

Analyse des variances : ANOVA et Kruskal-Wallis

Marie Vaugoyeau

7 November 2023

Table of contents

1	Définition de l'ANOVA	2
2	Les limites d'utilisations	5
2.1	Indépendance des données	5
2.2	Normalité des données	5
2.3	Homogénéité des variances	6
3	Réalisation d'une ANOVA	8
3.1	Vérification des données	8
3.2	Réalisation de l'ANOVA	9
3.3	Test post-hoc de Tukey	10
4	Réalisation d'une ANOVA non paramétrique : test de Kruskal-Wallis	13
4.1	Fonctionnement et limites	13
4.2	Utilisation de la fonction <code>kruskal.test()</code> du package <code>{stats}</code>	14
4.3	Test-post hoc de Nemenyi	14

*Ce support, produit pour le live du **11 novembre 2023 sur Twitch**, est mis à disposition selon les termes de la [Licence Creative Commons Attribution 4.0 International](#).*

1 Définition de l'ANOVA

💡 ANOVA : Analyse des variances

Permet de savoir si **deux échantillons ou plus** sont issus d'une **même population** ou pour le dire autrement, les groupes créés ont-ils la **même moyenne**.

L'ANOVA permet d'étudier l'influence d'au moins une **variable qualitative** ayant **deux modalités ou plus**, sur une **variable quantitative**.

D'un point de vue pratique, l'ANOVA cherche à savoir si les **moyennes des groupes** sont globalement **différentes** ou pour le dire autrement, si la **variation intragroupe** est **plus faible** que la **variation intergroupe**.

Le principe de l'ANOVA est de décomposer la **variabilité totale des données** en deux :

— la **variabilité factorielle** : la variabilité entre groupes, c'est-à-dire la différence entre la moyenne de toutes les données et les moyennes de chaque groupe (cf. Figure 1).

— la **variabilité résiduelle** : la variabilité qui reste une fois que la variabilité due au groupe est retirée c'est-à-dire la différence entre la moyenne du groupe et la valeur de chaque échantillon (cf. Figure 2).

```
library(tidyverse)

iris_moyenne <- iris |>
  group_by(Species) |>
  summarise(moyenne = mean(Sepal.Length))

ggplot(iris) +
  aes(x = Species, y = Sepal.Length, color = Species) +
  geom_jitter(alpha = 0.3) +
  geom_hline(
    aes(yintercept = mean(Sepal.Length))
  ) +
  geom_errorbar(
    aes(
      ymin = moyenne,
      y = 5.84, ymax = moyenne
    ),
    data = iris_moyenne,
    size = 1
  ) +
  geom_spoke(
```

```

aes(
  y = moyenne,
  radius = mean(iris$Sepal.Length) - moyenne,
  angle = 1.57
),
data = iris_moyenne,
linetype = "dashed"
) +
theme_classic() +
theme(legend.position = "none")

```

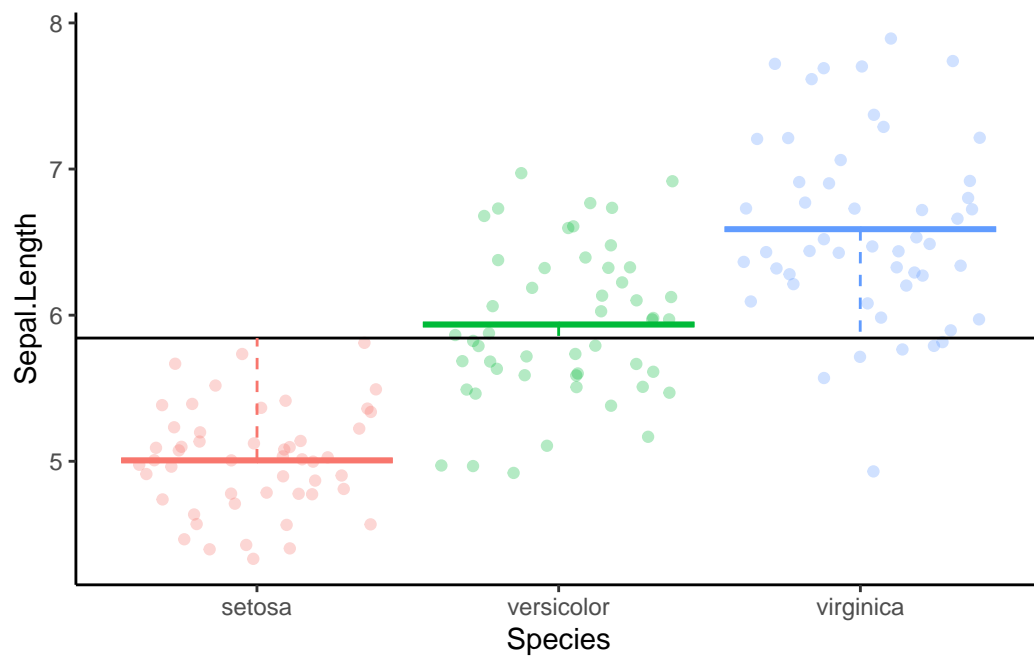


Figure 1: Variabilité factorielle

```

ggplot(iris) +
  aes(x = Species, y = Sepal.Length, color = Species) +
  geom_jitter(alpha = 0.3) +
  geom_errorbar(
    aes(
      ymin = moyenne,
      y = 5.84,
      ymax = moyenne
    )
  )

```

```

    ),
    data = iris_moyenne,
    size = 1
  ) +
  geom_spoke(
    aes(
      radius = iris_moyenne$moyenne - Sepal.Length,
      angle = 1.57
    ),
    data = iris |>
      group_by(Species) |>
      slice(3) |>
      ungroup(),
    linetype = "dashed"
  ) +
  geom_spoke(
    aes(
      radius = iris_moyenne$moyenne - Sepal.Length,
      angle = 1.57
    ),
    data = iris |>
      group_by(Species) |>
      slice(11) |>
      ungroup(),
    linetype = "dashed"
  ) +
  geom_point(
    data = iris |>
      group_by(Species) |>
      slice(c(3, 11)) |>
      ungroup()
  ) +
  theme_classic() +
  theme(legend.position = "none")

```

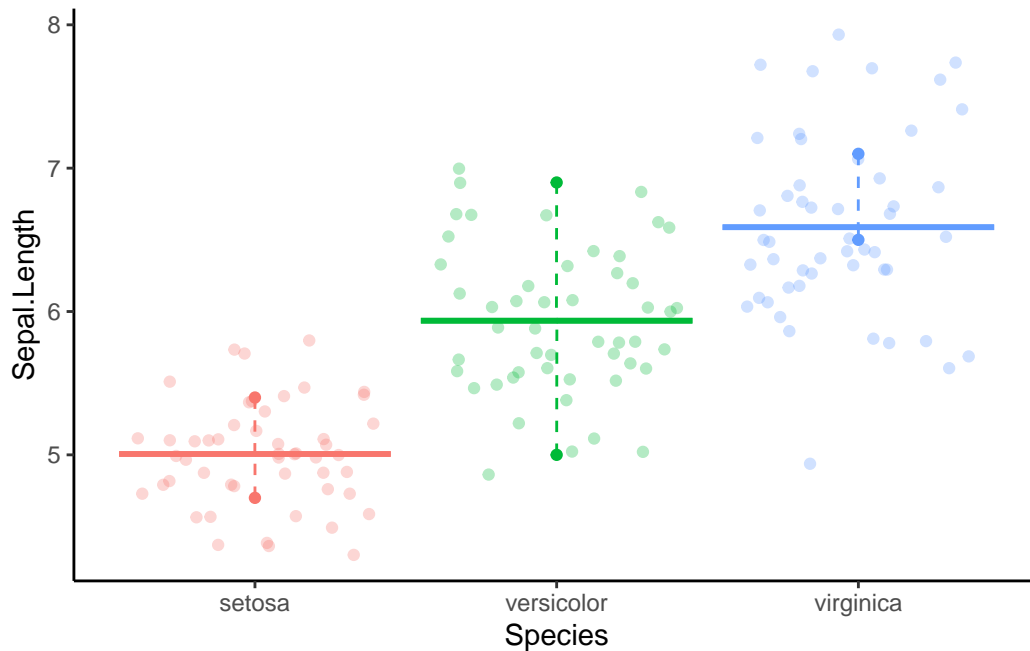


Figure 2: Variabilité résiduelle

2 Les limites d'utilisations

2.1 Indépendance des données

Les données doivent provenir d'un échantillonnage aléatoire et les groupes doivent-êtré indépendants entre eux.

2.2 Normalité des données

Les données au sein de chaque groupe doivent suivre une loi normale ou être approximé par une loi normale ($n > 30$).

```
shapiro.test(iris$Sepal.Length)
```

```
Shapiro-Wilk normality test
```

```
data:  iris$Sepal.Length
W = 0.97609, p-value = 0.01018
```

```
# les données ne suivent pas une loi normale

map(
  .x = c(iris$Species |> levels()),
  .f = ~shapiro.test(filter(iris, Species == .x)$Sepal.Length)
)
```

```
[[1]]
```

```
Shapiro-Wilk normality test
```

```
data: filter(iris, Species == .x)$Sepal.Length
W = 0.9777, p-value = 0.4595
```

```
[[2]]
```

```
Shapiro-Wilk normality test
```

```
data: filter(iris, Species == .x)$Sepal.Length
W = 0.97784, p-value = 0.4647
```

```
[[3]]
```

```
Shapiro-Wilk normality test
```

```
data: filter(iris, Species == .x)$Sepal.Length
W = 0.97118, p-value = 0.2583
```

```
# les données au sein de chaque groupe suivent des lois normales
```

2.3 Homogénéité des variances

Les groupes doivent avoir une variance similaire.

Le test de Bartlett permet de tester la variance de plus de deux groupes.

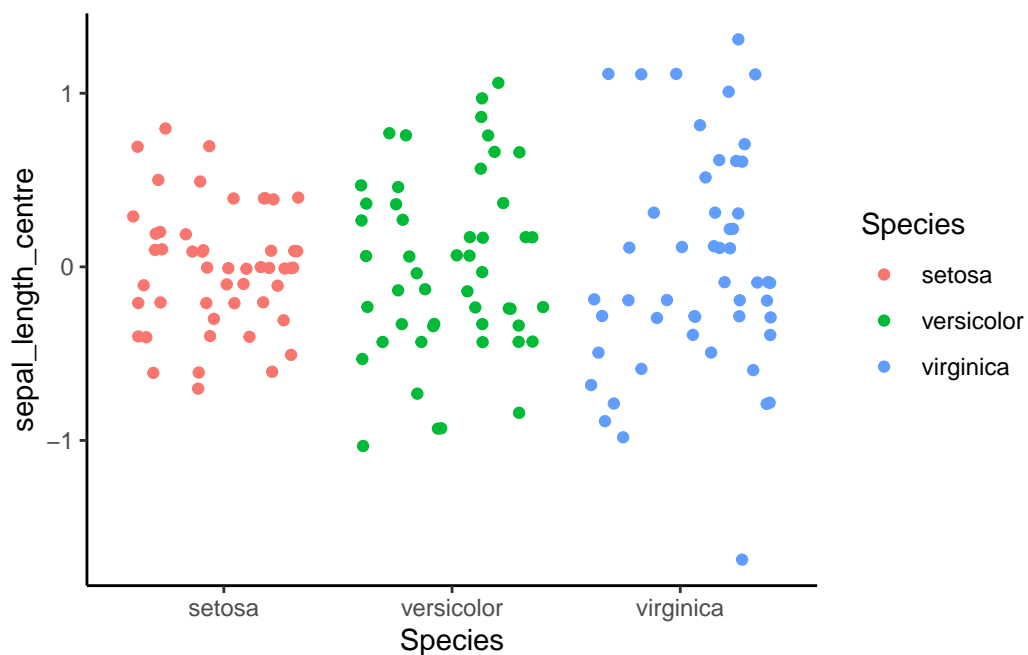
```
bartlett.test(Sepal.Length ~ Species, data = iris)
```

Bartlett test of homogeneity of variances

data: Sepal.Length by Species

Bartlett's K-squared = 16.006, df = 2, p-value = 0.0003345

```
iris |>
  left_join(iris_moyenne) |>
  mutate(sepal_length_centre = Sepal.Length - moyenne) |>
  ggplot() +
  aes(x = Species, color = Species, y = sepal_length_centre) +
  geom_jitter() +
  theme_classic()
```



Warning

Il ne faut pas faire d'ANOVA ici, les groupes n'ont pas la même variance !

3 Réalisation d'une ANOVA

Comme la longueur des sépales ne peut pas être utilisée, on va le faire sur la largeur des sépales.

3.1 Vérification des données

```
map(  
  .x = c(iris$Species |> levels()),  
  .f = ~shapiro.test(filter(iris, Species == .x)$Sepal.Width)  
)
```

[[1]]

Shapiro-Wilk normality test

data: filter(iris, Species == .x)\$Sepal.Width
W = 0.97172, p-value = 0.2715

[[2]]

Shapiro-Wilk normality test

data: filter(iris, Species == .x)\$Sepal.Width
W = 0.97413, p-value = 0.338

[[3]]

Shapiro-Wilk normality test

data: filter(iris, Species == .x)\$Sepal.Width
W = 0.96739, p-value = 0.1809

```
bartlett.test(Sepal.Width ~Species, data = iris)
```

Bartlett test of homogeneity of variances


```
data: Sepal.Width by Species
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Les données suivent des lois normales et les variances sont similaires.

3.2 Réalisation de l'ANOVA

! Important

La détermination d'un modèle ANOVA doit-être réalisé avec la fonction `aov()` du package `{stats}`. Les fonctions `anova()` du package `{stats}` ou `Anova()` du package `{car}` permet de réaliser une analyse de variance/déviance sur un modèle donc par exemple le résultat de `aov()` mais pas que ^^

```
anova_sepal_largeur <- aov(Sepal.Width ~ Species, data = iris)

summary(anova_sepal_largeur)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
Species        2  11.35    5.672   49.16 <2e-16 ***
Residuals     147  16.96    0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(anova_sepal_largeur)
```

Analysis of Variance Table

```
Response: Sepal.Width
              Df Sum Sq Mean Sq F value    Pr(>F)
Species        2 11.345   5.6725   49.16 < 2.2e-16 ***
Residuals     147 16.962   0.1154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

! Oublié dans le [live](#)

J'ai oublié de le préciser dans le [live](#) sur twitch mais il faut penser à vérifier que les résidus (la différence entre les valeurs prédites et observées) suivent une loi normales et soient homogènes !

```
shapiro.test(anova_sepal_largeur$residuals)
```

Shapiro-Wilk normality test

```
data: anova_sepal_largeur$residuals  
W = 0.98948, p-value = 0.323
```

```
bartlett.test(anova_sepal_largeur$residuals ~ iris$Species)
```

Bartlett test of homogeneity of variances

```
data: anova_sepal_largeur$residuals by iris$Species  
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Les **résidus** suivent bien une **loi normale** et sont **homogènes** donc le **modèle est validé**.

Il est aussi possible de le vérifier graphiquement grâce à la fonction `plot()` appliquée sur le modèle ajusté.

```
plot(anova_sepal_largeur)
```

Pour la vérification graphique :

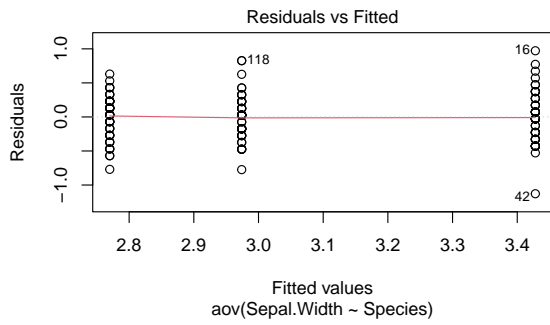
- __ Le QQ-plot, Figure [3b](#) montre que les résidus suivent une loi normale
- __ La Figure [3c](#) montre que les résidus sont homogènes

L'ANOVA réalisée est donc validée !

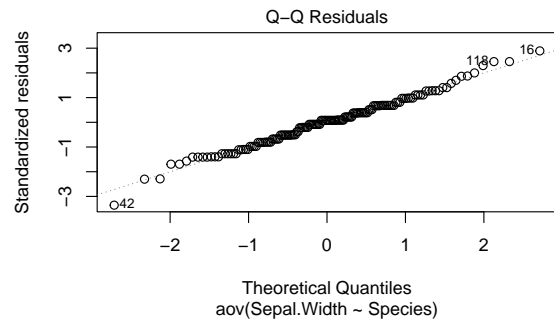
3.3 Test post-hoc de Tukey

Afin de savoir quel(s) groupe(s) est(sont) différent(s), il faut utiliser un test post-hoc de Tukey.

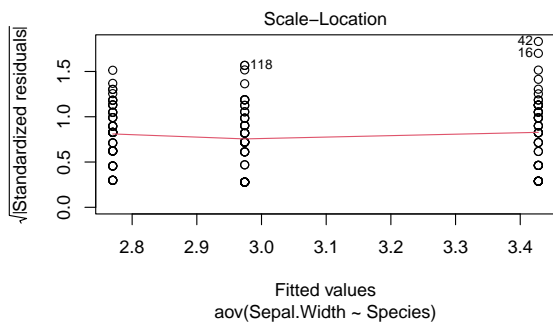
```
TukeyHSD(anova_sepal_largeur)
```



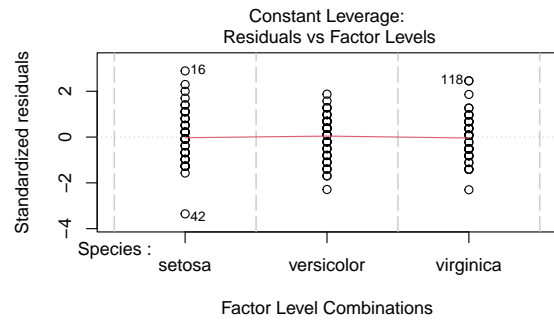
(a) Résidus en fonction des valeurs prédites



(b) QQplot : Vérification de la normalité des résidus



(c) Vérification de l'homogénéité des variances



(d) Vérification de l'effet de levier des individus

Figure 3: Vérification graphique du modèle

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = Sepal.Width ~ Species, data = iris)

\$Species

	diff	lwr	upr	p adj
versicolor-setosa	-0.658	-0.81885528	-0.4971447	0.0000000
virginica-setosa	-0.454	-0.61485528	-0.2931447	0.0000000
virginica-versicolor	0.204	0.04314472	0.3648553	0.0087802

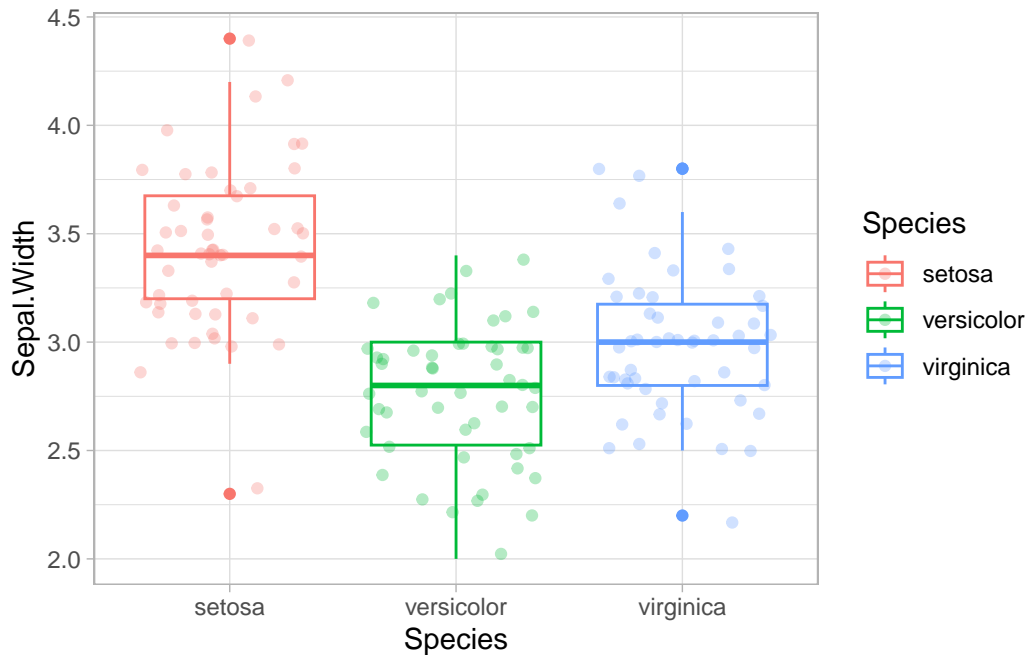
```
rstatix::tukey_hsd(anova_sepal_largeur)
```

A tibble: 3 x 9

	term	group1	group2	null.value	estimate	conf.low	conf.high	p.adj
*	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Species	setosa	versicolor	0	-0.658	-0.819	-0.497	3.10e-14
2	Species	setosa	virginica	0	-0.454	-0.615	-0.293	1.36e- 9
3	Species	versicolor	virginica	0	0.204	0.0431	0.365	8.78e- 3

i 1 more variable: p.adj.signif <chr>

```
ggplot(iris) +  
  aes(x = Species, color = Species, y = Sepal.Width) +  
  geom_boxplot() +  
  geom_jitter(alpha = 0.3) +  
  theme_light()
```



4 Réalisation d'une ANOVA non paramétrique : test de Kruskal-Wallis

Réalisation sur les longueurs de sépales qui ne sont pas homogènes entre les groupes

4.1 Fonctionnement et limites

💡 Kruskal-Wallis

Permet de savoir si **deux échantillons ou plus** sont issus d'une **même population** ou pour le dire autrement, les groupes créés ont-ils la **même médiane**.

Le test de Kruskal-Wallis se base sur le rang des données.

```
iris |>
  arrange(Sepal.Length) |>
  rowid_to_column(var = "rang") |>
  group_by(Species) |>
  summarise(somme_rang = sum(rang)) |>
  ungroup()
```

```
# A tibble: 3 x 2
  Species    somme_rang
  <fct>      <int>
1 setosa      1448
2 versicolor  4115
3 virginica   5762
```

Une fois que le rang de chaque groupe calculé, la statistique de test va être calculé et comparer à une valeur seuil.



Les limites

- __ échantillonnage aléatoire
- __ indépendance des groupes
- __ Plus de 5 observations par groupe

4.2 Utilisation de la fonction `kruskal.test()` du package `{stats}`

Comme la longueur des sépales n'avaient pas la même variance en fonction de l'espèce, il n'est pas possible de réaliser une ANOVA.

Le test de Kruskal-Wallis est conseillé ici.

```
kruskal.test(Sepal.Length ~ Species, data = iris)
```

```
Kruskal-Wallis rank sum test
```

```
data: Sepal.Length by Species
```

```
Kruskal-Wallis chi-squared = 96.937, df = 2, p-value < 2.2e-16
```

4.3 Test-post hoc de Nemenyi

```
summary(
  PMCMRplus::kwAllPairsNemenyiTest(
    data = iris,
    Sepal.Length ~ Species
  )
)
```

	q value	Pr(> q)	
versicolor - setosa == 0	8.628	3.1659e-09	***
virginica - setosa == 0	13.764	3.4084e-14	***
virginica - versicolor == 0	5.137	0.00082097	***

```
ggplot(iris) +
  aes(x = Species, color = Species, y = Sepal.Length) +
  geom_boxplot() +
  geom_jitter(alpha = 0.3) +
  theme_bw()
```

