesis: 16.41 and 20.24 are outliers
```

3. Check the existence of Area outliers for the Kama modality by grubbs test:

H0: There is no outlier in our Area column
H1: There is an outlier in our Area column

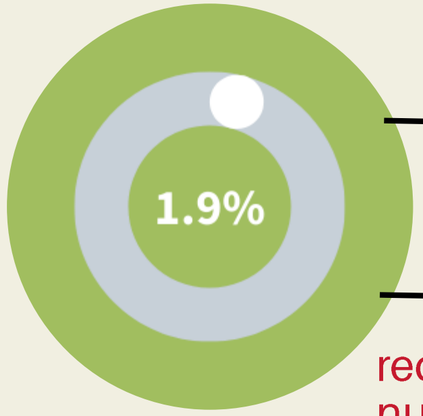
➡ p-value = 0.2918 > 0.05

✓ H0 ✗ H1

Data preprocessing

1. Visualization of Outlier values:

```
> Taux
[1] 0.01995211
> colSums(is.na(seed))
       A        P        C       Lk       Wk       Ac      Lkg varietie
      11        1        4        3        0        4        2        0
```

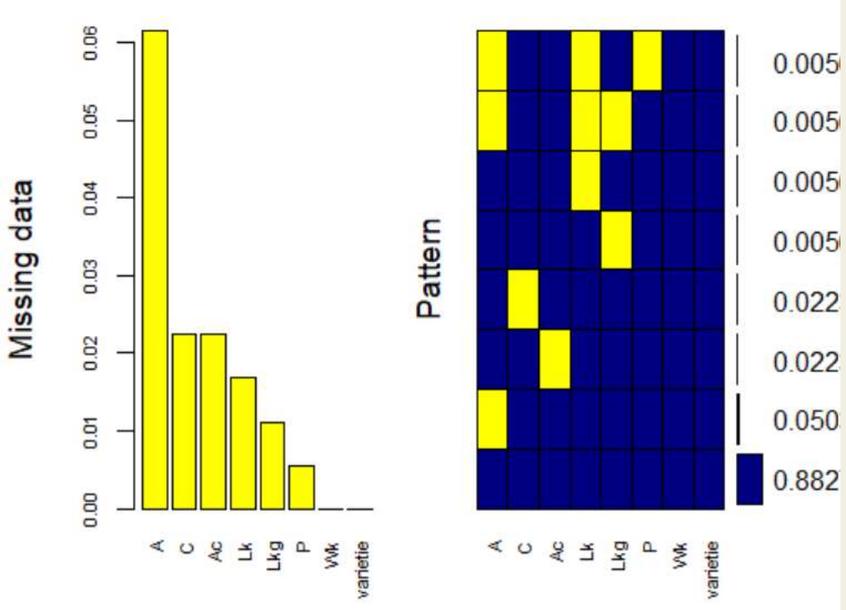


< 5%

elimination of observations from the base

reduced number of observations

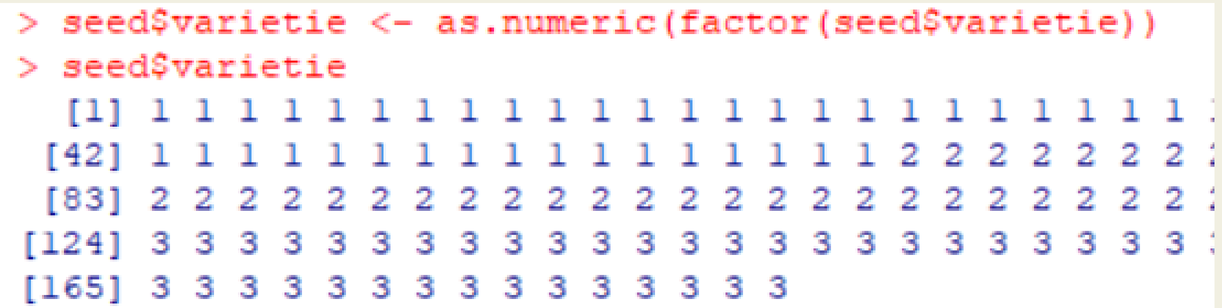
Imputation to avoid loss of information



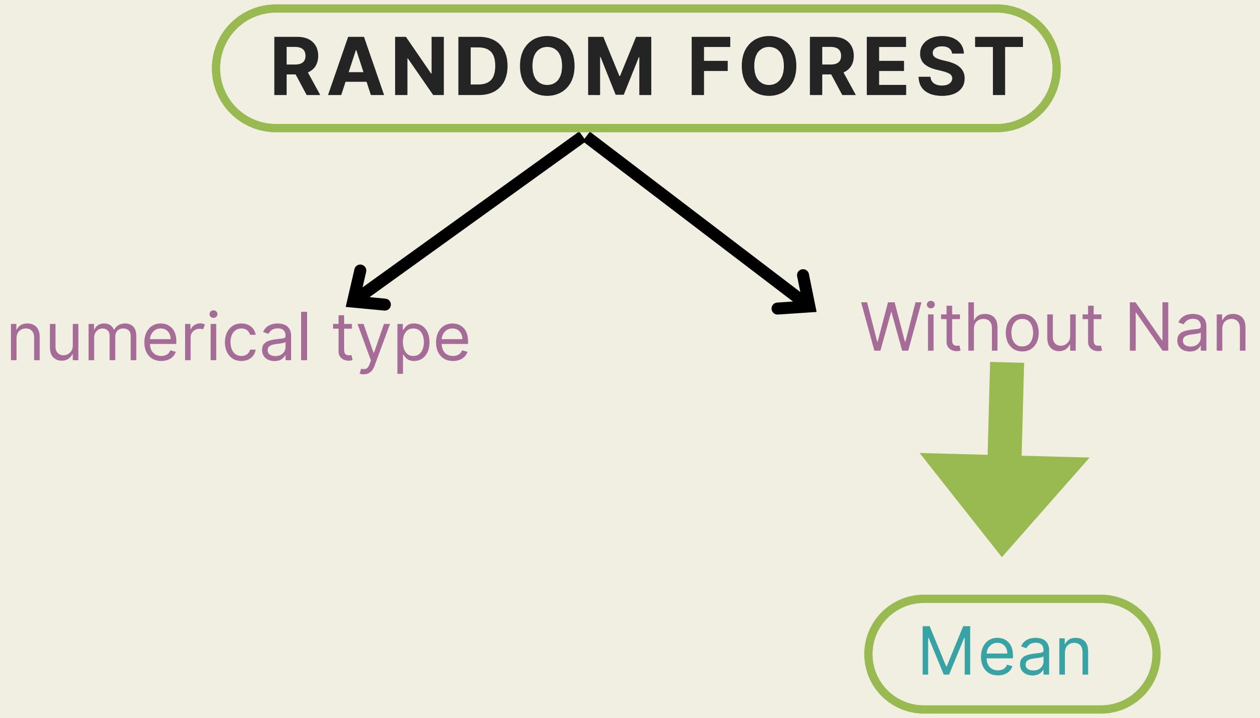
03

Data preprocessing

2. Processing Missing values :



```
> seed$varietie <- as.numeric(factor(seed$varietie))
> seed$varietie
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [42] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
 [83] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[124] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[165] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```



RANDOM FOREST

numerical type

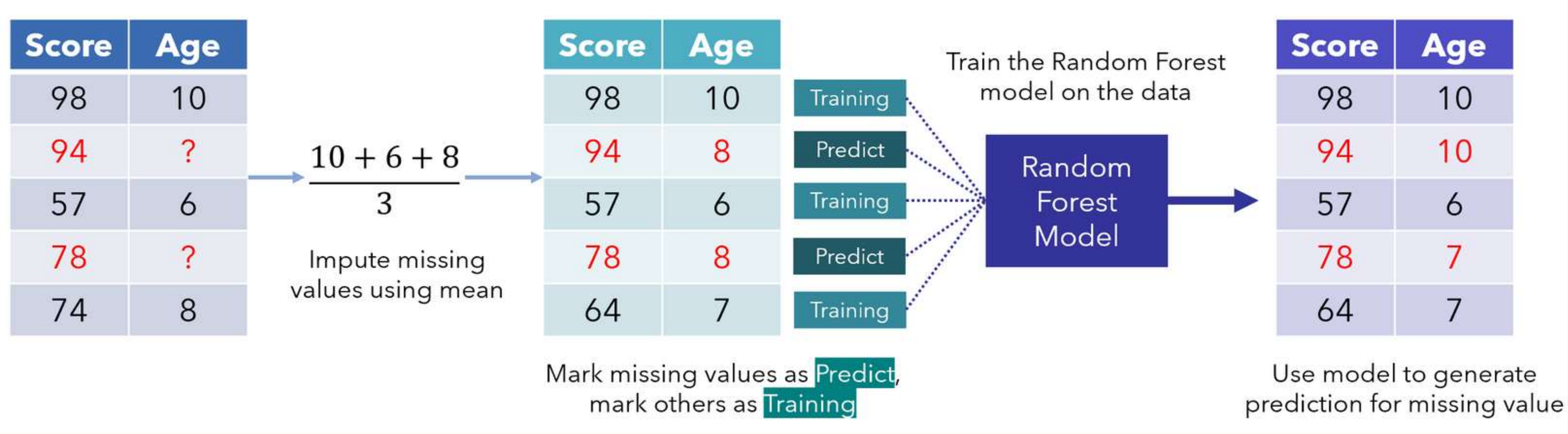
Without Nan

Mean

➕ Mean

Random Forest

🟰 Miss Forest



Score	Age
98	10
94	?
57	6
78	?
74	8

$$\frac{10+6+8}{3}$$

Impute missing values using mean

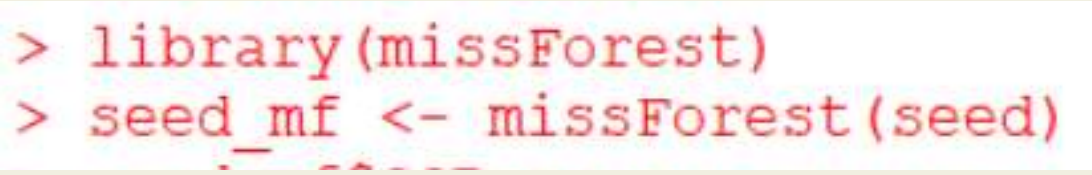
Score	Age	
98	10	Training
94	8	Predict
57	6	Training
78	8	Predict
64	7	Training

Mark missing values as Predict, mark others as Training

Train the Random Forest model on the data

Score	Age
98	10
94	10
57	6
78	7
64	7

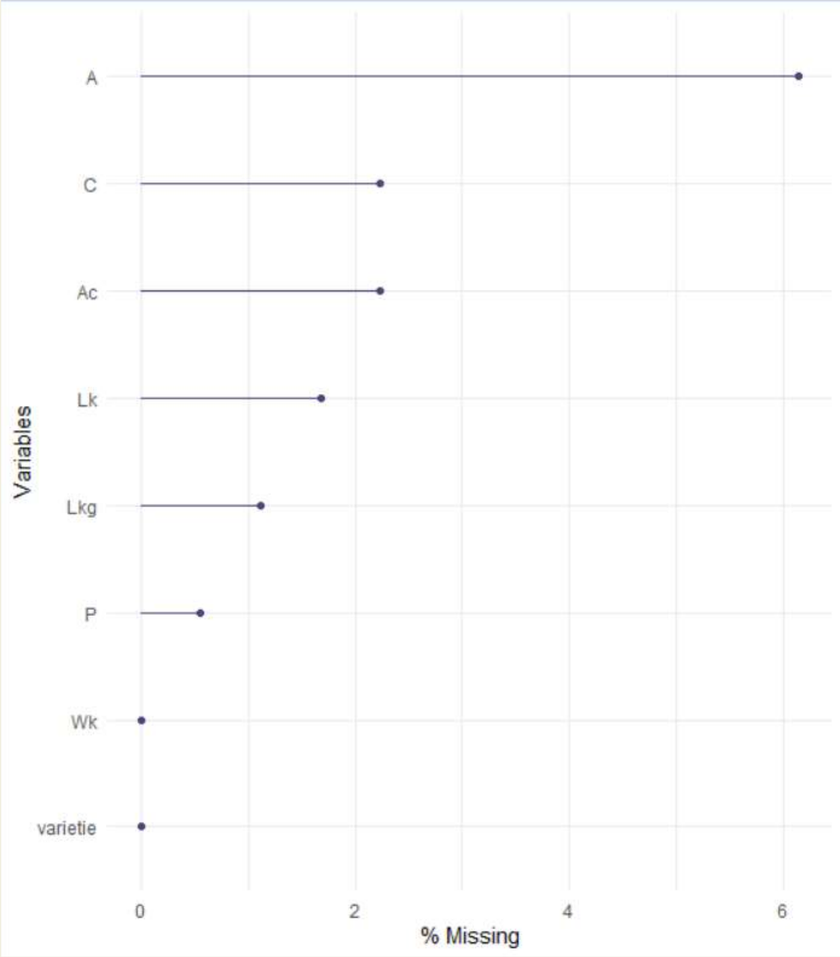
Use model to generate prediction for missing value

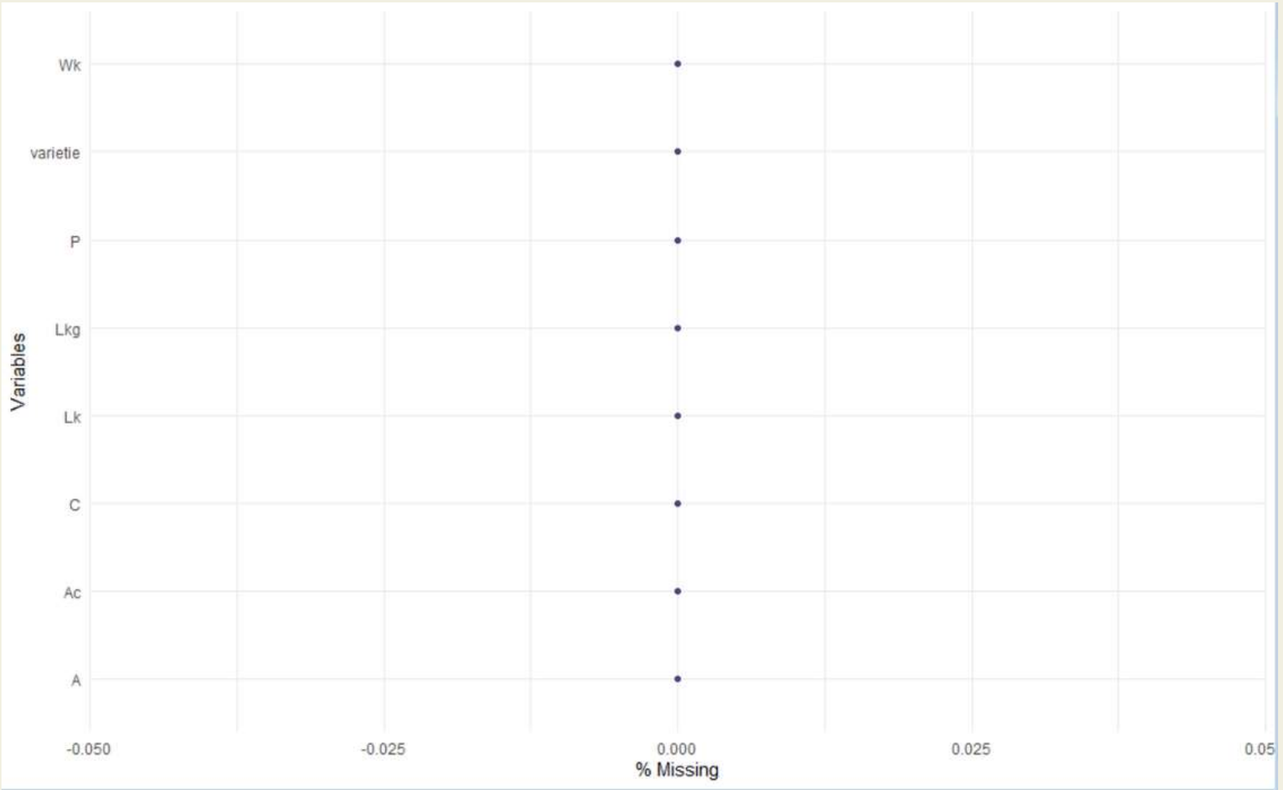


```
> library(missForest)
> seed_mf <- missForest(seed)
```

Data preprocessing

3. Visualization after imputation of Missing values:



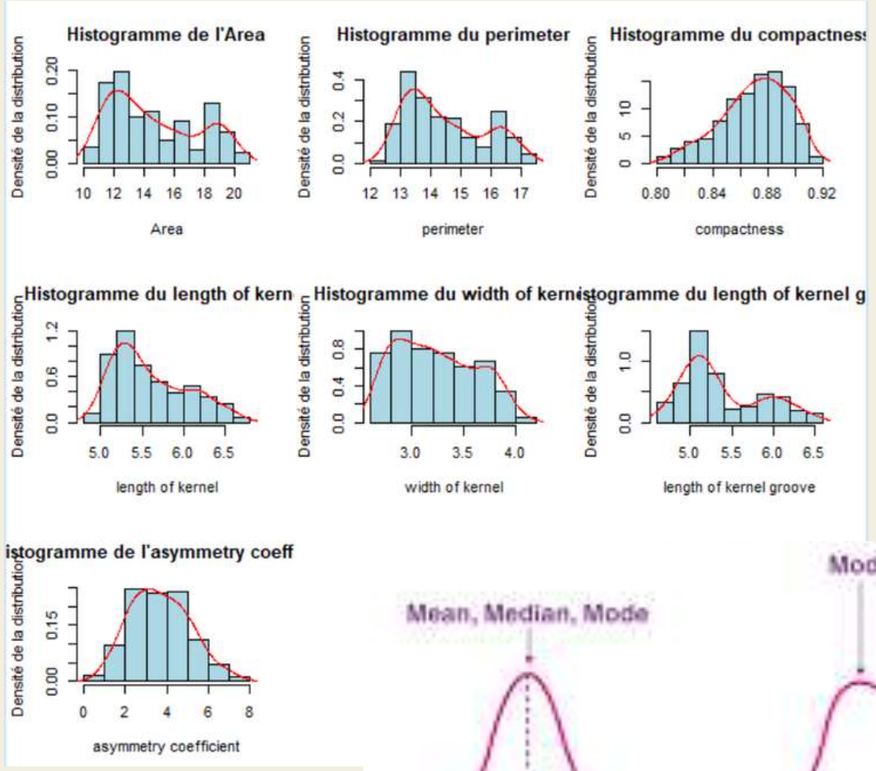


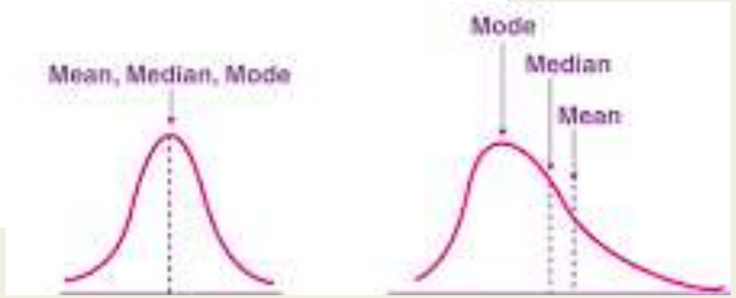
04

Univariate analysis

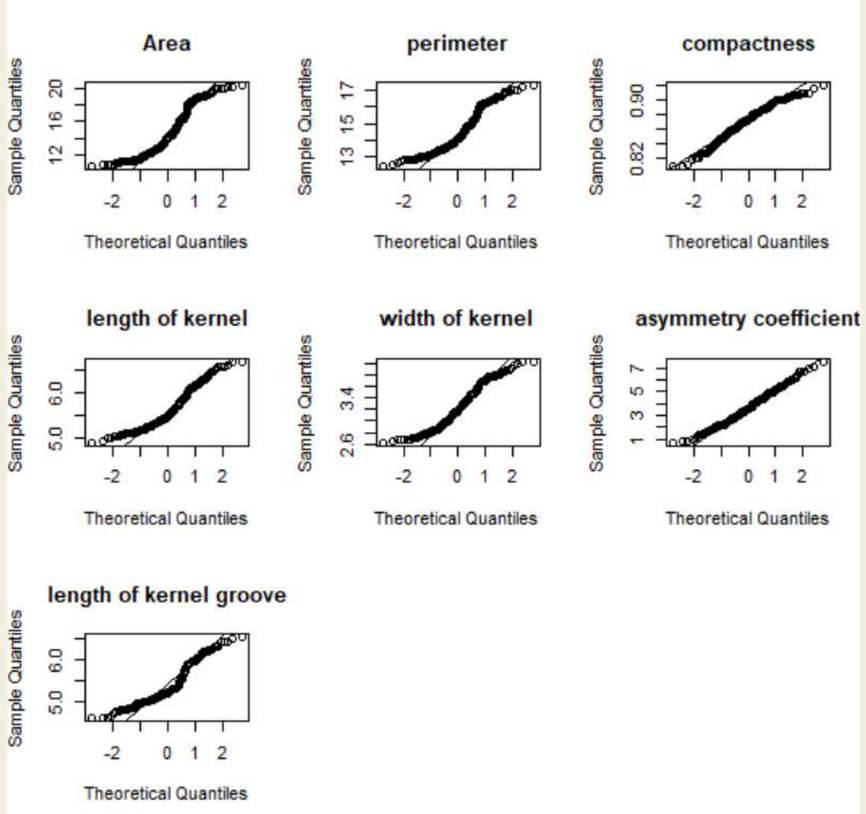
4.1- Study of normality

a- Graphical representation:





Histogram



Henry's line

Univariate analysis

b- By Calculation:

Shapiro test:

```
>
> #-----------------------------Pour la colonne width of kernel---------------------
> #H0 : Wk suit une normale
> #H1 : Wk ne suit pas une normale
> shapiro.test(seed_imputed$Wk)

        Shapiro-Wilk normality test

data:  seed_imputed$Wk
W = 0.95014, p-value = 6.197e-06

> #p_value<0.05 : on rejette H0 et on accepte H1
> #La variable Wk ne suit pas la loi normale
>
>
> #---------------------------Pour la colonne asymmetry coefficient-----------------
> #H0 : Ac suit une normale
> #H1 : Ac ne suit pas une normale
> shapiro.test(seed_imputed$Ac)

        Shapiro-Wilk normality test

data:  seed_imputed$Ac
W = 0.98745, p-value = 0.1122

> #p_value>0.05 : on accepte H0 et on rejette H1
> #La variable Ac suit la loi normale
>
```

Univariate analysis

Conclusion

NORMALITY TEST

- Graphically
  - Histogram
  - Henry's line
- By calculation: shapiro

Univariate analysis

4.2- Study of the modality of the variable variety:

```
> unique(seed_imputed$varietie)
[1] 1 2 3
> table(seed_imputed$varietie)

 1  2  3
58 53 68
> by(seed_imputed,seed$varietie,nrow)
seed$varietie: 1
[1] 58
------------------------------------------------------------------------
seed$varietie: 2
[1] 53
------------------------------------------------------------------------
seed$varietie: 3
[1] 68
>
> #on a trois modalité :
> #"Canadian wheat" avec 58 observations
> #"Kama wheat" avec 53 observations
> #"Rosa wheat" avec 68 observations
> #On a une database balanced : les modalités sont réparties de facon équitable
```

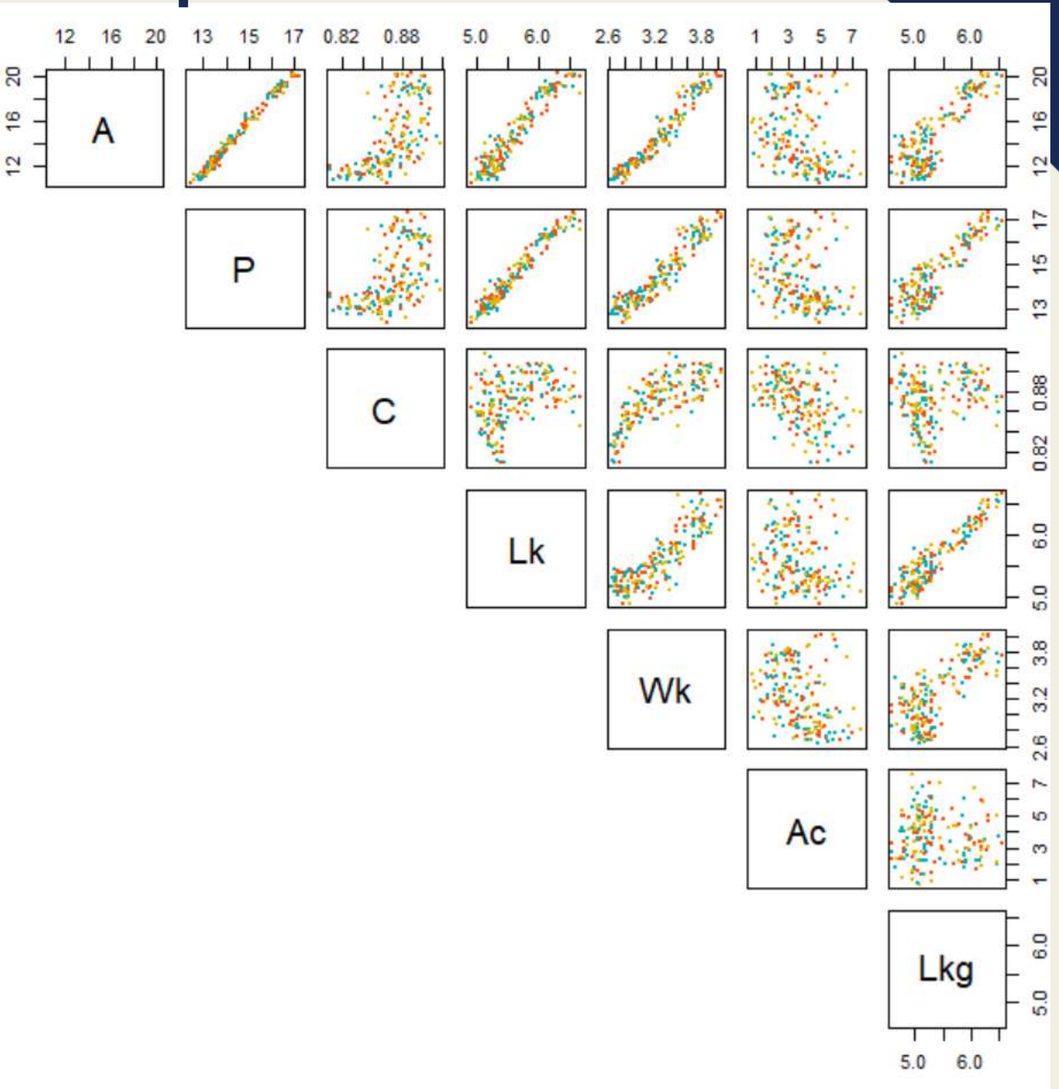
05

Bivariate analysis :

5.1. Study of the dependence relationship between quantitative variables:

a- Graphical representation:

Relation linéaire positive	Liaison monotone positive non linéaire	Absence de liaison
A-P A-Lk A-Wk A-Lkg P-Lk P-Wk P-Lkg Lk-Lkg Wk-Lkg	A-C P-C C-Lk C-Wk Lk-Wk	Ac { p, A, C, Lk, Wk, Lkg



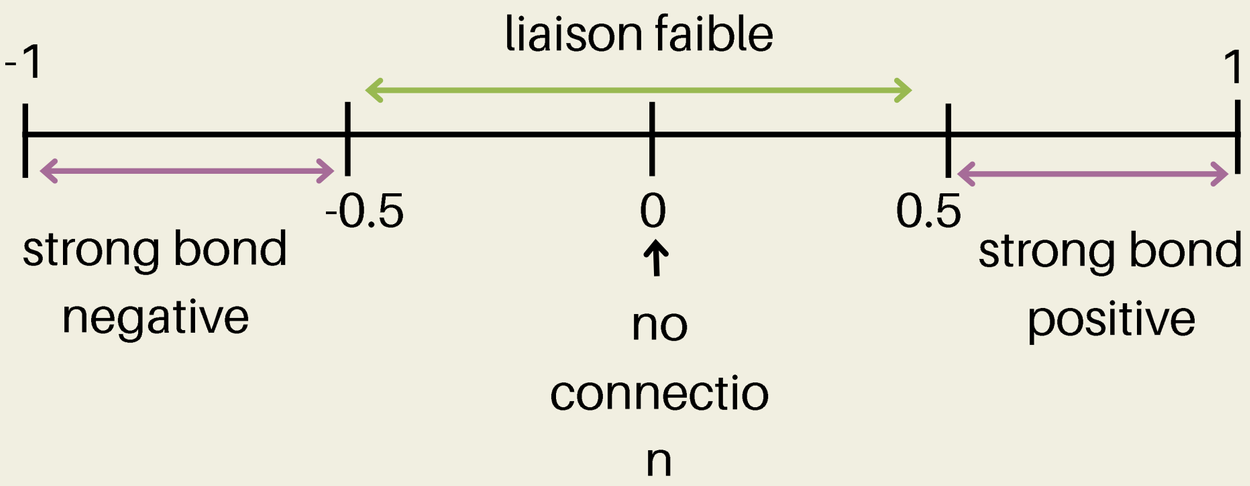
05

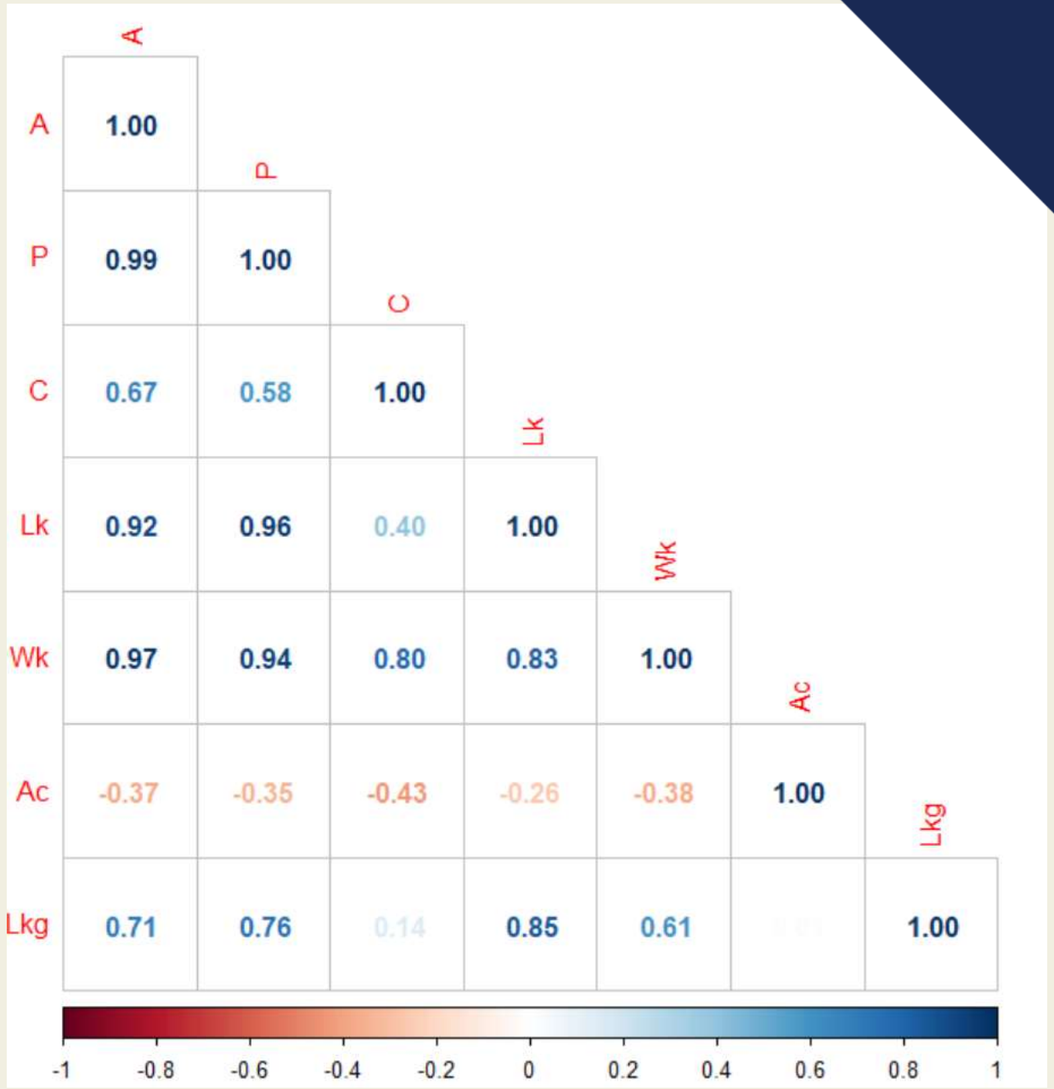
Bivariate analysis :

b- Calculation of statistical indicators

Spearman's correlation coefficient:

```
> RAC=cor(seed_imputed$A,seed_imputed$C,method="spearman")
```






```
> corrplot(cor(seed_imputed[,1:7],method="s"), type = "lower", method = "number")
```

05

Bivariate analysis :

c - Correlation test

```
#H0:r=0 => absence de dépendance
#H1:r<>0 => présence de dépendance

cor.test(seed_imputed$A,seed_imputed$P,method="spearman")

#p-value=2.2e-16<<<<<0.05 => p-value est négligeable=>accepter H1
#=>A et P sont fortement corrélées

cor.test(seed_imputed$Ac,seed_imputed$Lk,method="spearman")

#p-value= 0.0003664 <0.05 =>accepter H1
#=>Ac et Lk sont faiblement corrélées car p-value est proche de 0.05
```

05

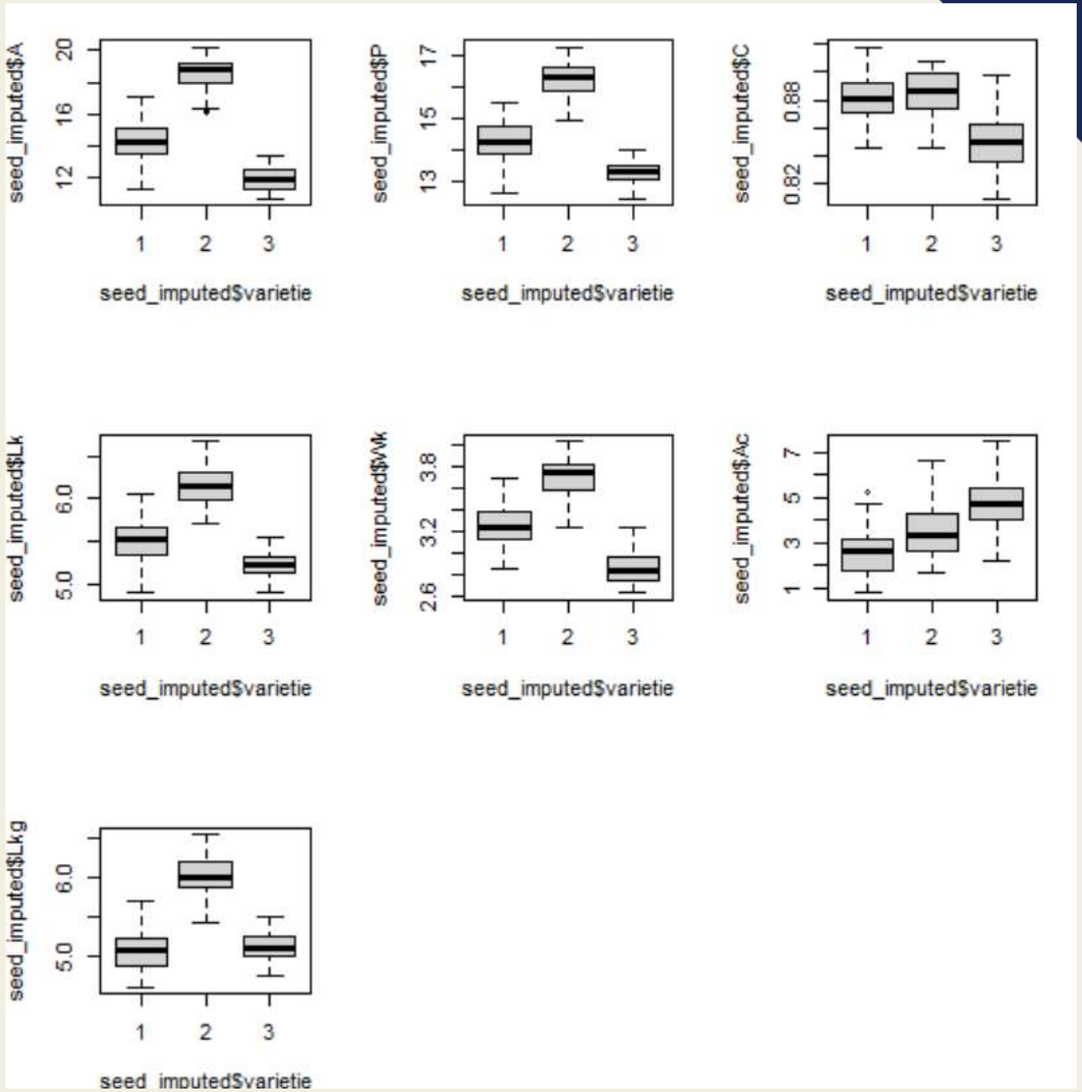
Bivariate analysis :

5.2- 5.2- Study of the dependence relationship between the quantitative variables and the qualitative variable:

a- Graphical representation

By changing the modality of the variable varietie, we notice a change in the reference values of the other variables.

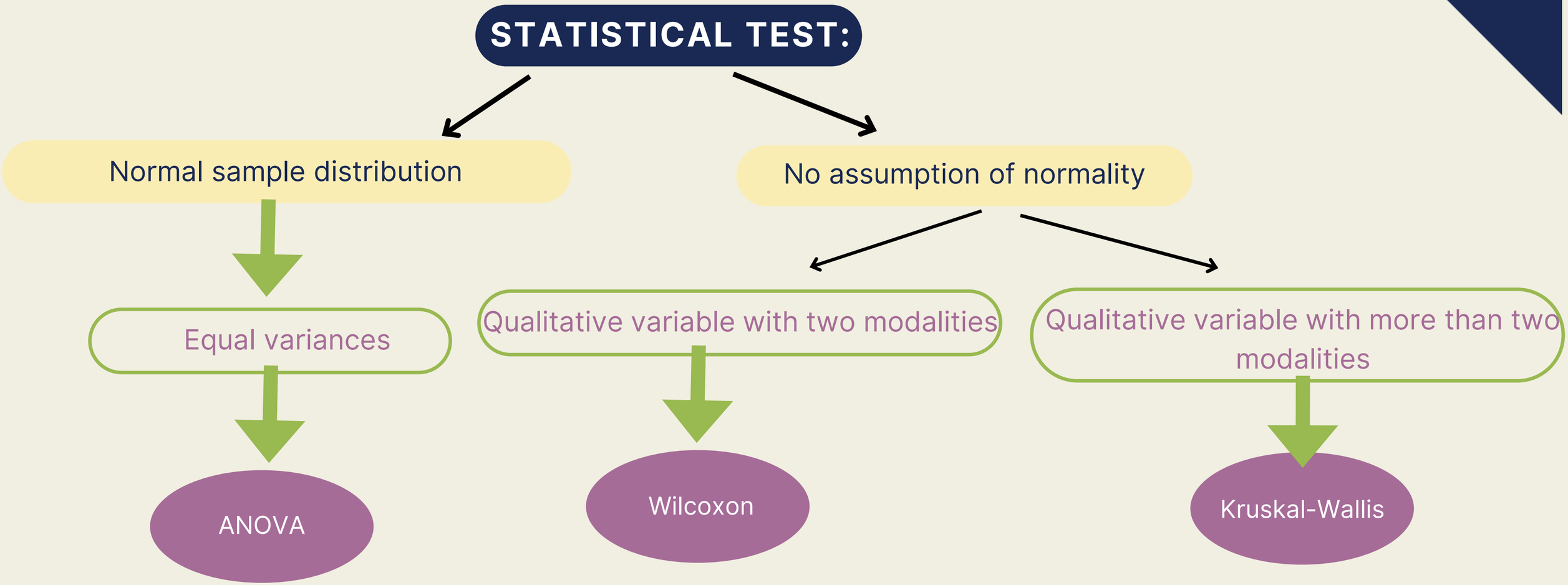
There is an effect of the target variable varietie on all quantitative variables.



05

Bivariate analysis :

b- Statistical Test:



05

Bivariate analysis :

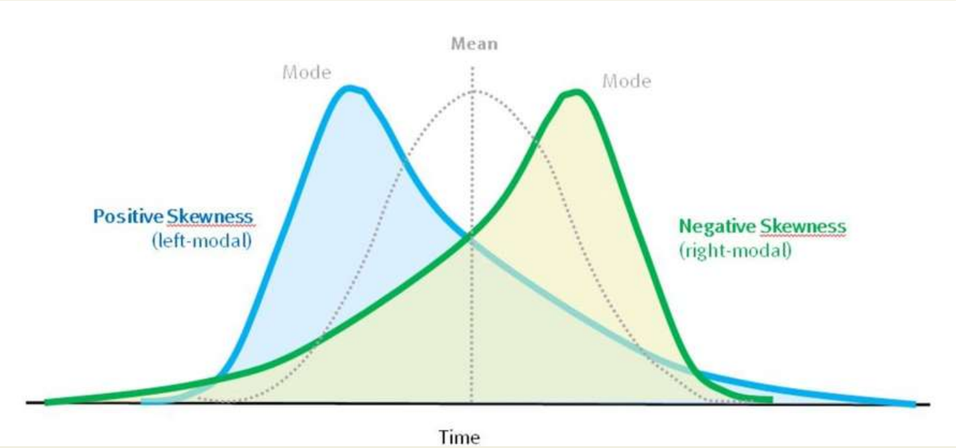
```
> #On est dans le cas de deux variables une quantitative
> #A/P/C/Lk/Wk/Ac/Lkg et une qualitative varietie
> #Le premier test possible à appliquer est l'ANOVA
> #Vérifiant alors ces conditions d'application:
>
> tapply(seed_imputed$A,seed_imputed$varietie,shapiro.test)
$`1`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.99097, p-value = 0.944


$`2`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.90785, p-value = 0.0006049


$`3`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.96808, p-value = 0.0785


>
> #pvalue = 0.944 > 0.05, alors on a la normalité pour le groupe "canadian wheat".
> #pvalue = 0.0006571 < 0.05, alors n'on pas la normalité pour le groupe "kama wheat".
> #pvalue = 0.0785 > 0.05, alors a la normalité pour le groupe "Rosa wheat".
> #==> on n'a pas la normalité
```

```
> #la variable qualitative varietie possède plus que deux modalités (3)
> #=>On applique le test de Kruskall-wallis
> #H0:variables indépendantes
> #H1:variables liées(les distributions des échantillons ne sont pas les mêmes)
>
> kruskal.test(seed_imputed$A ~ seed_imputed$varietie)

        Kruskal-Wallis rank sum test

data:  seed_imputed$A by seed_imputed$varietie
Kruskal-Wallis chi-squared = 147.02, df = 2, p-value < 2.2e-16

>
> #p-value=2.2e-16 <<<< 0.05
> #on accepte H1 : les distributions des échantillons ne sont pas les mêmes,
> #il existe alors une différence entre les différentes modalités
> #donc l'effet est présent et par la suite
> #on ne peut pas ignorer la relation entre les deux variables varietie et A.
```

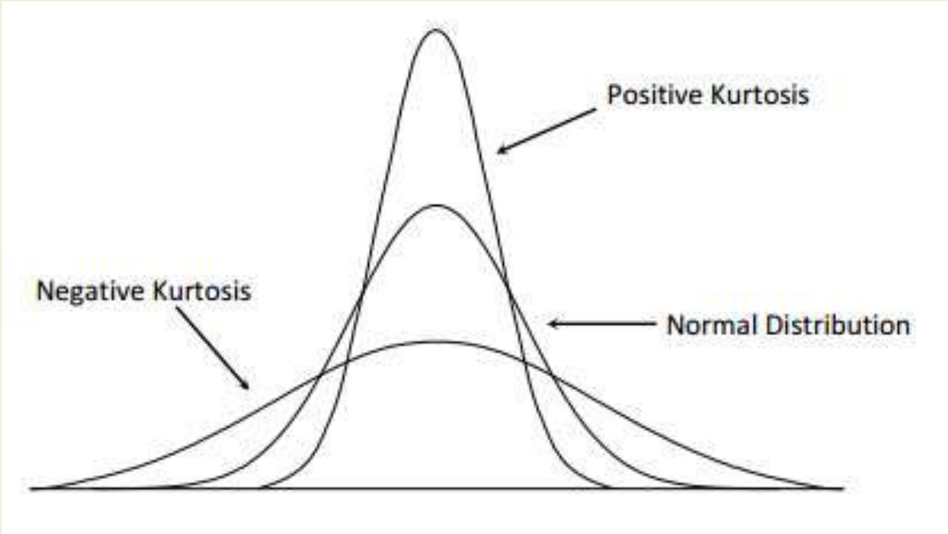
H0 H1

05 Bivariate analysis :

```
> tapply(seed_imputed$Ac,seed_imputed$varietie,shapiro.test)
$`1`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98231, p-value = 0.5567


$`2`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.95733, p-value = 0.05606


$`3`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98957, p-value = 0.8446


> #on a la normalité
```




```
> skewness(seed_imputed$Ac)
[1] 0.2689857
```




```
> kurtosis(seed_imputed$Ac)
[1] 2.539139
```

05

Bivariate analysis :

Normality → Shapiro.test

Equal variances → Bartlet

→ **ANOVA: DEPENDANCE**

```
> bartlett.test(seed_imputed$Ac ~ seed_imputed$varietie)

        Bartlett test of homogeneity of variances

data:  seed_imputed$Ac by seed_imputed$varietie
Bartlett's K-squared = 1.381, df = 2, p-value = 0.5013

> #p-value = 0.5013>0.05 =>On accepte H0
>
> #==> On peut appliquer le test statistique ANOVA
> #H0: variables indépendantes
> #H1: présence de dépendance entre les deux varibles
> fit <- aov(seed_imputed$Ac ~ seed_imputed$varietie)
> summary(fit)
                       Df Sum Sq Mean Sq F value Pr(>F)
seed_imputed$varietie   1  140.8  140.77   112.6 <2e-16 ***
Residuals             177  221.2    1.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #p-value<2e-16 <0.001
> # On accepte H1
> # au moins un niveau(groupe) avec une moyenne significativement différente
> # Il existe un effet de varietie sur Ac
```

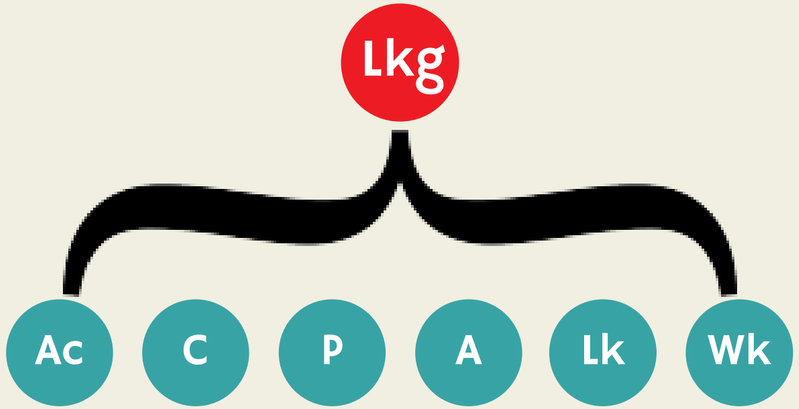
✓ H0 ✗ H1

✗ H0 ✓ H1

06

Linear regression :

6.1- Regression of the quantitative target variable Lkg according to the others:



(R squared) = Sum of Squares Explained/ Sum of Squares Total

```
> D=lm(Lkg ~ A+P+C+Lk+Wk+Ac ,data=seed_imputed)
> summary(D)

Call:
lm(formula = Lkg ~ A + P + C + Lk + Wk + Ac, data = seed_imputed)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41102 -0.08040  0.00099  0.09716  0.29221 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.903466   2.172037   5.480 1.49e-07 ***
A            0.418297   0.057306   7.299 1.02e-11 ***
P           -0.626575   0.145726  -4.300 2.86e-05 ***
C           -7.269379   1.965942  -3.698 0.000292 ***
Lk           0.672920   0.156694   4.294 2.92e-05 ***
Wk          -0.372342   0.295683  -1.259 0.209642    
Ac           0.046892   0.008604   5.450 1.72e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.136 on 172 degrees of freedom
Multiple R-squared:  0.9236,	Adjusted R-squared:  0.9209 
F-statistic: 346.6 on 6 and 172 DF,  p-value: < 2.2e-16
```

06 Linear regression :

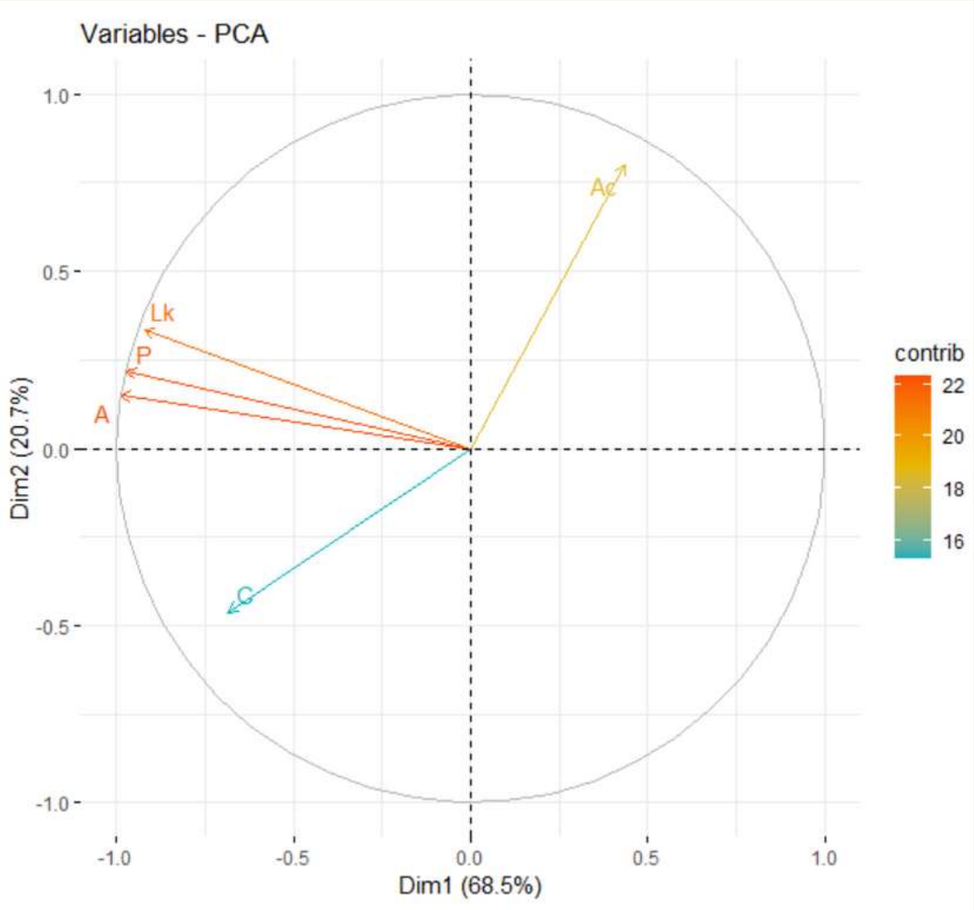
6.2- Improvement of the performance of the regression model:

```
> DSWk=lm(Lkg ~ A+P+C+Lk+Ac ,data=seed_imputed)
> summary(DSWk)

Call:
lm(formula = Lkg ~ A + P + C + Lk + Ac, data = seed_imputed)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41296 -0.08306 -0.00036  0.09821  0.29515 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.552810   1.735628   7.809 5.31e-13 ***
A            0.412312   0.057205   7.208 1.69e-11 ***
P           -0.711640   0.129342  -5.502 1.33e-07 ***
C           -9.288077   1.139932  -8.148 7.13e-14 ***
Lk           0.715176   0.153318   4.665 6.16e-06 ***
Ac           0.042825   0.007988   5.361 2.62e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1362 on 173 degrees of freedom
Multiple R-squared:  0.9229,	Adjusted R-squared:  0.9207 
F-statistic: 414.2 on 5 and 173 DF,  p-value: < 2.2e-16

>
> #Le modèle ne perd pas de qualité
> #=> on a toujours R2=0.9229
> #le même taux d'information expliquée de la variabilité de Lkg.
> #la variable éliminée n'a pas d'importance
```

```
> DSLk=lm(Lkg ~ A+P+C+Ac ,data=seed_imputed)
> summary(DSLk)

Call:
lm(formula = Lkg ~ A + P + C + Ac, data = seed_imputed)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37028 -0.09825 -0.00697  0.09316  0.36179 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.325692   1.725188   9.463  < 2e-16 ***
A             0.421362   0.060487   6.966 6.44e-11 ***
P            -0.467177   0.125103  -3.734 0.000255 ***
C           -12.081246   1.026227 -11.772  < 2e-16 ***
Ac            0.041996   0.008449   4.971 1.59e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1441 on 174 degrees of freedom
Multiple R-squared:  0.9132,	Adjusted R-squared:  0.9112 
F-statistic: 457.7 on 4 and 174 DF,  p-value: < 2.2e-16

> # R2:  0.9132 =>le modèle perd de qualité
> #=>la varible Lk est significative
> #=>On va considérer le modèle précédant DSWk comme modèle optimal
```

06 Linear regression :

6.3- principal component analysis:



Wk : insignificant variable (According to linear regression)

Lkg : the target quantitative variable

varietie : target qualitative variable

→ To eliminate variables to be projected by the PCA

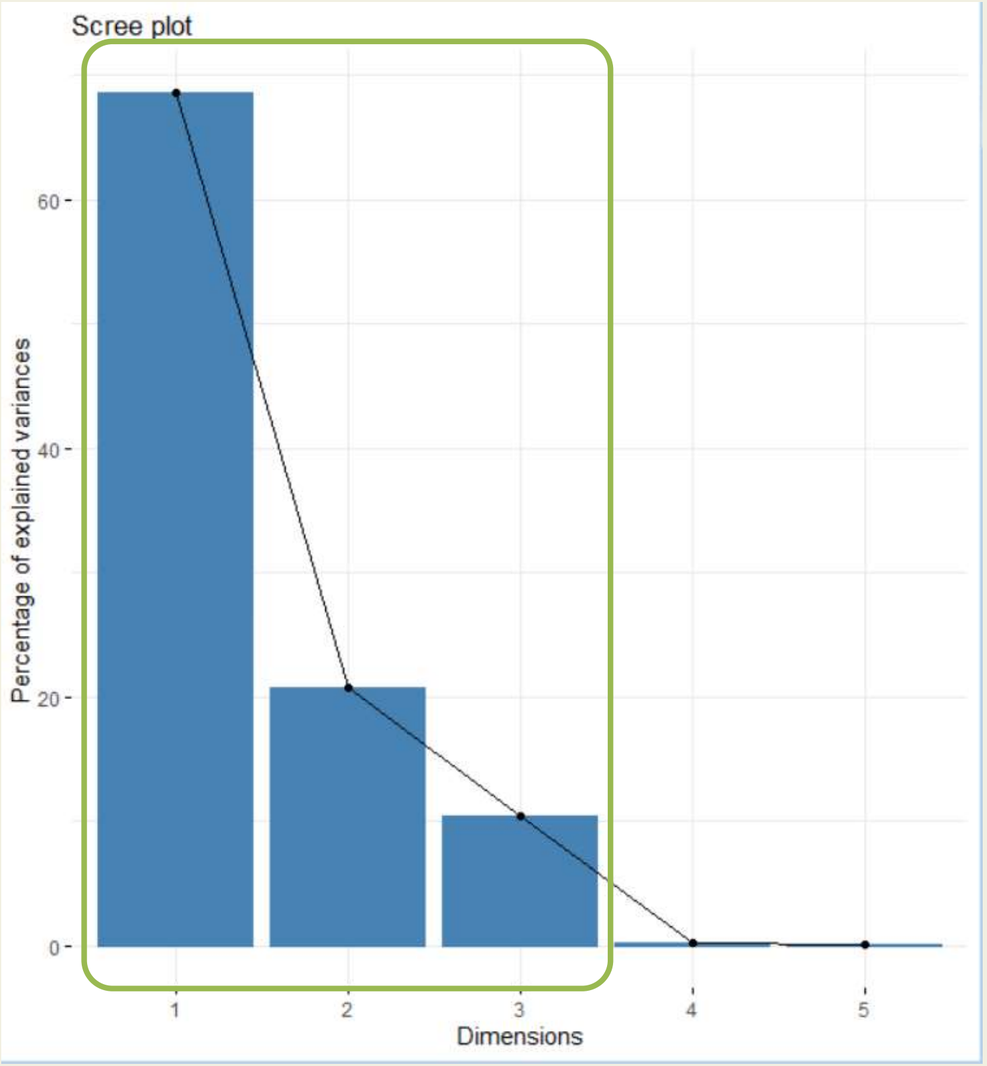
```
> results$rotation
          PC1        PC2         PC3          PC4         PC5
A  -0.5329520  0.1479898 -0.02333715  0.460678135  0.69375226
P  -0.5256544  0.2132081  0.07541820  0.402892332 -0.71429657
C  -0.3697605 -0.4569437 -0.77633620 -0.216961097 -0.06862654
Lk -0.4975099  0.3273642  0.25137922 -0.760494534  0.06142566
Ac  0.2353946  0.7852875 -0.57260638 -0.005419556 -0.00234499
```

→ The first 5 principal components

06

Linear regression :

6.3- Principal component analysis:



This graph represents the percentages of variances explained by each main axis.

Les trois premières dimensions sont les variances expliquées les plus importantes.

→ To keep

06

Linear regression :

6.3- Principal component analysis:

```
> NEW_DATA=results$x[, 1:3]
> NEW_MODEL=lm(seed_imputed$Lkg ~ NEW_DATA )
> summary(NEW_MODEL)

Call:
lm(formula = seed_imputed$Lkg ~ NEW_DATA)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.46235 -0.09641  0.00799  0.09683  0.35030 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.372542   0.011340 473.761  < 2e-16 ***
NEW_DATAPC1 -0.207583   0.006143 -33.790  < 2e-16 ***
NEW_DATAPC2  0.232842   0.011184  20.818  < 2e-16 ***
NEW_DATAPC3  0.120737   0.015716   7.682 1.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1517 on 175 degrees of freedom
Multiple R-squared:  0.9033,	Adjusted R-squared:  0.9016 
F-statistic: 544.7 on 3 and 175 DF,  p-value: < 2.2e-16

>
> #R-squared: 0.9032
> #90.32% de la variation de Lkg est expliquée par les variables résultantes
> #de la PCA
> #Malgré la diminution de R2, il reste un bon modèle
>
```

R-squared=0.9033

07

Generalized linear regression :

GLM is a flexible generalization of linear regression.
The dependent variable is linearly related to the variables via a precise link function.

Why we use GLMs

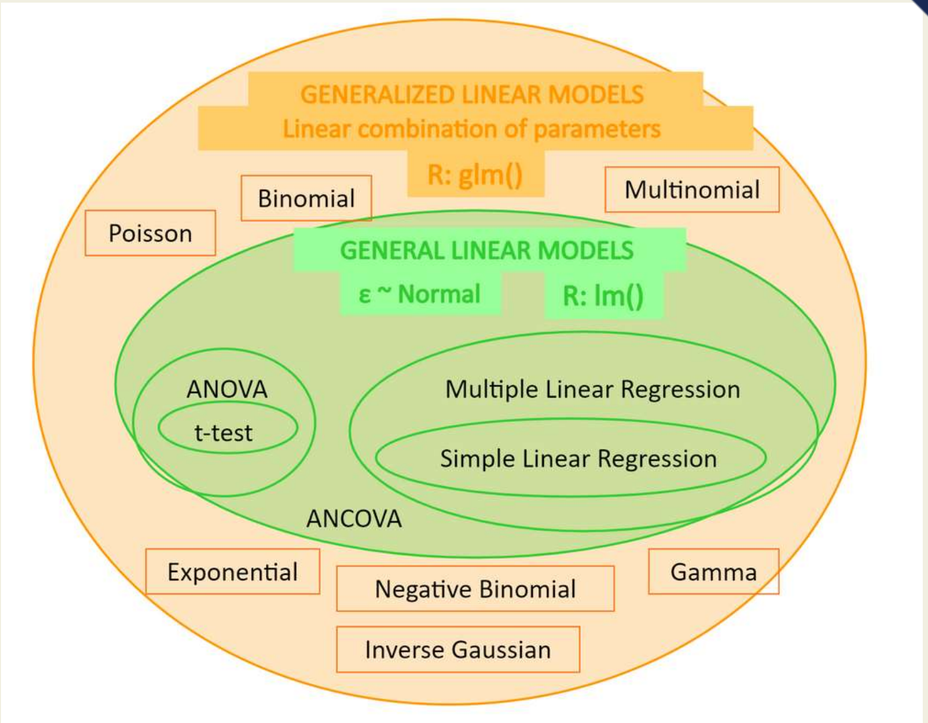
They provide a way to model dependent variables:

Non-normal distributions

no linear relationship with the explanatory variables

No need to check:

- The normality of the residuals
- Homoscedasticity
- That the dependent variables are continuous



07

Generalized linear regression :

TYPE OF GENERALIZED REGRESSION

Gaussian(normal)

If our variable is normally distributed

Fish(Total)

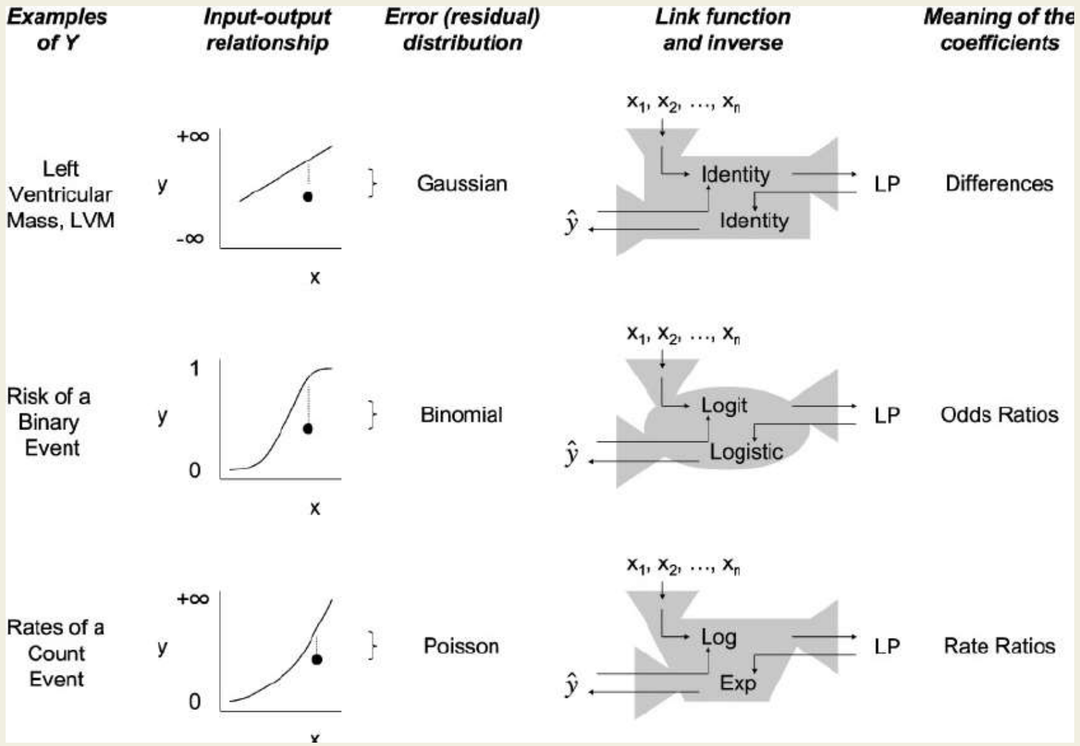
if our dependent variable is discrete and represents the number of occurrences of an event

Gamma(Continuous and +)

if values of dependent variables continue positive

Logistics(Binary)

If our dependent variable is of binary type



07

Generalized linear regression :

```
> GENERALIZED_MODEL <- glm(Lkg ~ A+P+C+Lk+Wk+Ac, data=seed_imputed,family=Gamma())
> #gamma : valeurs des variables dépendantes continue positifs
>
> summary(GENERALIZED_MODEL)

Call:
glm(formula = Lkg ~ A + P + C + Lk + Wk + Ac, family = Gamma(),
    data = seed_imputed)

Deviance Residuals:
      Min        1Q    Median        3Q       Max
-0.074866  -0.016248  0.001226  0.017230  0.058137

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0280443  0.0732571  -0.383 0.702326
A           -0.0128009  0.0018663  -6.859 1.20e-10 ***
P            0.0190367  0.0048757   3.904 0.000135 ***
C            0.2558217  0.0676622   3.781 0.000215 ***
Lk          -0.0211096  0.0052306  -4.036 8.18e-05 ***
Wk           0.0088303  0.0101111   0.873 0.383702
Ac          -0.0015049  0.0003017  -4.987 1.49e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.000658218)

    Null deviance: 1.39491  on 178  degrees of freedom
Residual deviance: 0.11374  on 172  degrees of freedom
AIC: -193.15

Number of Fisher Scoring iterations: 3
```

```
> AIC(GENERALIZED_MODEL)
[1] -193.1547
> AIC(NEW_MODEL)
[1] -161.5193
> #AIC(GENERALIZED_MODEL)< AIC(NEW_MODEL)
> #==> Le modèle linéaire généralisé a une qualité d'ajustement meillsseure
> #que le modèle obtenu par la modélisation linéaire des données obtenues
> #de la PCA
```

Thank you for your Attention