Table des matières

Chapitre 1

Statistiques Unidimensionnelle

Les statistiques descriptives sont utilisées pour résumer les données de manière à donner un aperçu des informations contenues dans celles-ci.

1.1 Définitions

1.1.1 Statistiques

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. Il ne faut pas confondre la statistique qui est la science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis. Nous distinguons :

- La statistique descriptive concerne la synthèse des données. Nous avons un jeu de données et nous aimerions le décrire de plusieurs manières. Habituellement, cela implique de calculer des mesures descriptives, telles que pourcentages, sommes, moyennes,... etc.

- La statistique inférentielle fait plus. Il y a une inférence associée à l'ensemble de données, une conclusion tirée à propos de la population à l'origine des données.

1.1.2 Population

Un ensemble d'objets ou de personnes d'une étude statistique est appelé population. L'ensemble sur lequel porte l'activité statistique s'appelle la population.

1.1.3 Échantillon

Un échantillon est un sous-ensemble de n individus (d'effectif n) extraits de la population pour lesquels on a mesuré (observé) un caractère quantitatif ou qualitatif. L'obtention d'un échantillon est une étape clé dans la démarche statistique car cet ensemble de données permet de conclure sur les propriétés d'une population donnée. Si l'échantillon est extrait dans de "bonnes conditions" on parlera d'échantillon représentatif de la population.

1.1.4 Individus

Les éléments de la population sont appelés individus ou unités statistiques. Un individu doit être défini sans ambiguité. Pour chaque individu, on dispose d'une ou plusieurs observations. Les individus d'une population peuvent être de nature très diverses : Souvent, un être humain, parfois aussi un animal, une plante ou autre chose.

1.1.5 Variable statistique

C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative. (âge, sexe, taille, poids, nombre de plantes, etc.).

1.1.6 Types de données

1. Données numériques
   (a) Données discrètes :
   La variable ne peut prendre que des valeurs entières (0, 1, 2, etc.) exemples : nombre de malades, nombre fleurs, etc.
   (b) Données continues :
   Toutes les valeurs réelles (souvent comprises dans une certaine classe) sont possible par exemple : poids, âge, durée de vie, etc.
2. Données qualitatives (non numériques, catégoriques)
   (a) Données nominales : catégories non ordonnées
   exemples : groupe sanguin, couleur des yeux
   (b) Données ordinales ou ordonnées : catégories ordonnées
   exemples : niveau de tabagisme, attitudes (bon-modéré-mauvais)
   Le codage numérique des données nominales ou ordonnées ne rend pas les données numériques !

1.2 Paramètres de la statistique descriptive

En général une série statistique à caractère discret se présente sous la forme :

Valeurs	x_1	x_2	.....	x_p
Effectifs	n_1	n_2	.....	n_p
Fréquences	f_1	f_2	.....	f_p

On écrira souvent : la série (x_i, n_i). (On n'indique pas le nombre de valeurs lorsqu'il n'y a pas d'ambiguité). Souvent on notera N l'effectif total de cette série donc $N = n_1 + n_2 + \cdots + n_p$.

Lorsqu'une série comporte un grand nombre de valeurs, on cherche à la résumer, si possible, à l'aide de quelques nombres significatifs appelés paramètres.

Dans cette section, on présente quelques paramètres permettant de résumer des séries à caractère quantitatif qui seront illustrés à l'aide de l'exemple suivant :

Série 1

Une étude sur le nombre de Statisticiens dans les laboratoires de recherche du centre universitaire a donné les résultats suivants :

Nombre de biologistes	1	2	3	4	5	6	7	8
Effectif	12	17	25	19	13	18	9	5

1.2.1 Paramètres de position

1.2.1.1 Paramètres de position de tendance centrale

Mode

Le mode d'une série statistique est la valeur du caractère qui correspond au plus grand effectif. Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, la classe modale est la classe du plus grand effectif.

Remarque : Il peut y avoir plusieurs modes ou classes modales.

Exemple

Déterminer le mode de la série 1.

Statisticiens

1 2 3 4 5 6 7 8

12 17 25 19 13 18 9 5

Pour la série 1, le mode est 3.

Médiane

La médiane d'une série statistique est un réel noté M_e tel que au moins 50% des valeurs sont inférieures ou égales à M_e et au moins 50% des valeurs sont supérieures ou égales à M_e.

Dans le cas d'une série à caractère discret, la médiane s'obtient en ordonnant les valeurs dans l'ordre croissant et en prenant la valeur centrale si N est impair et la moyenne des valeurs centrales si N est pair.

Dans le cas d'une série à caractère continu, la médiane peut s'obtenir de manière graphique en prenant la valeur correspondant à $0,5$ sur le polygone des fréquences cumulées croissantes.

L'avantage, mais parfois aussi l'inconvénient de la médiane, est qu'elle n'est pas affectée par les valeurs extrêmes des données.

Exemple

Déterminer la médiane de la série 1.

La médiane de la série 1 est 4 qui correspond à la moyenne de la 238ème et de la 239ème valeur.

Moyenne

La moyenne d'une série statistique (x_i, n_i) est le réel, noté $\bar{x}$ défini par :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{N} = \frac{\sum_{i=1}^{p} n_i x_i}{N}$$

Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, x_i désigne le centre de chaque classe.

Exemple

Déterminer la moyenne de la série 1.

$$\bar{x}_1 = \frac{12 \times 1 + 17 \times 2 + 25 \times 3 + 19 \times 4 + 13 \times 5 + 18 \times 6 + 9 \times 7 + 6 \times 8}{12 + 17 + 25 + 19 + 13 + 18 + 9 + 5} = 4.008475$$

1.2.1.2 Paramètres de position non centrale

Quartiles

Le premier quartile Q_1 est la plus petite valeur du caractère telle qu'au moins 25% des termes de la série aient une valeur qui lui soit inférieure ou égale. Le troisième quartile Q_3 est la plus petite valeur du caractère telle qu'au moins 75% des termes de la série aient une valeur qui lui soit inférieure ou égale.

Dans le cas d'une série à caractère discret, les quartiles s'obtiennent en ordonnant les valeurs dans l'ordre croissant puis :

— Si N est multiple de 4 alors Q_1 est la valeur de rang $\frac{N}{4}$ et Q_3 est la valeur de rang $\frac{3N}{4}$.

— Si N n'est pas multiple de 4 alors Q_1 est la valeur de rang immédiatement supérieur à $\frac{N}{4}$ et Q_3 est la valeur de rang immédiatement supérieur à $\frac{3N}{4}$.

Exemple

Déterminer les quartiles de la série 1.

Le nombre de données (118) n'est multiple de 4. Le premier quartile est donc la 30^{e} valeur et le troisième quartile la 89^{e} valeur.

On a ainsi : $Q_1 = 3$ et $Q_3 = 6$.

Dans le cas d'une série à caractère continu, les quartiles peuvent s'obtenir à partir du polygone des fréquences cumulées croissantes où Q_1 est la valeur correspondant à la fréquence cumulée croissante égale $0,25$ et Q_3 est la valeur correspondant à la fréquence cumulée croissante égale $0,75$.

1.2.2 Paramètres de dispersion

Étendue

L'étendue est la différence entre la plus grande valeur du caractère et la plus petite. Ce paramètre est très sensible aux valeurs extrêmes.

Exemple

Déterminer l'étendue de la série 1.

L'étendue de la série 1 est $8 - 1 = 7$.

Écart interquartile

L'intervalle interquartile est l'intervalle $[Q_1; Q_3]$.

L'écart interquartile $IQR = Q_3 - Q_1$ est la longueur de l'intervalle interquartile.

Remarque : Contrairement à l'étendue, l'écart interquartile élimine les valeurs extrêmes, qui peut être un avantage. En revanche il ne prend en compte que 50% de l'effectif, qui peut être un inconvénient.

Exemple

Déterminer l'intervalle interquartile et l'écart interquartile de la série 1.

L'intervalle interquartile est $[3;6]$. L'écart interquartile est donc 3.

Variance et écart-type

Pour mesurer la dispersion d'une série, on peut s'intéresser à la moyenne des distances des valeurs à la moyenne. On utilise plutôt les carrés des distances.

On appelle variance d'une série quelconque à caractère quantitatif discret le nombre :

$$V_x = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2$$

On appelle écart-type de cette série le nombre $\sigma_x = \sqrt{V_x}$.

Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, x_i désigne le centre de chaque classe.

Proposition :
On peut calculer la variance de la façon suivante :

$$V_x = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$$

Démonstration :

$$\begin{aligned} V_x &= \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2n_i x_i \bar{x} + n_i \bar{x}^2 \\ &= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2\bar{x} \times \frac{1}{N}\sum_{i=1}^{n} n_i x_i + \bar{x}^2 \times \frac{1}{N}\sum_{i=1}^{n} n_i = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2 \end{aligned}$$

Exemple

Déterminer la variances et l'écart-type de la série 1.
On a : $V_1 = 3.82044$ et $\sigma_1 = 1.954595$.

1.3 Quelques graphiques avec R

Les utilisateurs de l'informatique statistique devront produire des graphiques de leurs données et des résultats de leurs calculs. Dans ce chapitre, nous commençons par un aperçu général de la façon dont cela est effectué à l'aide du logiciel R et apprenons à dessiner des graphiques de base.

Il existe également d'autres graphiques disponibles dans R : des graphiques interactifs, des affichages 3D, etc.

1.3.1 Graphiques linéaires

Nous allons d'abord produire un graphique très simple en utilisant les valeurs du vecteur Amaryllis (nom d'une fleur) :

```
# Définir le vecteur Amaryllis avec 5 valeurs
Amaryllis <- c(2, 4, 5, 4, 8)
# Représenter graphiquement le vecteur de Amaryllis avec tous les défauts
plot(Amaryllis)
```

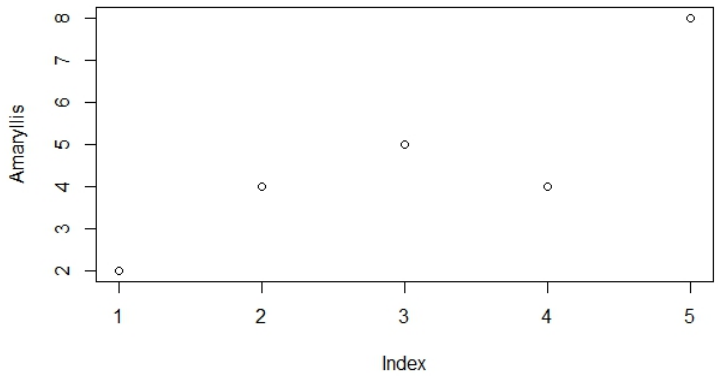

Fig. 1. Graphe du vecteur Amaryllis

Ajoutons un titre, une ligne pour relier les points et de la couleur :

```
# Tracer des Amaryllis en utilisant des points bleus superposés par une ligne

plot(Amaryllis, type="o", col="blue")

# Créer un titre avec une police rouge, gras / italique

title(main="Amaryllis", col.main="red", font.main=4)
```

Ajoutons maintenant une ligne rouge pour les Astéracées (genre de plante) et spécifions directement la plage d'axe y afin qu'elle soit assez grande pour contenir les données du Astéracées :

```
# Définir 2 vecteurs

Amaryllis <- c(2, 4, 5, 4, 8)
Asteraceae <- c(3, 5, 6, 3, 10)
```

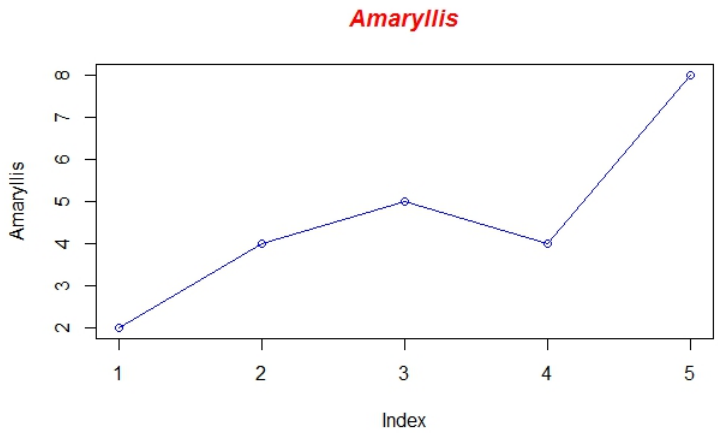

Fig. 2. Graphe Amaryllis avec couleurs et titre

Représenter graphiquement Amaryllis en utilisant un axe y compris entre 0 et 12

plot(Amaryllis, type="o", col="blue", ylim=c(0,12))
# Graphique Asteraceae avec une ligne pointillée rouge et des points carrés

lines(Asteraceae, type="o", pch=22, lty=2, col="red")

Créer un titre avec une police rouge, gras / italique

title(main="Amaryllis et Asteraceae", col.main="red", font.main=4)

Ensuite, changeons les étiquettes des axes pour faire correspondre nos données et ajoutons une légende. Nous allons également calculer les valeurs de l'axe des y à l'aide de la fonction max pour que toute modification de nos données soit automatiquement reflétée dans notre graphique.

Calculez la plage de 0 à la valeur maximale des Amaryllis et des Asteraceae

grange <- range(0, Amaryllis, Asteraceae)

Graphique des autos en utilisant l'axe des ordonnées allant de 0 à max dans le vecteur Amaryllis ou Asteraceae. Désactiver les axes et les annotations (étiquettes d'axe) afin que nous puissions les spécifier nous-mêmes

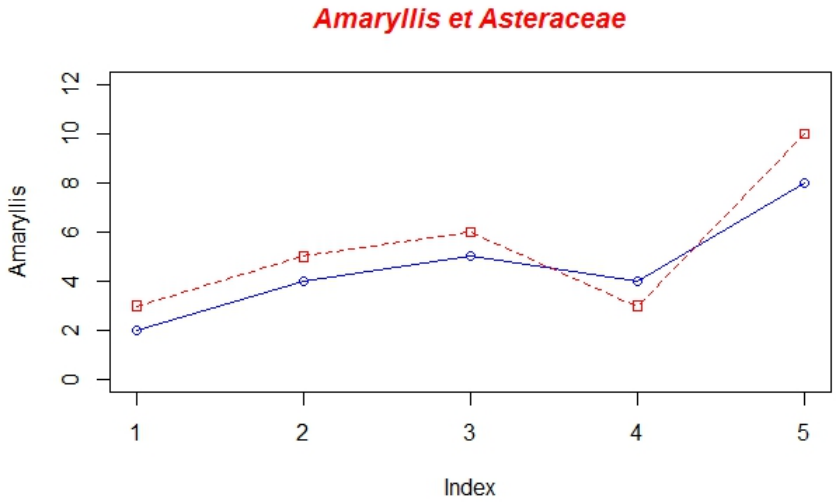

Fig. 3. Graphe Amaryllis et Asteraceae

plot(Amaryllis, type="o", col="blue", ylim=grange, axes=FALSE, ann=FALSE)

Faire un axe x en utilisant les étiquettes Lun-Ven

axis(1, at=1 :5, lab=c("Lun","Mar","Mer","Jeu","Ven"))

Créez l'axe y avec des étiquettes horizontales affichant des graduations tous les 4 points. 4 * 0 : grange [2] est équivalent à c (0,4,8,12).

axis(2, las=1, at=4*0 :grange[2])

Graphique Asteraceae avec une ligne pointillée rouge et des points carrés

lines(Asteraceae, type="o", pch=22, lty=2, col="red")

Créer un titre avec une police rouge, gras / italique

title(main="Amaryllis et Asteraceae", col.main="red", font.main=4)

étiquetez les axes x et y avec du texte vert foncé

title(xlab="Jours", col.lab=rgb(0,0.5,0))
title(ylab="Total", col.lab=rgb(0,0.5,0))

Créez une légende à (1, grange [2]) légèrement plus petite (cex) et utilisant les mêmes couleurs de ligne et les mêmes points utilisés par les tracés réels.

legend(1, grange[2], c("Amaryllis","Asteraceae"), cex=0.8, col=c("blue","red"), pch=21 :22, lty=1 :2) ;

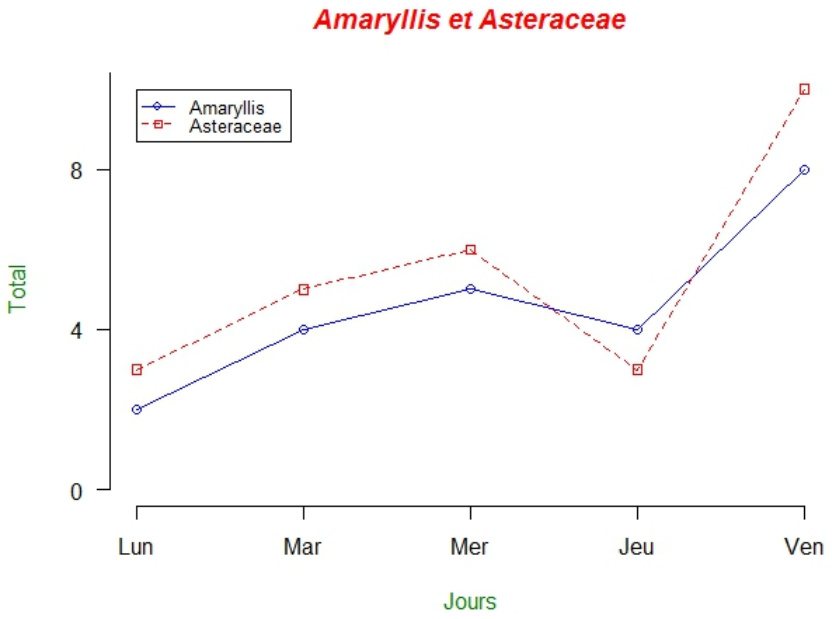

Fig. 4. Graphe Amaryllis et Asteraceae avec une légende

1.3.2 Diagramme à barres (à bande)

Le type le plus élémentaire de graphique est celui qui affiche un seul ensemble de nombres. Les graphiques à barres et les graphiques à points le font en affichant une barre ou un point dont la longueur ou la position correspond au nombre.

Commençons par un simple diagramme à barres illustrant le vecteur Amaryllis :

```
# Définir le vecteur Amaryllis avec 5 valeurs

Amaryllis <- c(2, 4, 5, 4, 8)

# Graphique Amaryllis

barplot(Amaryllis)
```

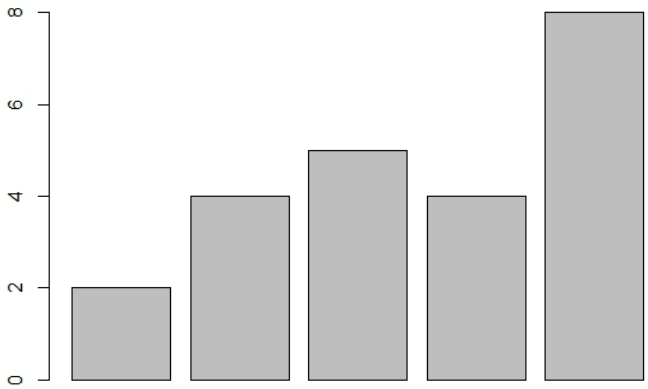

Fig. 5. Diagramme à barres d'Amaryllis

```
# Graph Amaryllis avec des étiquettes spécifiées pour les axes. Utilisez le bleu
# bordures et lignes diagonales dans les barres

barplot(Amaryllis, main="Diagramme à barres d'Amaryllis", xlab="Jours", ylab="Total",
names.arg=c("Lun","Mar","Mer","Jeu","Ven"), border="blue", density=c(10,20,30,40,50))
```

1.3.3 Histogramme

Une autre façon de regarder la distribution consiste à tracer un histogramme des données. Un histogramme est un type spécial de graphique à barres utilisé pour montrer la distribution de fréquence d'un ensemble de nombres. Chaque rectangle représente le nombre de valeurs x comprises dans la plage indiquée par la base du rectangle. Habituellement, tous les rectangles doivent avoir la même largeur. c'est la valeur par défaut de R. Dans

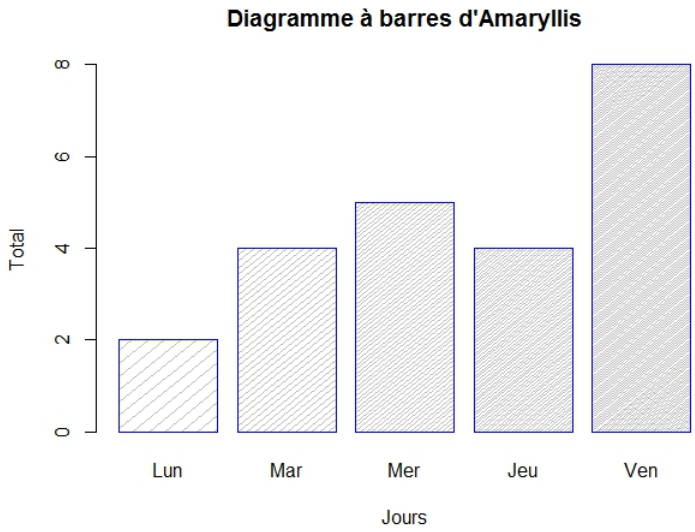

Fig. 6. Diagramme à barres modifié d'Amaryllis

ce cas, la hauteur de chaque rectangle est proportionnelle au nombre d'observations dans l'intervalle correspondant. Si les rectangles ont des largeurs différentes, l'aire du rectangle doit être proportionnelle au nombre ; de cette manière, la hauteur représente la densité (c'est-à-dire la fréquence par unité de x).

Commençons par un simple histogramme représentant graphiquement la distribution des Asteraceae

Créer un histogramme pour les Asteraceae

hist (Asteraceae)

Traçons maintenant un histogramme en couleur des données combinées des Amaryllis et d'Asteraceae.

Concaténer les deux vecteurs

combine <- c(Amaryllis, Asteraceae)

Créez un histogramme pour les Amaryllis en bleu clair avec l'axe des ordonnées allant de 0 à 10

hist(combine, col="lightblue", ylim=c(0,10))

Modifiez maintenant les sauts afin qu'aucune des valeurs ne soit regroupée et inversez les étiquettes de l'axe des ordonnées horizontalement.

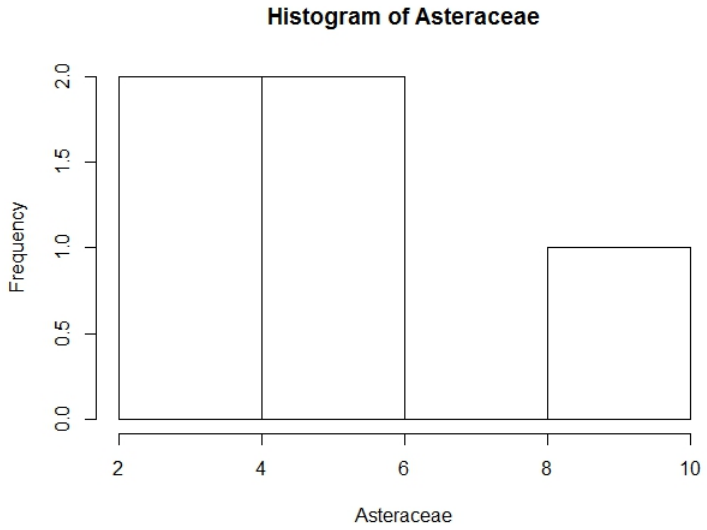

Fig. 7. Histogramme d'Asteraceae

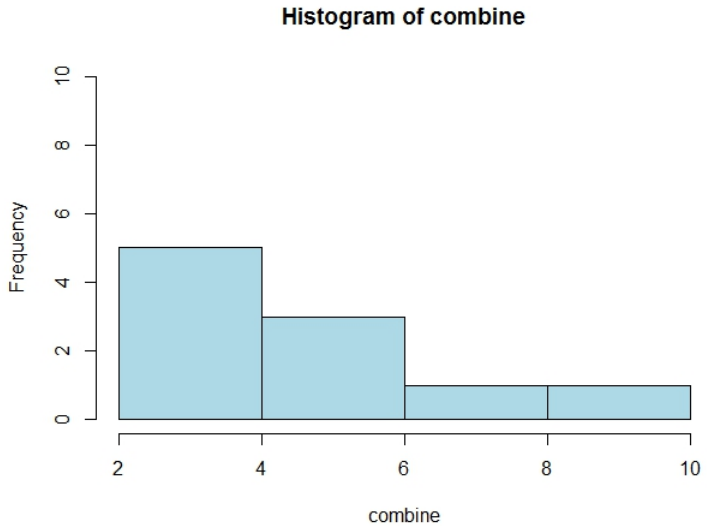

Fig. 8. Histogramme des Amaryllis et d'Asteraceae

Calcule la plus grande valeur y utilisée dans les autos

maxnum <- max (combine)

Créez un histogramme pour les Amaryllis avec des couleurs d'incendie, définissez des sauts de manière à ce que chaque nombre soit dans son propre groupe, définissez l'axe des abscisses à partir de 0-maxnum, désactivez la fermeture à droite des intervalles de cellules, définissez le titre et positionnez les étiquettes de l'axe des ordonnées à l'horizontale.

hist(combine, col=heat.colors(maxnum), breaks=maxnum, xlim=c(0,maxnum), right=F, main="Histogramme modifié des Amaryllis et d'Asteraceae", las=1)

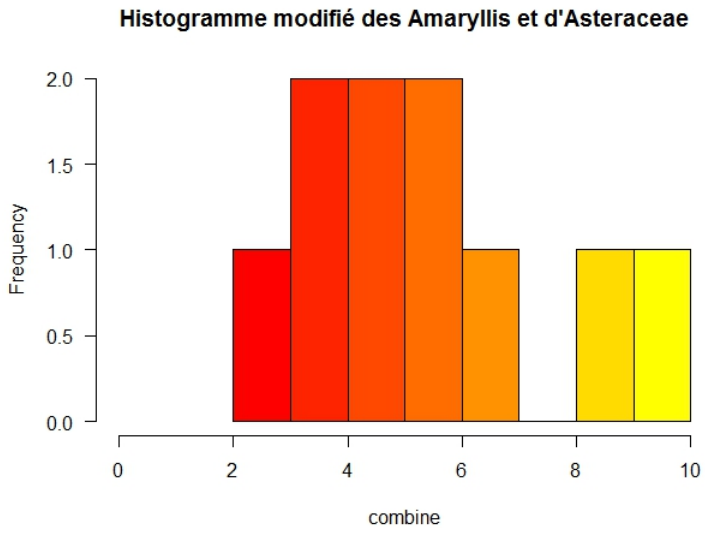

Fig. 9. Histogramme modifié des Amaryllis et d'Asteraceae

1.3.4 Diagramme en secteurs (camembert)

Pour les données nominales, la moyenne et l'écart type n'ont pas beaucoup de sens ; ni la médiane, ni les centiles. Pour voir la distribution des données, consultez le tableau des fréquences et le diagramme en barres (ou camembert) est une représentation graphique.

Remarque : les camemberts sont trompeurs et doivent être évités !
Commençons par un simple diagramme à secteurs illustrant le vecteur des Amaryllis :
# Créer un camembert pour les Amaryllis
pie(Amaryllis)

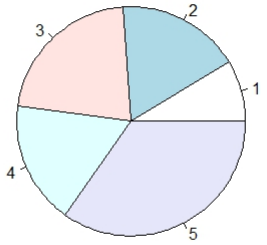

Fig. 10. Camembert des Amaryllis

Ajoutons maintenant une en-tête, changeons les couleurs et définissons nos propres étiquettes : # Créer un graphique à secteurs avec un titre défini, des couleurs et des étiquettes personnalisées :

pie(Amaryllis, main="Amaryllis", col=rainbow(length(Amaryllis)), labels=c("Lun","Mar","Mer","Jeu","Ven"))

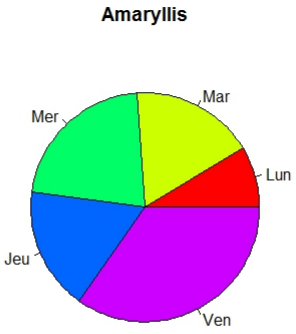

Fig. 11. Camembert modifié des Amaryllis

1.3.5 Boîte à moustaches

Une boîte à moustaches est une représentation graphique de la médiane et des quartiles, elle donne un aperçu de la distribution des données. Ce type de graphe est souvent utilisé pour comparer des données entre différents groupes.

Les boîtes à moustaches ne sont pas censées être très informatives, mais elles sont souvent utiles pour obtenir une idée approximative d'un ensemble de données. L'interprétation d'une boîte à moustaches nécessite quelques connaissances. Le diagramme suivant pourrait aider.

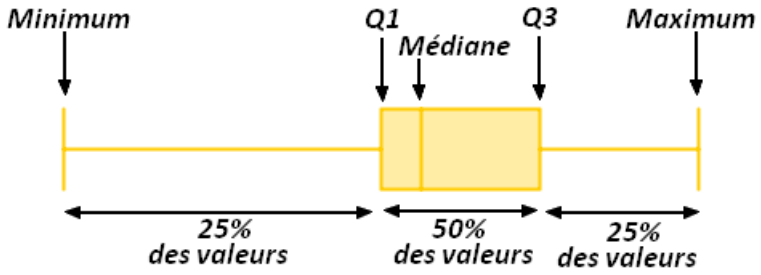

Fig. 12. Signification d'une Boite à moustache