TP ANOVA

ENSAE de Dakar
ISE 2
2024-2025

Les auteurs

NOTE

Ce travail a ete realise par xxxxxxx et xxxxxx

Table des matières

Chapitre 1

Presentation et statistiques descriptives

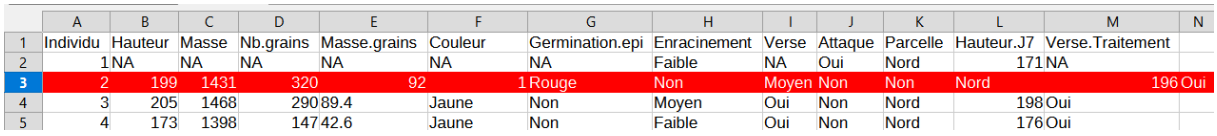
Presentation des données

Contexte et justification

Nous disposons de 13 variables dont une (la variable individus) qui fait office d'identifiant unique.Notre base contient 100 individus.Il s'agit de …… (description des variables,…)

Premiére inspection

Une premiere inspection de la base nous as conduit a modifier l'individus numero 2 .En effet les valeurs prises par cette individus pour des variables comme **Germination.epi** ou **Enracinement** ne sont pas coherente. ##



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Individu	Hauteur	Masse	Nb.grains	Masse.grains	Couleur	Germination.epi	Enracinement	Verse	Attaque	Parcelle	Hauteur.J7	Verse.Traitement	
2	1	NA	NA	NA	NA	NA	NA	Faible	NA	Oui	Nord	171	NA	
3	2	199	1431	320	92	1	Rouge	Non	Moyen	Non	Non	Nord	196	Oui
4	3	205	1468	290	89.4	Jaune	Non	Moyen	Oui	Non	Nord	198	Oui	
5	4	173	1398	147	42.6	Jaune	Non	Faible	Oui	Non	Nord	176	Oui	

Premiére inspection

Nous avons juge que cela est due a des problemes lors du saisi des donnée. nous vons corrige par :

Valeurs de l'individu numéro 2 par variable

Individu	Hauteur	Masse	Nb.grains	Masse.grains	Couleur	Germination.epi	Enracinement
2	199	1431	320	921	Rouge	Non	Moyen

Analyse descriptive

Generalite sur les variables quantitatives

Nous disposons de ….

Statistiques des Variables Quantitatives

Variable	valeur_NA	Moyenne	Variance
Individu	0	50.50000	841.6667
Hauteur	3	259.36082	1965.8580
Masse	3	1811.61856	102708.9051
Nb.grains	3	292.63918	10283.1914
Masse.grains	3	96.54742	8073.7931
Hauteur.J7	0	257.36000	1934.8994

Figure 1

Generalite sur les variables qualitative

Quant aux variables qualitatives

Variables Qualitatives, Facteurs, Fréquences et Nombre de Valeurs Manquantes

Variable	Facteurs	Fréquences	Valeurs_manquantes
Couleur	NA, Rouge, Jaune, Jaune.rouge	Jaune (48), Jaune.rouge (22), Rouge (29)	1
Germination.epi	NA, Non, Oui	Non (90), Oui (9)	1
Enracinement	Faible, Moyen, Tres.fort, Fort	Faible (19), Fort (26), Moyen (28), Tres.fort (27)	0
Verse	NA, Non, Oui	Non (57), Oui (42)	1
Attaque	Oui, Non	Non (54), Oui (46)	0
Parcelle	Nord, Sud, Est, Ouest	Est (33), Nord (17), Ouest (36), Sud (14)	0
Verse.Traitement	NA, Oui, Non	Non (44), Oui (55)	1

Figure 2

Teste de corrélation entre la masse et la quantité de grain

Ce teste permet de ….

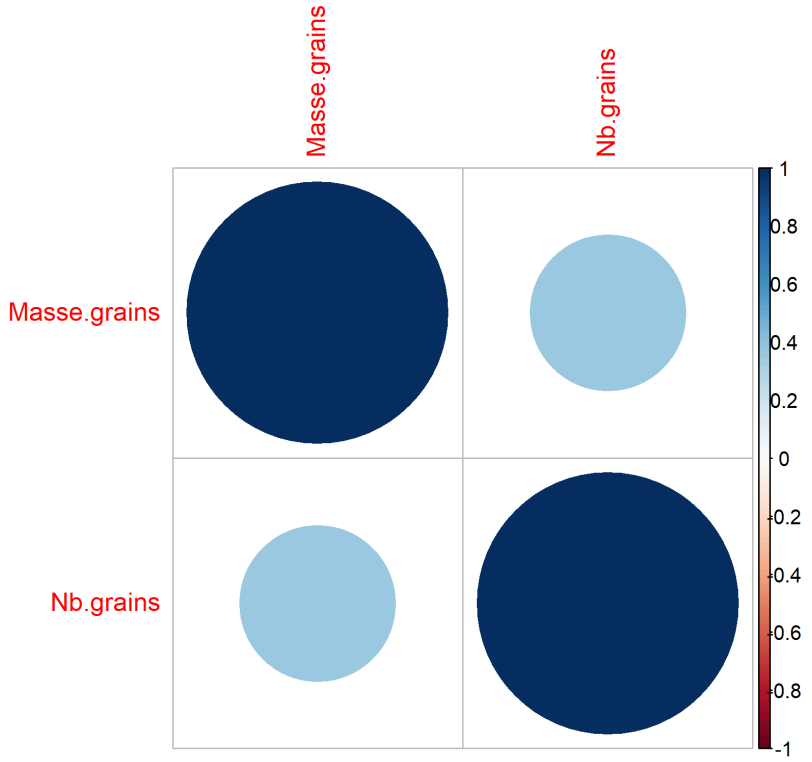


Figure 3

Le graphique montre une corrélation positive modérée entre **Masse.grains** et **Nb.grains**, indiquant que lorsque l'un augmente, l'autre tend également à augmenter. La force de cette relation est représentée par un cercle bleu clair, suggérant une liaison existante mais non parfaite.

Poursuivons par un autre teste pour voir la corrélation de rangs a defaut d'une correlation lineaire importante

Teste de corrélation de Spearman entre la masse et la quantité de grain

```
R> cor.test(data$Masse.grains, data$Nb.grains, method = "spearman")
```

```
    Spearman's rank correlation rho

data:  data$Masse.grains and data$Nb.grains
S = 4681.2, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9692218
```

Figure 4

Le coefficient de corrélation de Spearman est de 0,969, ce qui indique une très forte corrélation positive entre les deux variables bien que le teste de correlation lineaire indique une non lineaire relation entre nos differents variables. La p-value est inferieure a 0,01 ce qui indique que la correlation est hautement significative .nous rejetons l'hypothèse nulle selon laquelle il n'y a pas de relation entre les deux variables, et il existe donc une forte probabilité que la relation observée soit réelle et non due au hasard.

Conclusion

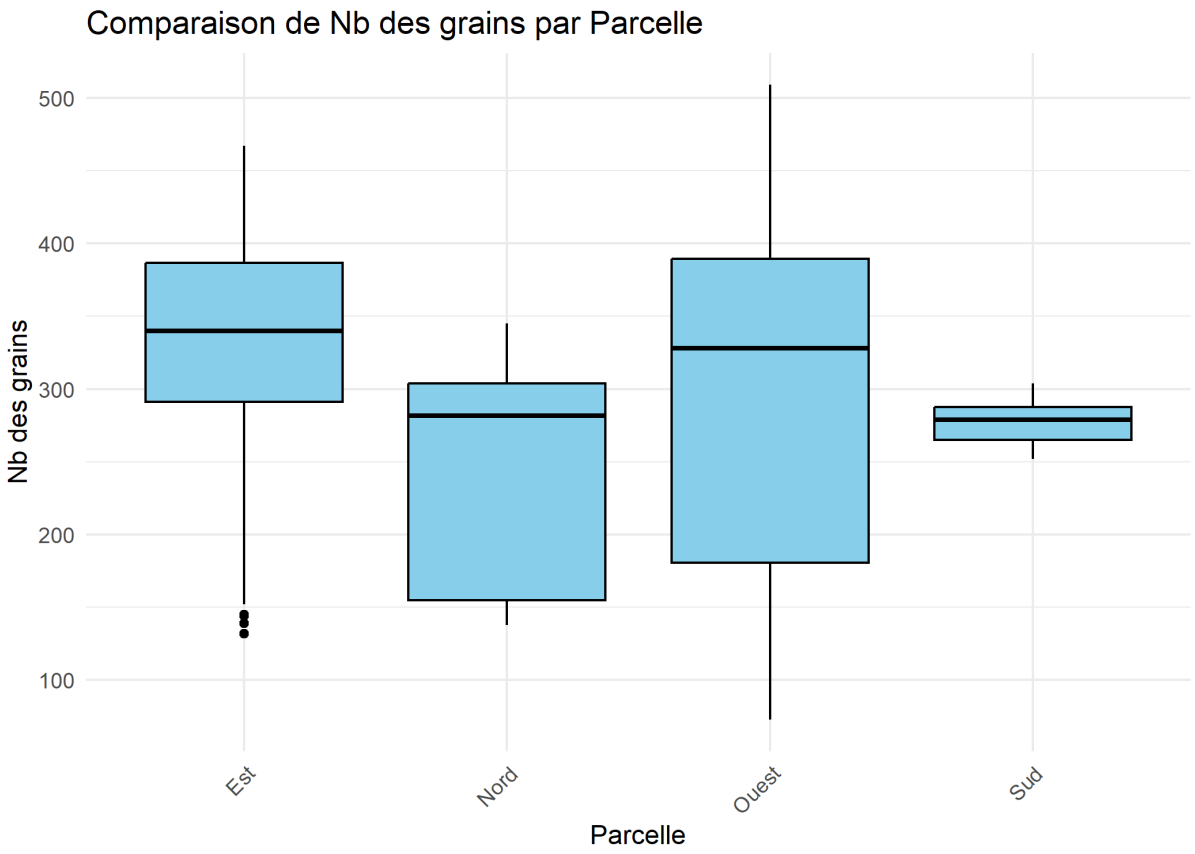
IMPORTANT

Compte tenu de la très forte corrélation entre masse des grains et Nombre de grains, nous pouvons conclure qu'une nombre élevée des grains est fortement associée à un nombre élevé de grains. En raison de cette forte association, nous choisissons le nombre de grains comme variable cible pour notre ANOVA.

La variable **nombre.grain** a trois valeur manquantes ce qui est assez faible,nous traiteront les valeurs manquantes apres choix des variables categorielles.

Analyse bivarie

Comparaison entre “Parcelle” et “Nb.grains”



On observe que

informations du parcelle en fonction de le nombre

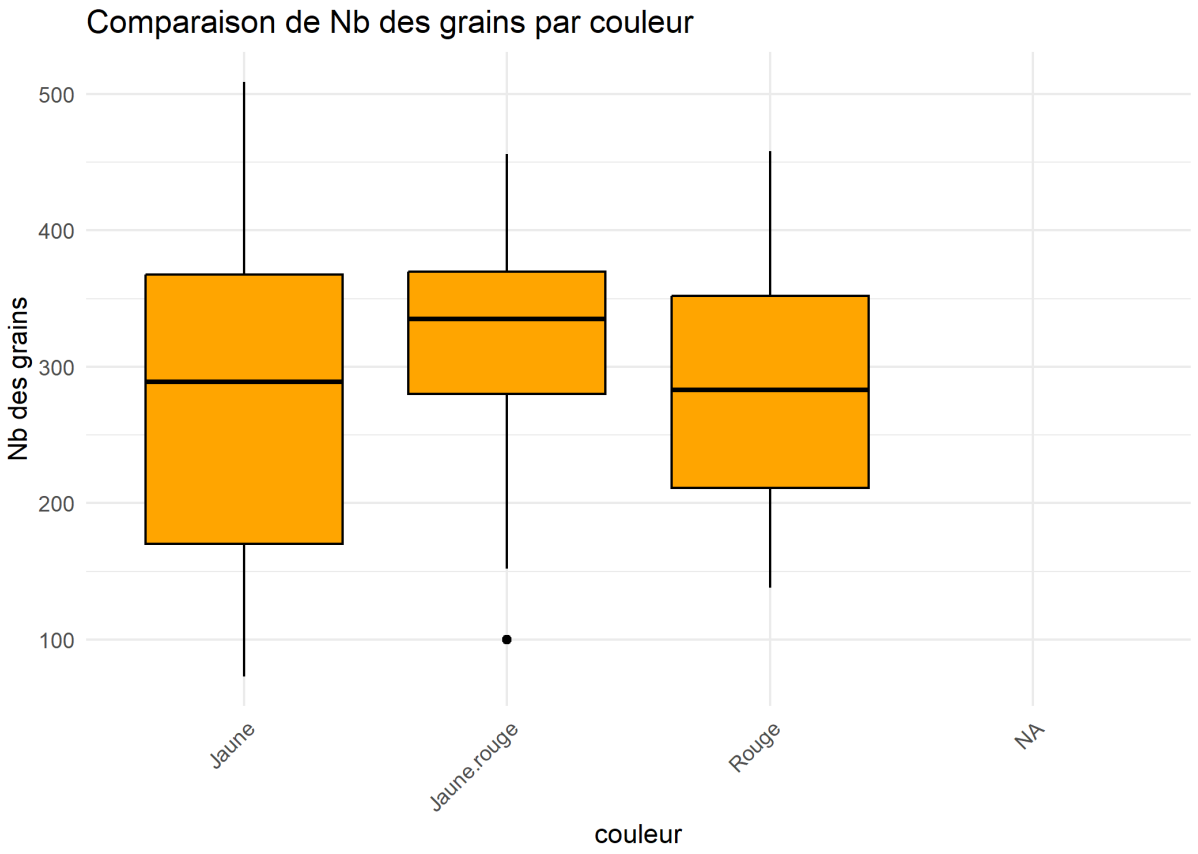
Individu	Parcelle	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	Nord	NA	TRUE	NA
32	Est	NA	TRUE	NA
65	Ouest	NA	TRUE	NA

Rapport de correlation

```
[1] 0.08406623
```

Le rapport de correlation entre ces deux variables indique que

Comparaison entre “couleur” et “Nb.grains”



On observe que

informations de la couleur en fonction de le nombre

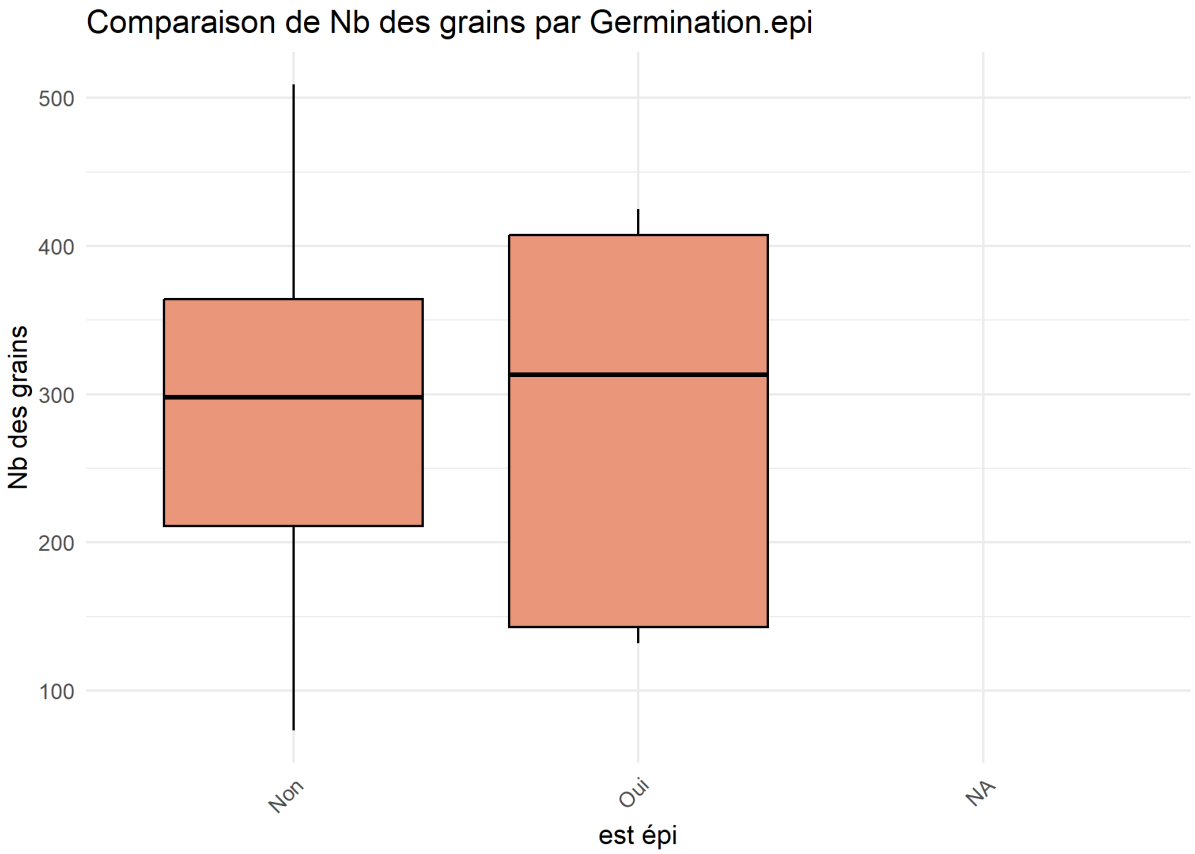
Individu	Couleur	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	NA	NA	TRUE	NA
32	Jaune	NA	TRUE	NA
65	Jaune	NA	TRUE	NA

Rapport de correlation

```
[1] 0.01083218
```

Le rapport de correlation entre ces deux variables indique que

Comparaison entre “épigité de la germination” et “Nb.grains”



On observe que

informations d’épigité de la graine en fonction de le nombre

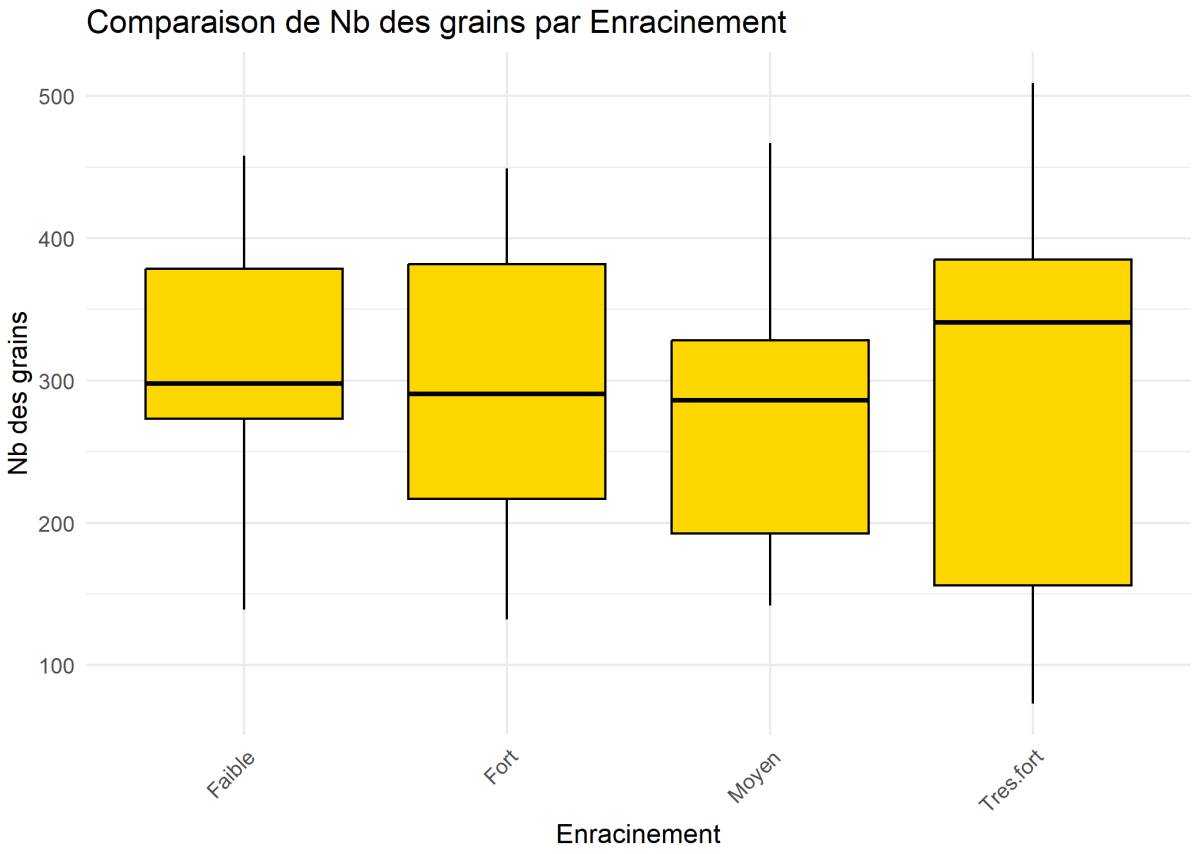
Individu	Germination.epi	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	NA	NA	TRUE	NA
32	Oui	NA	TRUE	NA
65	Non	NA	TRUE	NA

Rapport de correlation

```
[1] 0.0003605472
```

Le rapport de correlation entre ces deux variables indique que

Comparaison entre “Enracinement” et “Nb.grains”



On observe que

informations d'Enracinement en fonction de le nombre

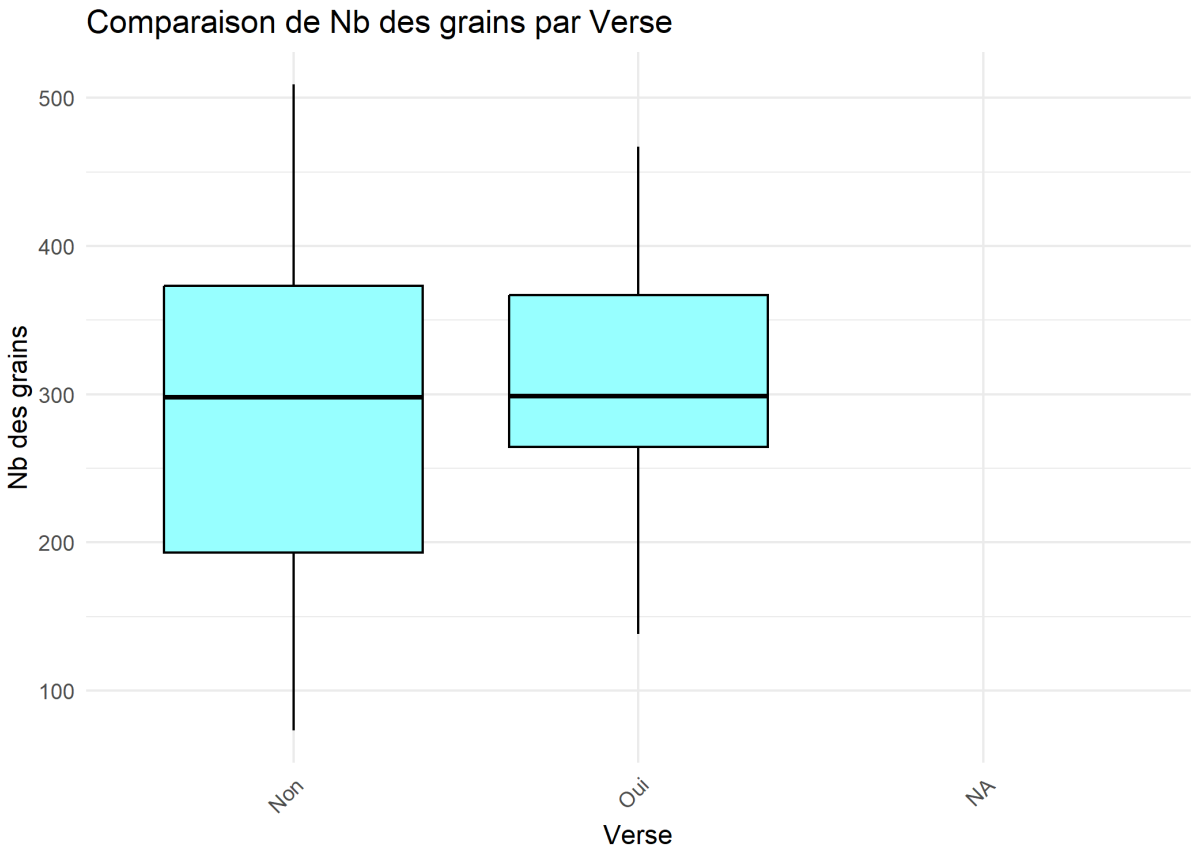
Individu	Enracinement	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	Faible	NA	TRUE	NA
32	Tres.fort	NA	TRUE	NA
65	Tres.fort	NA	TRUE	NA

Rapport de correlation

```
[1] 0.01728187
```

Le rapport de correlation entre ces deux variables indique que

Comparaison entre “courbure (verse) de la plante” et “Nb.grains”



On observe que

informations de Verse en fonction de le nombre

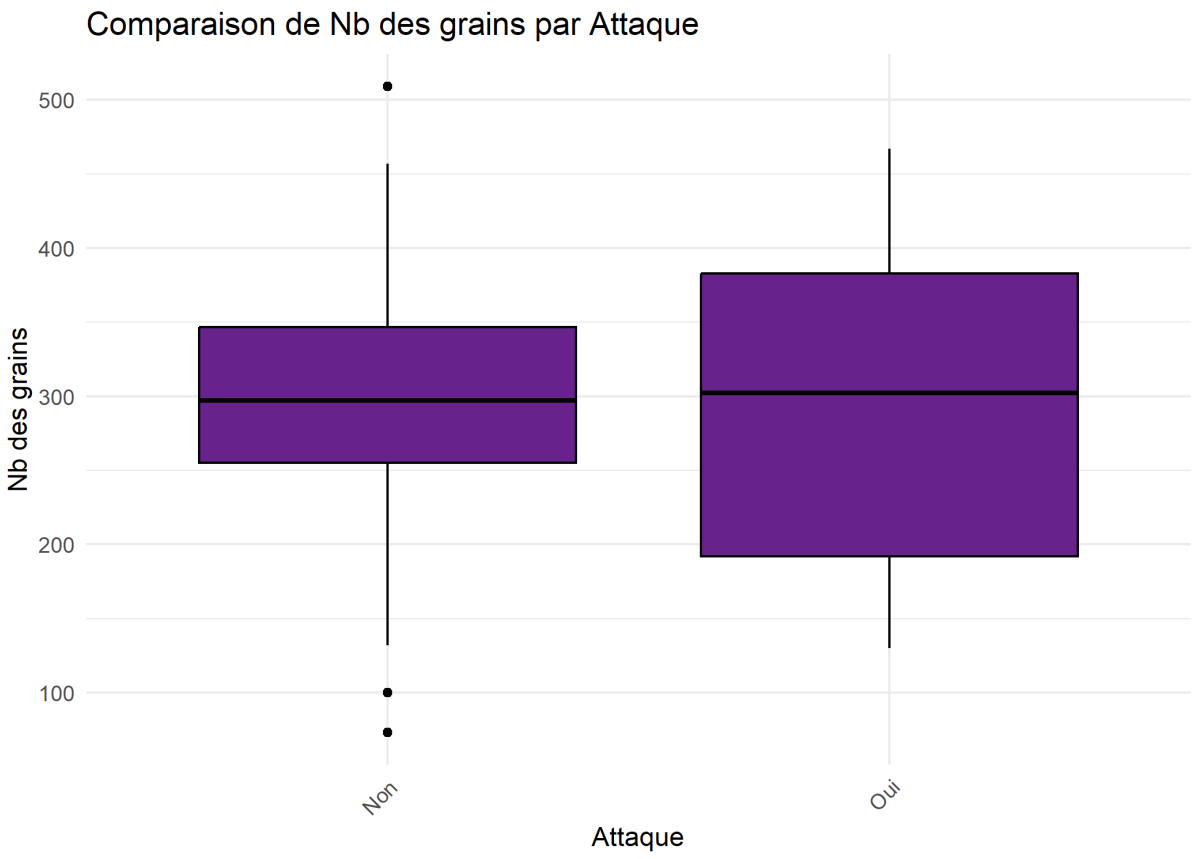
Individu	Verse	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	NA	NA	TRUE	NA
32	Non	NA	TRUE	NA
65	Non	NA	TRUE	NA

Rapport de correlation

```
[1] 0.0005838208
```

Le rapport de correlation entre ces deux variables indique que

Comparaison entre ” est attaqué par des insectes” et “Nb.grains”



On observe que

informations d'Attaque en fonction de le nombre

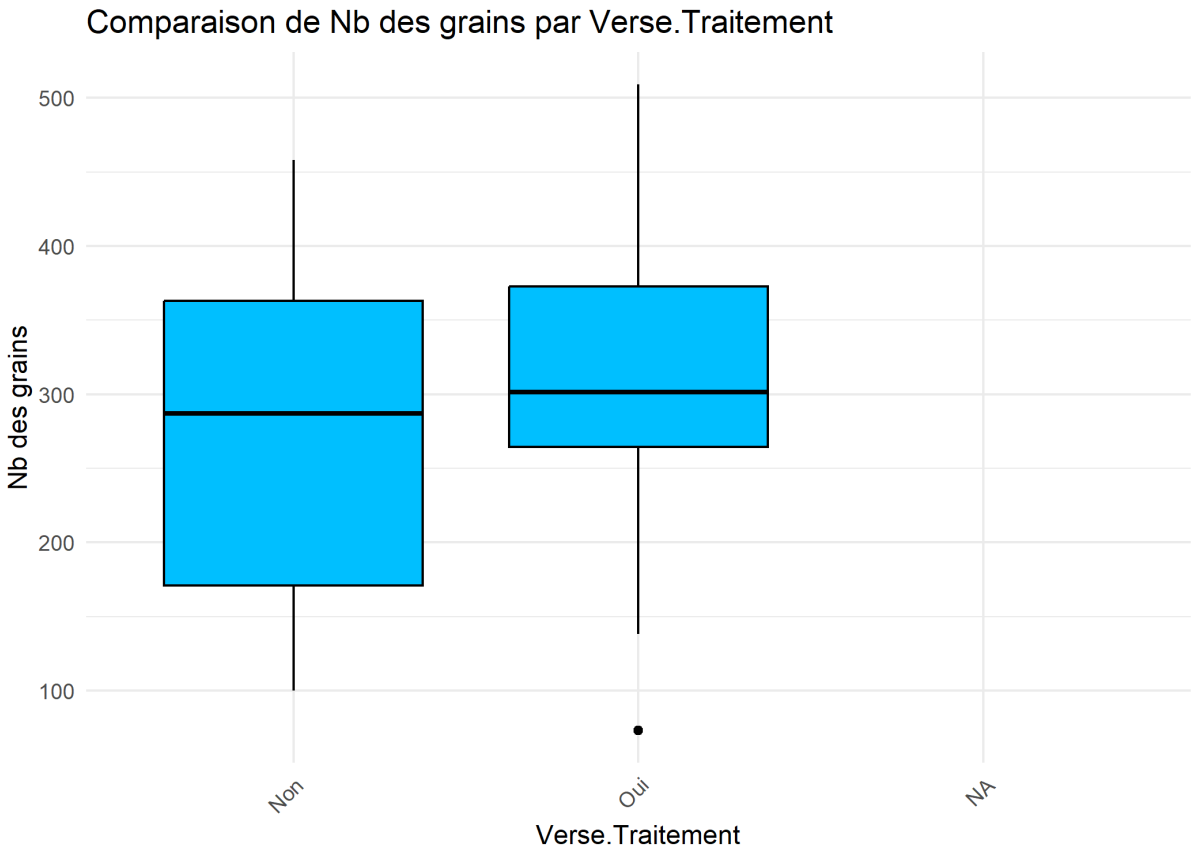
Individu	Attaque	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	Oui	NA	TRUE	NA
32	Non	NA	TRUE	NA
65	Non	NA	TRUE	NA

Rapport de correlation

```
[1] 0.0003195651
```

Le rapport de correlation entre ces deux variables indique que

Comparaison entre ” la courbure est traité” et “Nb.grains”



On observe que

informations courbure traité en fonction de le nombre

Individu	Verse.Traitement	Nb.grains	Nb_grains_missing	Nb_grains_aberrant
1	NA	NA	TRUE	NA
32	Non	NA	TRUE	NA
65	Oui	NA	TRUE	NA

Rapport de correlation

```
[1] 0.01254531
```

Le rapport de correlation entre ces deux variables indique que

Présentation et Philosophie

WEBIN-R

Ce chapitre est évoqué dans le webin-R #01 (premier contact avec R & RStudio) sur YouTube.

Présentation de R

R est un langage orienté vers le traitement de données et l'analyse statistique dérivé du langage **S**. Il est développé depuis une vingtaine d'années par un groupe de volontaires de différents pays. C'est un logiciel libre[1], publié sous licence GNU GPL.

L'utilisation de **R** présente plusieurs avantages :

- c'est un logiciel multiplateforme, qui fonctionne aussi bien sur des sytèmes **Linux**, **Mac OS X** ou **Windows** ;
- c'est un logiciel libre, développé par ses utilisateurs et modifiable par tout un chacun ;
- c'est un logiciel gratuit ;
- c'est un logiciel très puissant, dont les fonctionnalités de base peuvent être étendues à l'aide de plusieurs milliers d'extensions ;
- c'est un logiciel dont le développement est très actif et dont la communauté d'utilisateurs ne cesse de s'élargir ;
- les possibilités de manipulation de données sous **R** sont en général largement supérieures à celles des autres logiciels usuels d'analyse statistique ;
- c'est un logiciel avec d'excellentes capacités graphiques et de nombreuses possibilités d'export ;
- avec **Rmarkdown**[2], il est devenu très aisé de produire des rapports automatisés dans divers

1. Pour plus d'informations sur ce qu'est un logiciel libre, voir : http://www.gnu.org/philosophy/free-sw.fr.html.

format (**Word**, **PDF**, **HTML**, …) ;
- **R** est de plus utilisé dans tous les secteurs scientifiques, y compris dans le domaine des analyses d'enquêtes et, plus généralement, des sciences sociales.

Comme rien n'est parfait, on peut également trouver quelques inconvénients :

- le logiciel, la documentation de référence et les principales ressources sont en anglais. Il est toutefois parfaitement possible d'utiliser **R** sans spécialement maîtriser cette langue ;
- il n'existe pas encore d'interface graphique pour **R** équivalente à celle d'autres logiciels comme **SPSS** ou **Modalisa**. **R** fonctionne à l'aide de scripts (des petits programmes) édités et exécutés au fur et à mesure de l'analyse et se rapprocherait davantage de **SAS** dans son utilisation (mais avec une syntaxe et une philosophie très différentes). Ce point, qui peut apparaître comme un gros handicap, s'avère après un temps d'apprentissage être un mode d'utilisation d'une grande souplesse ;
- comme **R** s'apparente davantage à un langage de programmation qu'à un logiciel proprement dit, la courbe d'apprentissage peut être un peu « raide », notamment pour ceux n'ayant jamais programmé auparavant.

Il est à noter que le développement autour de **R** a été particulièrement actif ces dernières années. On trouvera dès lors aujourd'hui de nombreuses extensions permettant de se « faciliter la vie » au quotidien, ce qui n'était pas vraiment encore le cas il y a 5 ans.

Philosophie de R

Quelques points particuliers dans le fonctionnement de **R** peuvent parfois dérouter les utilisateurs habitués à d'autres logiciels :

- Sous **R**, en général, on ne voit pas directement les données sur lesquelles on travaille ; on ne dispose pas en permanence d'une vue des données sous forme de tableau[3], comme sous **Modalisa ou SPSS**. Ceci peut être déroutant au début, mais on se rend vite compte qu'on n'a pas besoin de voir en permanence les données pour les analyser.
- Alors qu'avec la plupart des logiciels on réfléchira avec un fichier de données ouvert à la fois, sous **R** chaque fichier de données correspondra à un objet différent chargé en mémoire, permettant de manipuler très facilement plusieurs objets à la fois (par exemple dans le cadre de fusion de tables[4]).
- Avec les autres logiciels, en général la production d'une analyse génère un grand nombre de résultats de toutes sortes dans lesquels l'utilisateur est censé retrouver et isoler ceux qui l'intéressent. Avec **R**, c'est l'inverse : par défaut l'affichage est réduit au minimum et c'est l'utilisateur qui demande à voir des résultats supplémentaires ou plus détaillés.

---

2. Voir http://rmarkdown.rstudio.com/.

3. On verra qu'il est possible avec **RStudio** de disposer d'une telle vue.

4. Voir par exemple la section dédiée à ce sujet dans le chapitre sur la manipulation de données.

- Sous **R**, les résultats des analyses sont eux aussi stockés dans des objets et sont dès lors manipulables.

Inhabituel au début, ce fonctionnement permet en fait assez rapidement de gagner du temps dans la conduite des analyses.

Présentation de RStudio

L'interface de base de **R** est assez rudimentaire (voir figure ci-après).

RStudio est un environnement de développement intégré libre, gratuit, et qui fonctionne sous **Windows**, **Mac OS X** et **Linux**. Il complète **R** et fournit un éditeur de script avec coloration syntaxique, des fonctionnalités pratiques d'édition et d'exécution du code (comme l'autocomplétion), un affichage simultané du code, de la console **R**, des fichiers, graphiques et pages d'aide, une gestion des extensions, une intégration avec des systèmes de contrôle de versions comme **git**, etc. Il intègre de base divers outils comme par exemple la production de rapports au format **Rmarkdown**. Il est en développement actif et de nouvelles fonctionnalités sont ajoutées régulièrement. Son principal défaut est d'avoir une interface uniquement anglophone.

Pour une présentation plus générale de **RStudio** on pourra se référer au site du projet : http://www.rstudio.com/.

RStudio peut tout à fait être utilisé pour découvrir et démarrer avec **R**. Les différents chapitres d'**analyse-R** partent du principe que vous utilisez **R** avec **RStudio**. Cependant, à part les éléments portant sur l'interface de **RStudio**, l'ensemble du code et des fonctions **R** peuvent être utilisés directement dans **R**, même en l'absence de **RStudio**.

La documentation de **RStudio** (en anglais) est disponible en ligne à https://support.rstudio.com. Pour être tenu informé des dernières évolutions de **RStudio**, mais également de plusieurs extensions développées dans le cadre de ce projet, vous pouvez suivre le blog dédié http://blog.rstudio.org/.