

Anova

Eric Marcon

14 février 2024

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Anova à 1 facteur

Anova

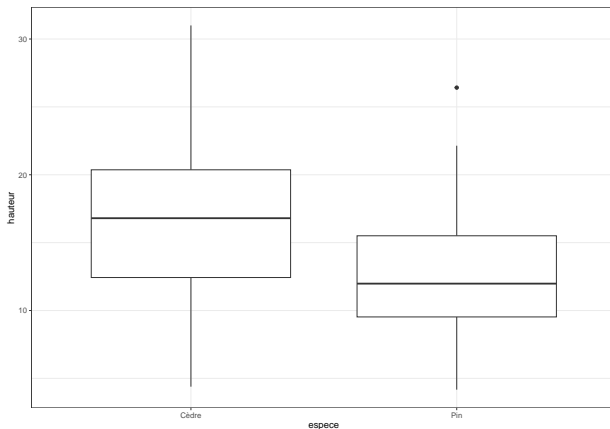
Eric Marcon

Anova à 1
facteur

Tests post-hoc

Les cèdres sont-ils plus haut que les pins ?

```
ventoux |>  
  ggplot(aes(x = espece, y = hauteur)) +  
  geom_boxplot()
```



Modèle de régression avec des covariables toutes catégorielles, codées sous forme d'indicatrices (autant d'indicatrices que de modalités - 1).

Exemple du Ventoux :

$$Y = \beta_0 + \beta_1 \mathbb{1}('Cedre') + E$$

Ici, deux modalités seulement → quelle autre méthode utiliser ?

L'Anova à un facteur étend le test de Student à plus de deux groupes, comme le test de Welch (`oneway.test()`), non traité ici.

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

```
aov(hauteur ~ espece, data = ventoux) %>% {. ->> ventoux_aov} |> summary
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## espece         1     824    823.6    33.1 2.9e-08
## Residuals    221     5499     24.9
##
## espece        ***
## Residuals
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La hauteur des arbres est différente entre les espèces.

La statistique de test est le rapport entre les sommes des carrés des écarts intergroupe et intragroupe, divisés par leurs degrés de liberté, qui suit une loi de Fisher (F).

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Celles du modèle linéaire.

- Homoscédasticité : la variance de l'erreur est identique entre les groupes.

Anova

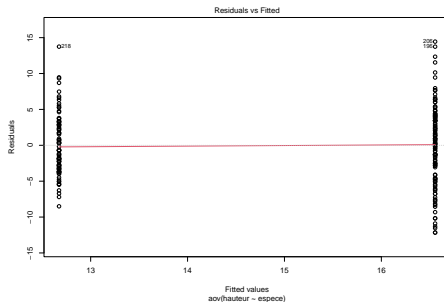
Eric Marcon

Anova à 1
facteur

Tests post-hoc

Graphique $E \sim Y^*$

```
plot(ventoux_aov, which = 1)
```



Les erreurs doivent être centrée sur 0 et uniformément réparties.

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Le test de Levene invalide l'hypothèse nulle d'égalité des variances.

```
library("car")
with(ventoux, leveneTest(hauteur ~ espece))

## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group    1  6.7887 0.009797 **
##           221
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Anova non paramétrique

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Il faut utiliser le test de Kruskal-Wallis, qui est un modèle linéaire sur les rangs.

```
kruskal.test(hauteur ~ espece, data = ventoux)
```

```
##  
##  Kruskal-Wallis rank sum test  
##  
## data:  hauteur by espece  
## Kruskal-Wallis chi-squared = 31.337, df = 1,  
## p-value = 2.169e-08
```

Le test de Kruskal-Wallis étend le test de Spearman à plus de deux groupes.

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Tests post-hoc

Anova

Eric Marcon

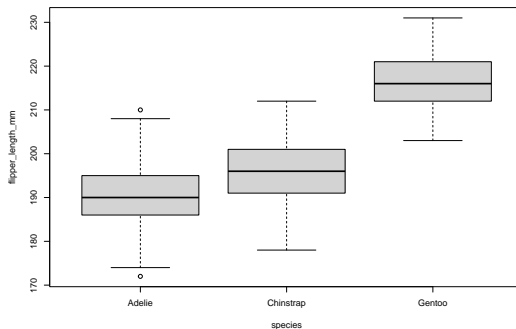
Anova à 1
facteur

Tests post-hoc

Exemple **traité en détail** par Antoine Soetewey.

Les données sont les longueurs des nageoires de trois espèces de manchots.

```
library("palmerpenguins")
with(penguins, boxplot(flipper_length_mm ~ species))
```



Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

```
aov(flipper_length_mm ~ species, data = penguins) %>%  
  {. ->> penguins_aov} |>  
  summary()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## species        2  52473   26237    594.8 <2e-16 ***  
## Residuals     339  14953        44  
## ---  
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness
```

→ Vérifiez le respect des hypothèses

Les trois espèces n'ont pas toutes les mêmes longueur de nageoires... Mais encore ? → Tests post-hoc.

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Le test de Tukey compare tous les groupes deux à deux.

```
library("multcomp")
penguins_aov |> glht(linfct = mcp(species = "Tukey")) %>%
  {. ->> penguins_tukey} |> summary()
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = flipper_length_mm ~ species, data = penguins)
##
## Linear Hypotheses:
```

	Estimate	Std. Error	
## Chinstrap - Adelie == 0	5.8699	0.9699	
## Gentoo - Adelie == 0	27.2333	0.8067	
## Gentoo - Chinstrap == 0	21.3635	1.0036	
	t value	Pr(> t)	
## Chinstrap - Adelie == 0	6.052	<1e-08	***
## Gentoo - Adelie == 0	33.760	<1e-08	***
## Gentoo - Chinstrap == 0	21.286	<1e-08	***
## ---			

Le problème des tests multiples

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Le seuil de risque de 5% signifie que 5% des tests seront des faux positifs.

Avec 7 groupes, on fait $6 \times 7/2 = 21$ tests d'égalité, donc on attend un faux positif.

Solution : réduire le seuil de risque α_m (pour *multiple*) :

$$\alpha_m = 1 - (1 - \alpha)^n \approx \alpha/n$$

C'est la correction de *Bonferroni*.

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Si un groupe est la référence (par exemple, le témoin), utiliser le test de Dunnett, plus puissant.

Le groupe de référence est le premier des facteurs.

```
str(penguins$species)
```

```
## Factor w/ 3 levels "Adelie","Chinstrap",...: 1 1 1 1 1 1 1 1 1 1 1 ...
```

Il peut être modifié :

```
penguins$species <- relevel(penguins$species, ref = "Gentoo")  
str(penguins$species)
```

```
## Factor w/ 3 levels "Gentoo","Adelie",...: 2 2 2 2 2 2 2 2 2 2 2 ...
```

```
penguins_aov |> glht(linfmt = mcp(species = "Dunnett")) |> summary()
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
##   Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = flipper_length_mm ~ species, data = penguins)
##
## Linear Hypotheses:
##
##              Estimate Std. Error
## Chinstrap - Adelie == 0    5.8699     0.9699
## Gentoo - Adelie == 0      27.2333     0.8067
##
##              t value Pr(>|t|)
## Chinstrap - Adelie == 0    6.052 7.59e-09 ***
## Gentoo - Adelie == 0      33.760 < 1e-10 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```


Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Objectif : ajouter les informations de l'Anova aux boîtes à moustache.

Avec *ggstatsplot* :

```
library("ggstatsplot")
penguins |>
  ggbetweenstats(
    x = species,
    y = flipper_length_mm,
    type = "parametric", # ANOVA or Kruskal-Wallis
    var.equal = TRUE, # ANOVA or Welch ANOVA
    plot.type = "box",
    pairwise.comparisons = TRUE,
    pairwise.display = "significant",
    centrality.plotting = FALSE,
    bf.message = FALSE
  )
```

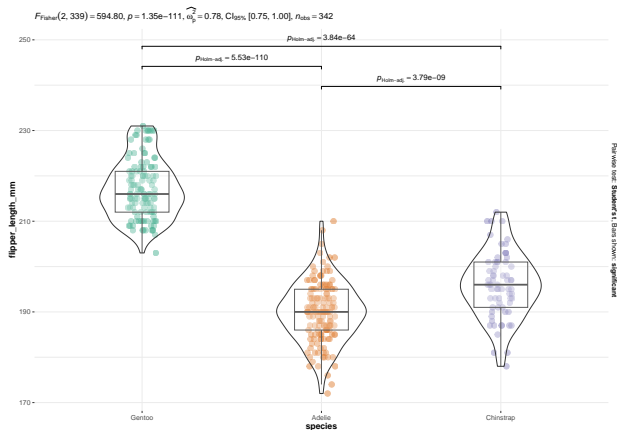
Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

ggstatsplot :



Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc

Graphique plus sobre, selon [Rosane Rech](#).

Les groupes identiques sont habituellement marqués par des lettres.

```
# Test de Tukey du package stats (moins détaillé que celui de multcomp)
penguins_tukey <- TukeyHSD(penguins_aov)
library("multcompView")
(penguins_letters <- multcompLetters4(penguins_aov, penguins_tukey))
```

```
## $species
##      Gentoo Chinstrap      Adelie
##      "a"          "b"          "c"
```

```
# Préparation d'un tibble contenant les lettres (format compliqué)
penguins_letters_tb <- tibble(
  species = names(penguins_letters[[1]]$Letters),
  letter = as.character(penguins_letters[[1]]$Letters)
)
```

Un tableau avec les groupes, leur lettre et leur 75ème centile est nécessaire pour la figure :

```
penguins |>
  group_by(species) %>%
  summarise(q_75 = quantile(flipper_length_mm, probs = 0.75, na.rm = TRUE))
  inner_join(penguins_letters_tb) -> penguins_letters_tbq
```

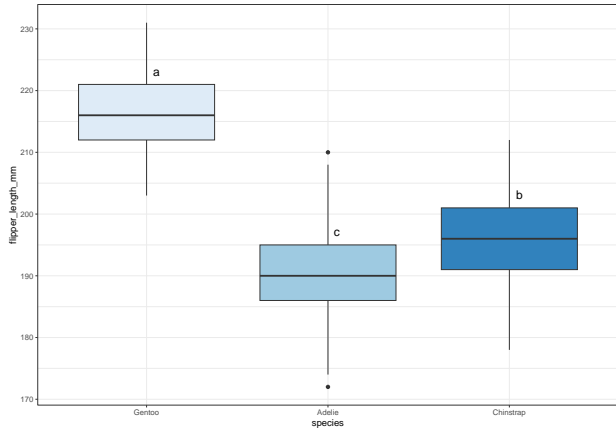
```
# Graphique
ggplot() +
  geom_boxplot(
    data = penguins,
    aes(x = species, y = flipper_length_mm, fill = species),
    show.legend = FALSE
  ) +
  geom_text(
    data = penguins_letters_tbq,
    aes(x = species, y = q_75, label = letter),
    size = 5, vjust=-1, hjust =-1
  ) +
  scale_fill_brewer(palette = "Blues")
```

Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc



Anova

Eric Marcon

Anova à 1
facteur

Tests post-hoc