Compte Rendu TP1 par JAFUNO Douba

Donnée Seeds

Informations sur le jeu de données:

Le groupe examiné comprenait des noyaux appartenant à trois variétés différentes de blé: Kama, Rosa et Canadian, de 70 éléments chacun, choisis au hasard pour l'expérience. Une visualisation de haute qualité de la structure interne du noyau a été détectée à l'aide d'une technique de rayons X douce. Il est non destructif et considérablement moins coûteux que d'autres techniques d'imagerie plus sophistiquées telles que la microscopie à balayage ou la technologie laser. Les images ont été enregistrées sur des plaques KODAK à rayons X de 13 x 18 cm. Les études ont été menées à l'aide de grains de blé récoltés dans des moissonneuses-batteuses provenant de champs expérimentaux et examinés à l'Institut d'agrophysique de l'Académie des sciences de Pologne à Lublin.

L'ensemble de données peut être utilisé pour les tâches de classification et d'analyse par grappes.

Informations d'attribut:

Pour construire les données, sept paramètres géométriques des grains de blé ont été mesurés:

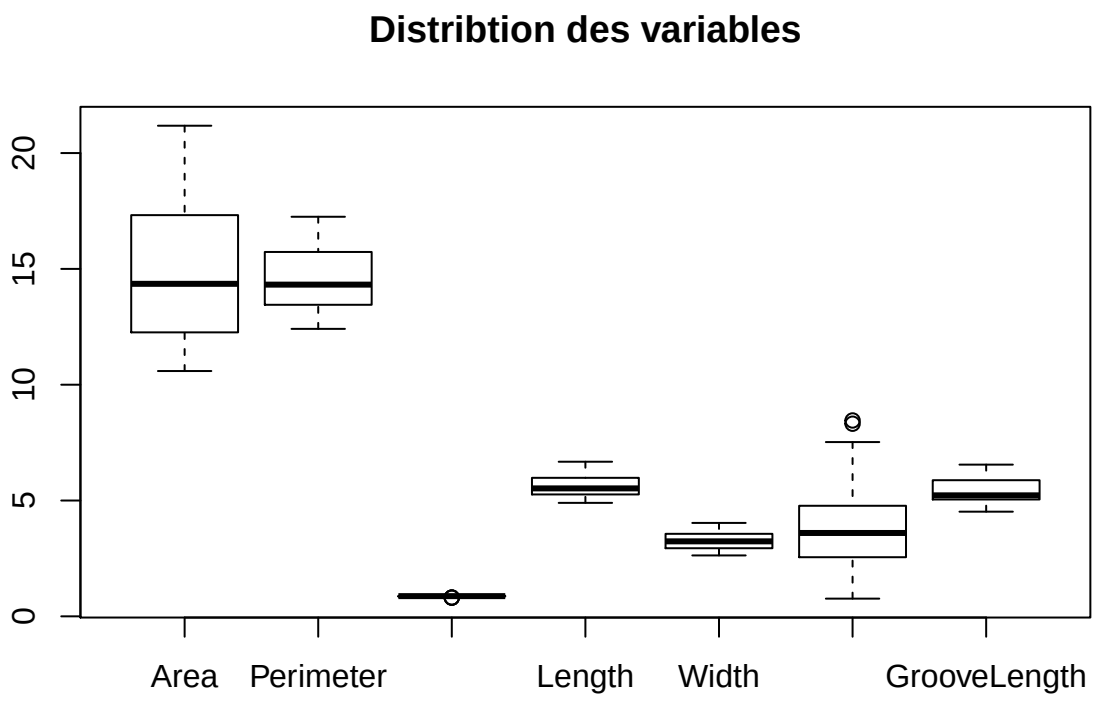
1. aire A,
2. périmètre P,
3. compacité C = 4 * pi * A / P ^ 2,
4. longueur du noyau,
5. largeur du noyau ,
6. coefficient d'asymétrie.
7. longueur de la gorge du noyau.

Tous ces paramètres étaient des valeurs réelles continues.

Nombres de valeur manquantes (nvm)

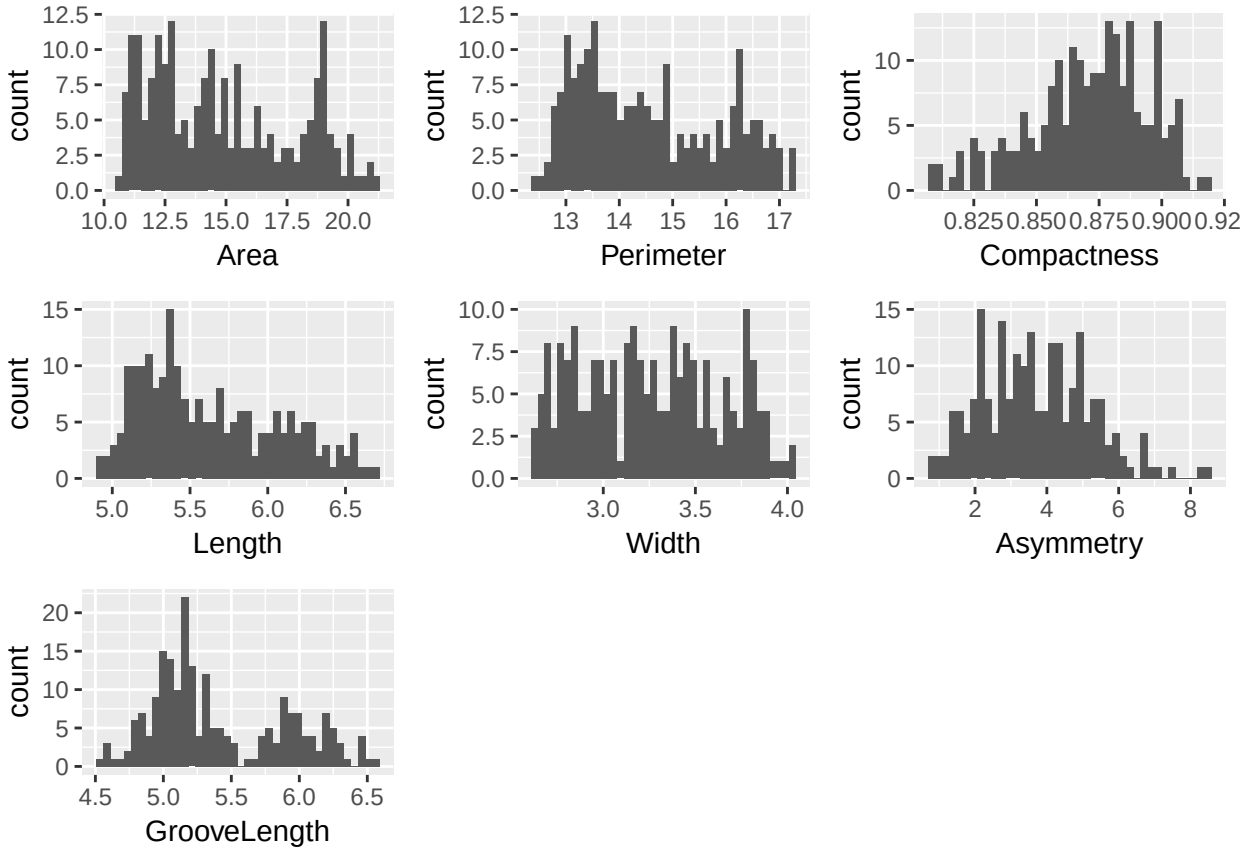
.	Area	Perimeter	Compactness	Length	Width	Asymmetry	GrooveLength
Nvm	0	0	0	0	0	0	0

Distributions des variables




```
     Area          Perimeter       Compactness         Length
Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262
Median :14.36   Median :14.32   Median :0.8734   Median :5.524
Mean   :14.85   Mean   :14.56   Mean   :0.8710   Mean   :5.629
3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980
Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675
    Width         Asymmetry       GrooveLength
Min.   :2.630   Min.   :0.7651   Min.   :4.519
1st Qu.:2.944   1st Qu.:2.5615   1st Qu.:5.045
Median :3.237   Median :3.5990   Median :5.223
Mean   :3.259   Mean   :3.7002   Mean   :5.408
3rd Qu.:3.562   3rd Qu.:4.7687   3rd Qu.:5.877
Max.   :4.033   Max.   :8.4560   Max.   :6.550
```

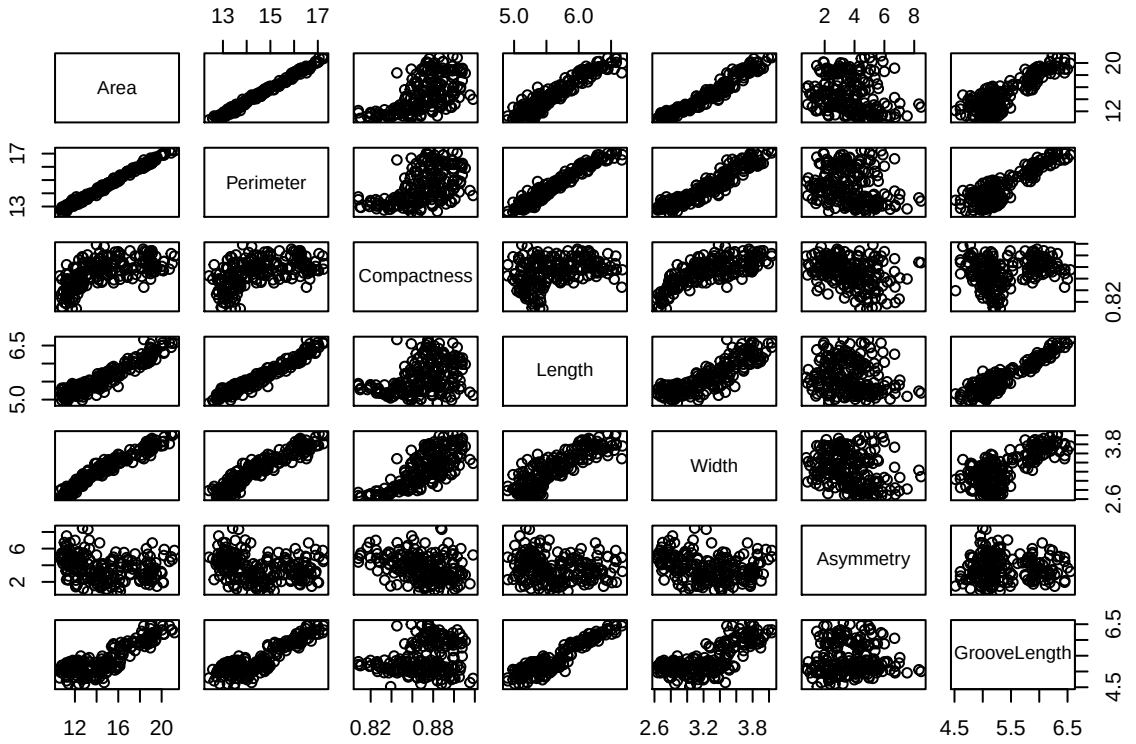
Histogrammes de chaque variable



la variable **Aréa** présente une plus grande variabilité de 10.59 à 21.18 un peu plus que la variable **Perimeter** et **Asymmetry** qui elle est entre 0 et 8, La distribution de la variable Compactness est très concentré autour de 0.8 et 0.9 c'est celle qui a les plus petites valeur. On voit aussi que pour nos 7 variables les histogrammes sont plutot **asymétriques** mais le boxplot montre bien les différences entre nos 7 variables comme on s'y attendais.

Corrélations entre nos variableS et nuages de points

Pour les corrélations comme il y a beaucoup de variables pour bien voir on peut se restreindre à quelques variables



.	Area	Perimeter	Compactness	Length	Width	Asymmetry	GrooveLength
Area	1,00	0,99	0,61	0,95	0,97	-0,23	0,86
Perimeter	0,99	1,00	0,53	0,97	0,94	-0,22	0,89
Compactness	0,61	0,53	1,00	0,37	0,76	-0,33	0,23
Length	0,95	0,97	0,37	1,00	0,86	-0,17	0,93
Width	0,97	0,94	0,76	0,86	1,00	-0,26	0,75
Asymmetry	-0,23	-0,22	-0,33	-0,17	-0,26	1,00	-0,01
GrooveLength	0,86	0,89	0,23	0,93	0,75	-0,01	1,00

Aréa, Perimeter ,Length ,Width ,GrooveLength corrélés 2 à 2 Compactcness et Asymmetry sont moins corrélé avec les autres variables ont le voit avec leurs nuages de point qui n'ont pas de tendance linéaire, pas de structure particulière

Analyse en Composantes Principales

L'objectif est de décrire sans à priori un jeu de données exclusivements de variables quantitatives, ici l'analyse sera donc effectuée sur 210 individus représentés par les grains de blé , décrits par nos 7 variables. Cette méthode permet de résumer l'information et d'en réduire la dimensionnalité.

Valeurs propres

Les valeurs propres (eigenvalues en anglais) mesurent la quantité de variance expliquée par chaque axe principal. Elles sont grandes pour les premiers axes et petites pour les axes suivants. Les premiers axes correspondent aux directions portant la quantité maximale de variation contenue dans le jeu de données. On peut examiner les valeurs propres pour déterminer le nombre de composantes principales à prendre en compte.

Les valeurs propres peuvent être utilisées pour déterminer le nombre d'axes principaux à conserver. Une valeur propre > 1 : la composante principale (PC) concernée représente plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. On peut regarder le nombre d'axes qui représente une certaine fraction de la variance totale. Par exemple, le nombre d'axe qui permet de parvenir à 70% de la variance totale expliquée.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	5.0312011860	71.87430266	71.87430
Dim.2	1.1975728470	17.10818353	88.98249
Dim.3	0.6780034386	9.68576341	98.66825
Dim.4	0.0683644770	0.97663539	99.64488
Dim.5	0.0187136090	0.26733727	99.91222
Dim.6	0.0053320457	0.07617208	99.98839
Dim.7	0.0008123968	0.01160567	100.00000

Nous voyons ici que **Dim 1** et **Dim 2** ont des valeurs propres > 1. **Dim1** explique à elle seule 71,9%de la variance. En gardant les 2 premières composantes, on explique **89%** de la variance, et avec les 3 premières composantes plus de 97%de la variance. On va donc prendre en compte ces 2 axes.

Variables

L'ACP permet également d'étudier les liaisons linéaires entre les variables.

Les objectifs sont de résumer la matrice des corrélations et de chercher des variables synthétiques : peut-on résumer les observations par un petit nombre de variables ?

Qualité et Contributions

Cos2

le cosinus carré des variables. Représente la qualité de représentation des variables sur le graphique de l'ACP. Il est calculé comme étant les coordonnées au carré : var.cos2=var.coord*var.coord. Avec var$coord les coordonnées des variables pour représenter le nuage de points.

Un cos2 élevé = bonne représentation de la variable sur les axes principaux → la variable est positionnée à proximité du bord du cercle de corrélation (ci dessous, il en est de meme pour la contibution).

Un faible cos2 = la variable n'est pas bien représentée par les axes principaux → la variable est proche du centre du cercle.

Contributions des variables aux axes principaux

Plus la valeur de la contribution est importante, plus la variable contribue à la composante principale en question.

Les contributions des variables sont exprimées en pourcentage.

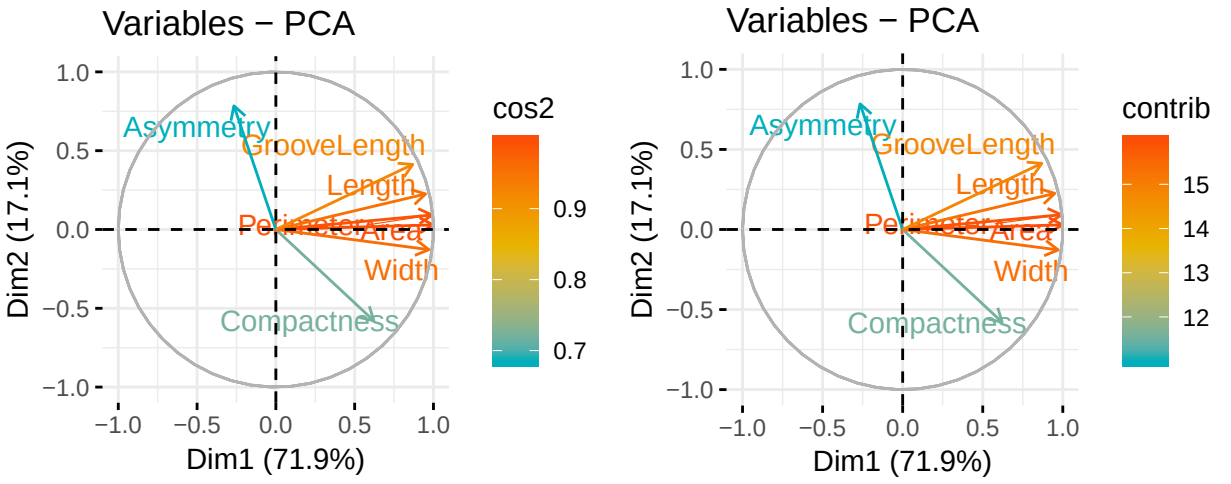
Les variables corrélées avec **Dim.1** et **Dim.2** sont les plus importantes pour expliquer la variabilité dans le jeu de données.

.	cos2Dim.1	cos2Dim.2	cos2Dim.3
Area	**0,99394755**	0,0008450341	0,0004537914
Perimeter	**0,98101057**	0,0084506414	0,0024277403
Compactness	0,38608745	**0,3353216539**	0,2688363174
Length	0,90262715	0,0508079572	0,0304376025
Width	**0,94250493**	0,0163067144	0,0317746687
Asymmetry	0,07087908	**0,6154564499**	0,3130532976
GrooveLength	0,75414445	**0,1703843961**	0,0310200207

.	contribDim.1	contribDim.2	contribDim.3
Area	**19,755671**	0,07056223	0,06693054
Perimeter	**19,498536**	0,70564738	0,35807198
Compactness	7,673862	**28,00010494**	39,65117316
Length	17,940589	4,24257759	4,48929913
Width	**18,733199**	1,36164697	4,68650554
Asymmetry	1,408790	**51,39198433**	46,17281857
GrooveLength	14,989352	**14,22747656**	4,57520109

Cercle de corrélation

Plus une variable est proche du cercle de corrélation, meilleure est sa représentation et elle est plus importante pour interpréter les composantes principales en considération Les variables qui sont proche du centre du graphique sont moins importantes pour interpréter les composantes.



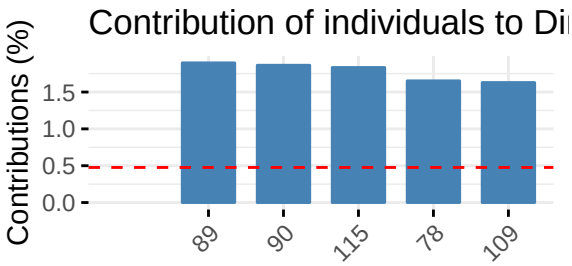
On voit bien que les variables **Area**, **Perimeter** et **Width** sont bien représentées sur l'axe **Dim1** et **GrooveLength**, **Asymmetry** et **Compactness** sont mieux représenté sur l'axe **Dim2** mais avec des cos2 inférieurs à 0.5 pour **GrooveLength** et **Compactness** , on voit sur le cercle que les 3 premieres variables: **Area**, **Perimeter** et **Width** contribuent le mieux sur les axes et ont une meilleur qualité de représentation des variables sur le graphique de l'ACP

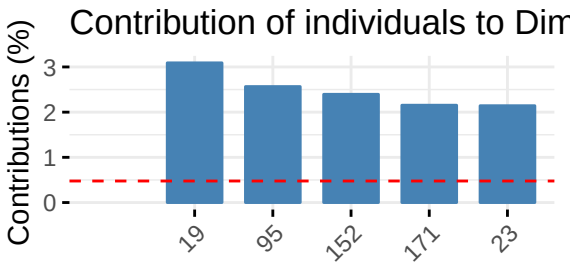
Individu

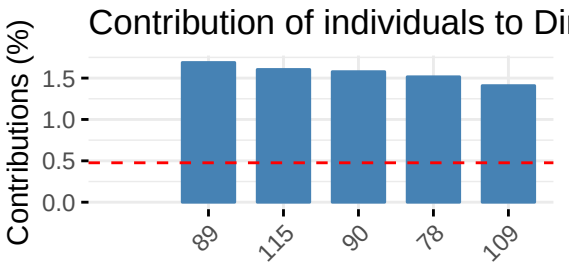
Graphique : qualité et contribution

Comme on a beaucoup trop d'individu regardons les graphiques des individus qur les 5 individus les mieux représsentés

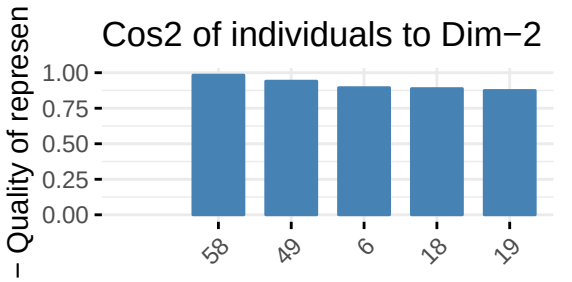
Contribution sur Dim1 et Dim2

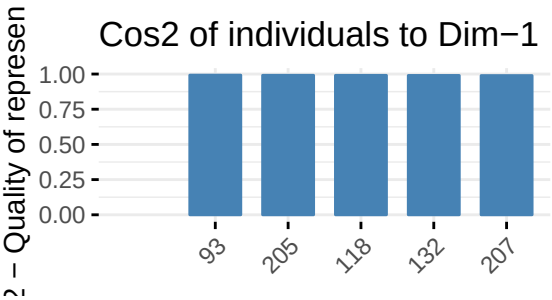


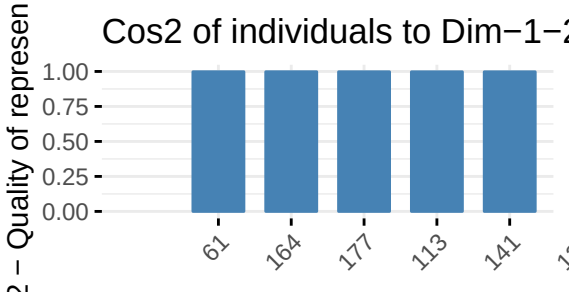




Qualité de représentation







Nous voyons que les grains **89 ,90 et 115** ont une meilleur contribution sur **Dim 1**, **19 95 et 152** sur **Dim 2** mais les 3 premiers contribuent mieux sur les 2 axes en meme temps. Nous voyons que les grains **58,49 et 6**ont unr meilleur qualité de représentation sur **Dim1**, **93 205 et 118** sur **Dim 2** et **61,164 et 177** sur les 2 axes en meme temps.

Conclusion

Nous avons pu synthétiser l'information relative à 210 graines en 2 dimensions à partir des différentes variables avec une précision de **89%**.

Données Crabes

Informations sur le jeu de données:

La base de données sur les crabes comporte 200 lignes et 8 colonnes, décrivant 5 mesures morphologiques sur 50 crabes de deux formes colorées et des deux sexes, de l'espèce Leptograpsus variegatus, collectées à Fremantle, Australie occidentale.

Ce cadre de données contient les colonnes suivantes :

sp - "B" ou "O" pour bleu ou orange.

le sexe comme il est dit.

Index 1:50 à l'intérieur de chacun des quatre groupes.

FL Taille du lobe frontal (mm).

RW largeur arrière (mm).

Longueur de la carapace CL (mm).

Largeur de la carapace CW (mm).

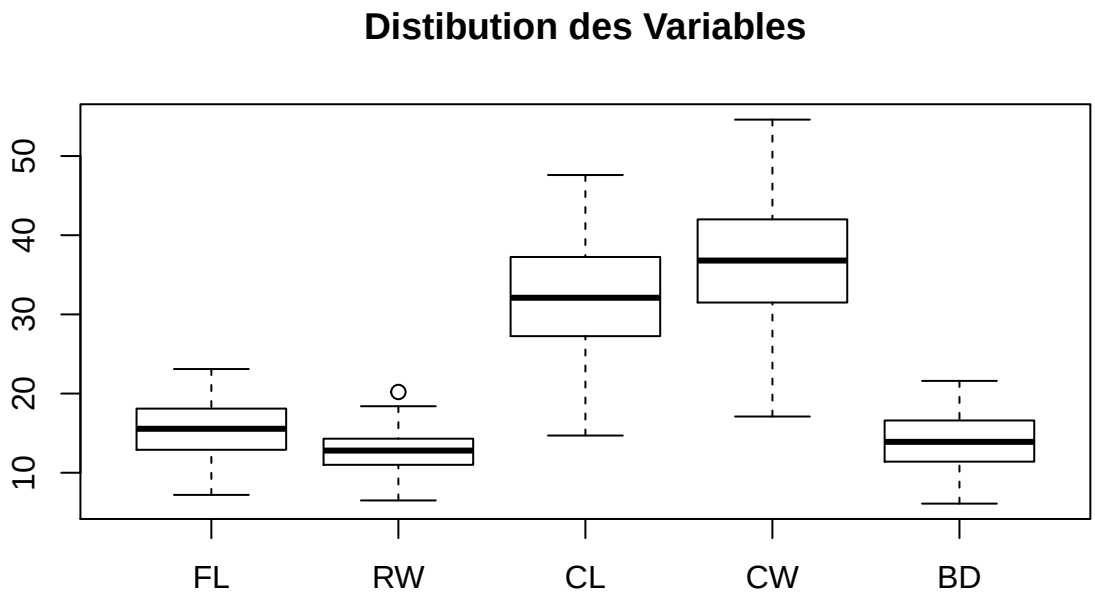
Profondeur du corps BD (mm).

tout au long de notre étude nous allons nous restreindre aux 5 variables **FL, RW, CL, CW et BD**

Nombres de Valeurs manquantes (nvm)

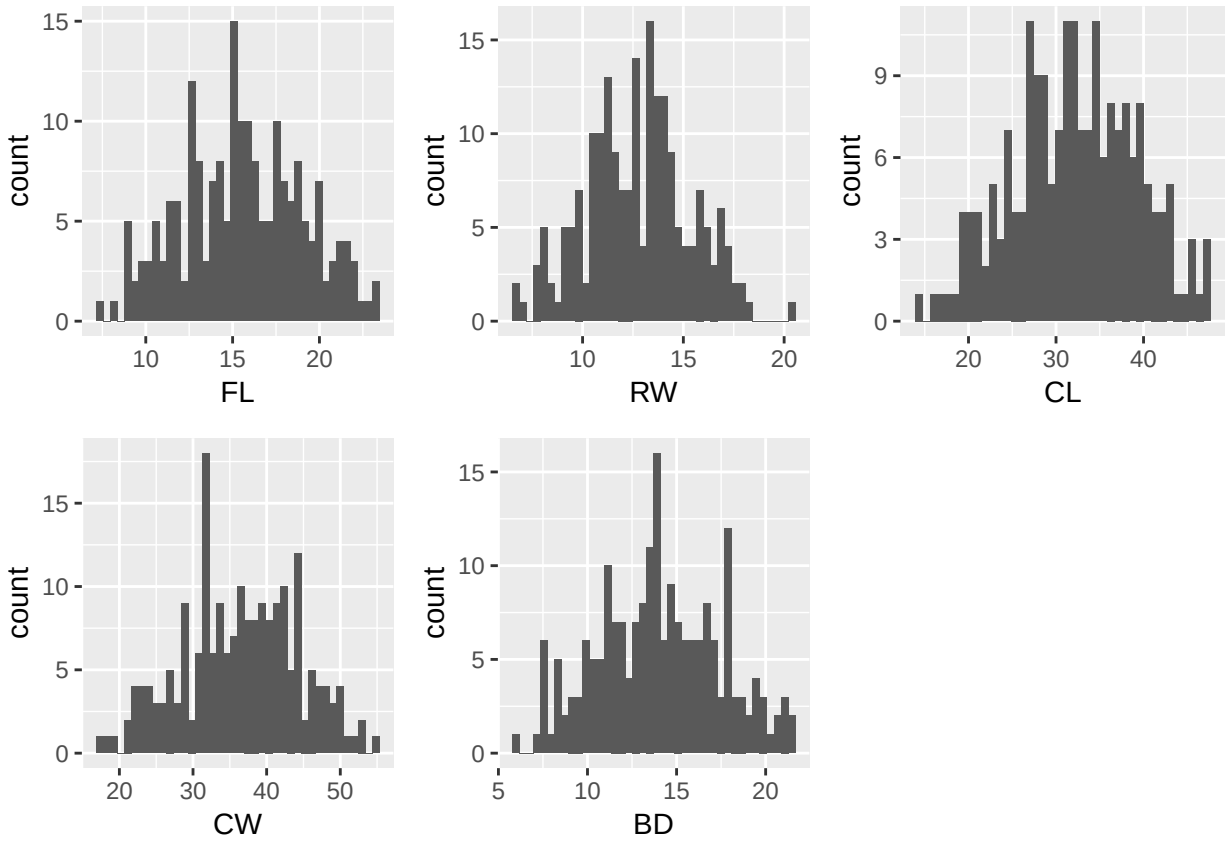
.	FL	RW	CL	CW
nvm	0	0	0	0

Distributions des variables



```
       FL               RW               CL               CW
 Min.   : 7.20   Min.   : 6.50   Min.   :14.70   Min.   :17.10
 1st Qu.:12.90   1st Qu.:11.00   1st Qu.:27.27   1st Qu.:31.50
 Median :15.55   Median :12.80   Median :32.10   Median :36.80
 Mean   :15.58   Mean   :12.74   Mean   :32.11   Mean   :36.41
 3rd Qu.:18.05   3rd Qu.:14.30   3rd Qu.:37.23   3rd Qu.:42.00
 Max.   :23.10   Max.   :20.20   Max.   :47.60   Max.   :54.60
       BD
 Min.   : 6.10
 1st Qu.:11.40
 Median :13.90
 Mean   :14.03
 3rd Qu.:16.60
 Max.   :21.60
```

Histogramme avec Ggplot



CL et **CW** ont une plus grande variabilité les boxplots semblent symétrique et montrent les différences des distributions; cependant le nombre de variable n'étant pas assez grand on ne peut pas émettre d'hypothèse concrète sur la normalité meme avec les histogrammes qui semblent etre symétrique

Corrélation entre nos variables et nuages de point

.	FL	RW	CL	CW	BD
FL	1,0000000	0,9069876	0,9788418	0,9649558	0,9876272
RW	0,9069876	1,0000000	0,8927430	0,9004021	0,8892054
CL	0,9788418	0,8927430	1,0000000	0,9950225	0,9832038
CW	0,9649558	0,9004021	0,9950225	1,0000000	0,9678117
BD	0,9876272	0,8892054	0,9832038	0,9678117	1,0000000

On remarque que toutes nos variables sont bien corrélées entre elles

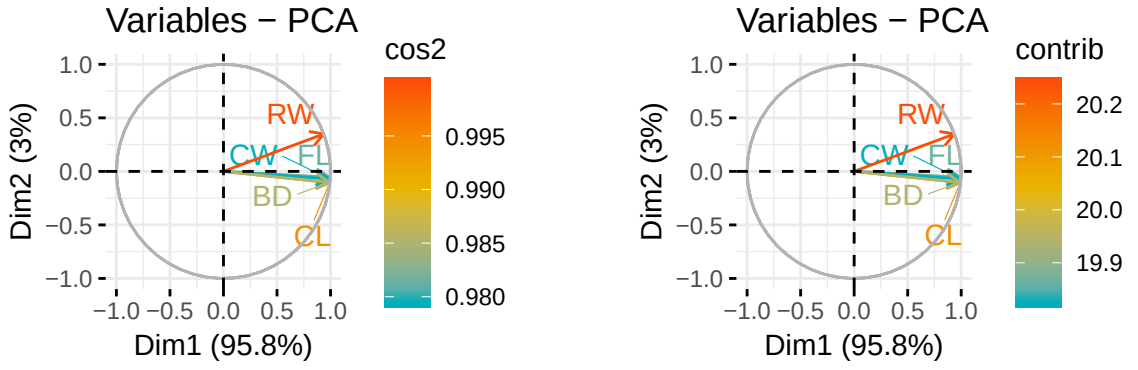
Analyse en Composantes Principales

```
       eigenvalue variance.percent cumulative.variance.percent
Dim.1 4.788834784      95.77669569                    95.77670
Dim.2 0.151685207       3.03370413                    98.81040
Dim.3 0.046632974       0.93265948                    99.74306
Dim.4 0.011135357       0.22270714                    99.96577
Dim.5 0.001711678       0.03423355                   100.00000
```

Nous voyons ici que **Dim 1** à une valeur propres > 1. **Dim1** explique à elle seule 95.7% de la variance. En gardant les 2 premières composantes, on explique **98%** de la variance, et avec les 3 premières composantes plus de **99%** de la variance. On va donc prendre en compte ces 2 axes.

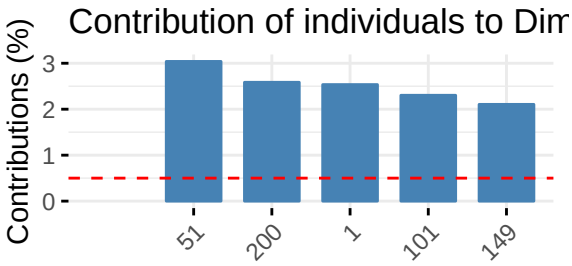
Variables

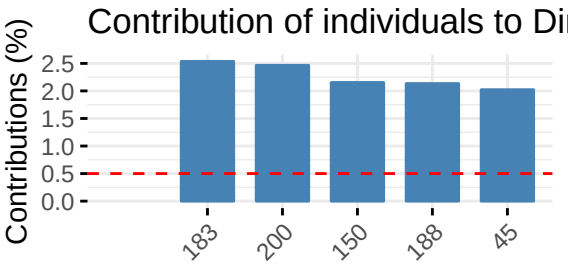
.	cos2Dim.1	cos2Dim.2	contribDim.1	contribDim.2
FL	**0,9785672**	0,002871189	**20,43435**	1,892860
RW	0,8775551	**0,122355169**	18,32502	**80,663877**
CL	**0,9835409**	0,010914009	**20,53821**	7,195170
CW	0,9745407	0,004947193	20,35027	3,261487
BD	**0,9746309**	0,010597646	**20,35215**	6,986605

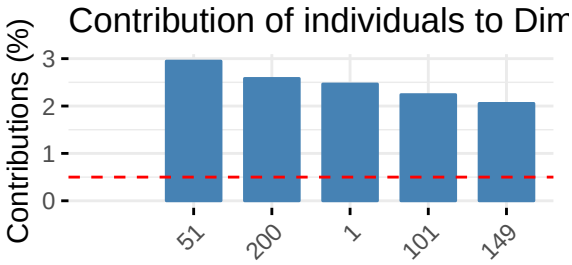


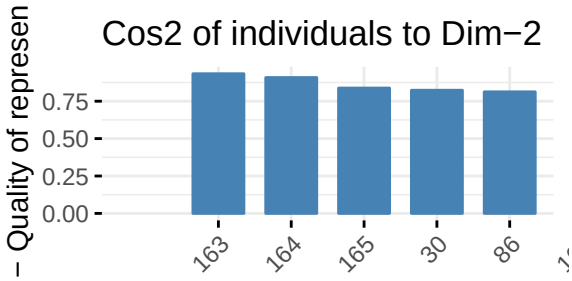
Au vue du tableau et des cercles de coorrélation ont voit que les variables **RW** et **CL** présentent une meilleure contribution et qualité de représentation sur nos 2 axes

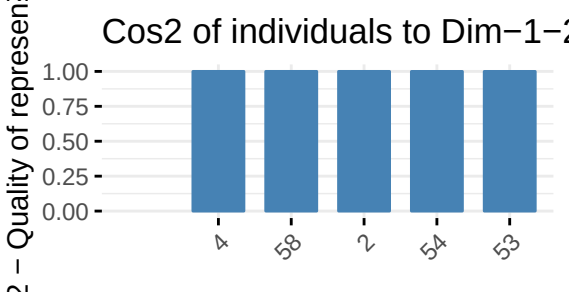
Graphes des Individus Contribution et qualité

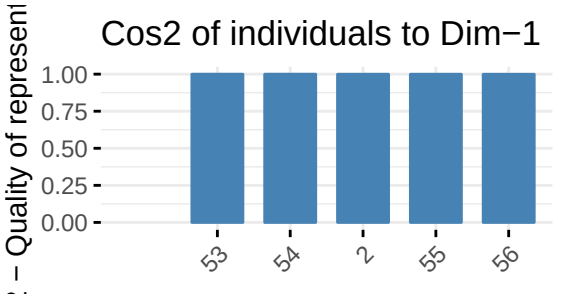












Nous voyons que les grains **89 ,90 et 115** ont une meilleur contribution sur **Dim 1**, **19 95 et 152** sur **Dim 2** mais les 3 premiers contribuent mieux sur les 2 axes en meme temps. Nous voyons que les grains **58,49 et 6**ont unr meilleur qualité de représentation sur **Dim1**, **93 205 et 118** sur **Dim 2** et **61,164 et 177** sur les 2 axes en meme temps.

Conclusion

Nous avons pu synthétiser l'information relative à 210 graines en 2 dimensions à partir des différentes variables avec une précision de **89%**.

Controverse sur la théorie de l'hérédité de Mendel

Danss cette partie j'ai seulemennt traité le jeux de données cotyledon

Test d'ajustement au paramètre d'une loi binomiale sans approximation normale de la binomiale.

Pour ce test selon la colonne Set on effectue un test sur chaque selon la variable **Set** on utilise la commmande binom.test(p,n) n, notre échantillon total est de 122 qui correspond à la longueur de notre colonne et p la proportion de chacun des 16 groupes de graine on obtient donc la probabilité de succès alternative. Hypothèse: H0 la probabilité de succes est égal à 0,5 et H1 le contraire

Classe	Proportion /122	Probabilité de Succès
1	8	0,06557377
2	7	0,05737701
3	4	0,03278689
4	6	0,04918033
5	11	0,09016393
6	8	0,06557377
7	7	0,05737705
8	9	0,07377049
9	10	0,08196721
10	9	0,07377049
11	3	0,02459016
12	10	0,08196721
13	9	0,07377049
14	7	0,05737705
15	9	0,07377049
16	5	0,04098361

et toutes les pvalue sont telle que p-value $<$ 2.2e-16. On rejette Ho et on accepte H1

Test de Kolmogorov Smirnov

il s'agit d'un test pour tester qu'une statistique bien choisie suit une loi normale N (0, 1) il s'agit de notre hypothèse H0. Mais cette statistiques (échantillon) ne doit pas avoir de doublon . Ici nous avons fait les test de kolmogorov de nos 16 groupes pour les graines à caractere dominant et récursif

POur les **Dominants** On obtient des p-values inférieur à 0.05 donc l'Hypothèse H0 est écartée. Pour les **récessifs** le **11** ème groupe a une p-value de 0.07 environ ce qui est supérieur l'hypotèse H0 n'est pas rejetée

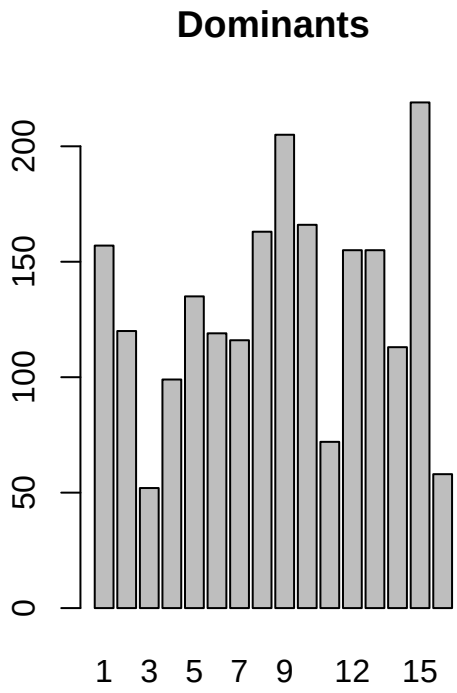
Test du khi2

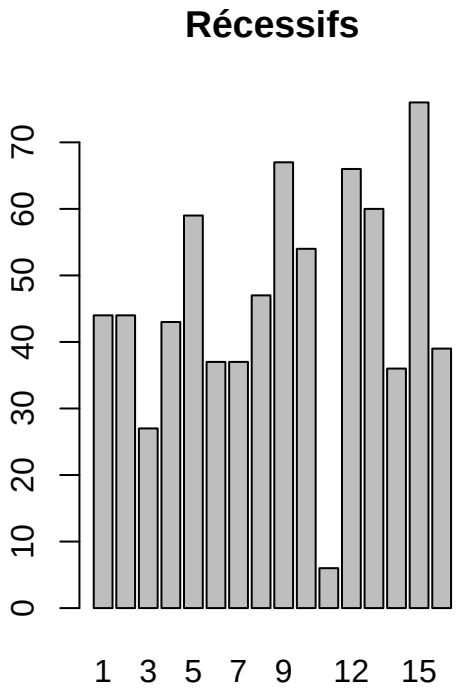
On étudie l'influence du caractère dominants ou récessifs sur les ensembles de graines grouperpar la variable Set

Pour tester l'indépendance de deux variables qualitatives, on teste l'hypothèse nulle H0: "les deux variables sont indépendantes" contre l'hypothèse alternative H1 : "les deux variables ne sont pas indépendantes". Pour cela, on construit la statistique de test suivante :

On va effectuer un test de chi 2 sur les proportions (parmi toutes les graines) de caractères dominants et récessifs qui sont ici nos 2 variables qualitatives.

```
            1   2  3  4   5   6   7   8   9  10 11  12  13  14  15 16
Dominants 157 120 52 99 135 119 116 163 205 166 72 155 155 113 219 58
Récessifs  44  44 27 43  59  37  37  47  67  54  6  66  60  36  76 39
```






```
    Pearson's Chi-squared test

data:  tab
X-squared = 36.587, df = 15, p-value = 0.001453
```

On obtient une p-value inférieur à 0.05 on rejette donc H0 au profit de H1.