TP Analyse de la variance (ANOVA) à deux facteurs

Mathilde Boissel

09/11/2020

Table of Contents

Théorie

Contexte

Quand utilise-t-on l'ANOVA à 2 facteurs (2 critères) ? On se limitera au cas suivant :

- Deux facteurs étudiés A et B.
- Le facteur A possède p modalités $A_1, A_2, \ldots, A_i, \ldots, A_p$.
- Le facteur B possède q modalités $B_1, B_2, \ldots, B_j, \ldots, B_q$.
- Le plan d'expérience est un plan factoriel complet, c.a.d que l'on expérimente toutes les combinaisons de modalités (A_i, B_j).
- On fait n répétitions par traitement, n peut être égal à 1 ce qui signifie qu'il n'y a pas de répétition. Le nombre d'unités expérimentales est donc $N = npq$.
- L'expérimentation est un dispositif en randomisation totale.

Dans le cas où $n > 1$, on dispose des données suivantes :

	B_1	$\cdots$	B_j	$\cdots$	B_q
A_1	y_{111} $\cdots$ y_{11k} $\cdots$ y_{11n}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{1j1} $\cdots$ y_{1jk} $\cdots$ y_{1jn}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{1q1} $\cdots$ y_{1qk} $\cdots$ y_{1qn}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	y_{i11} $\cdots$ y_{i1k} $\cdots$ y_{i1n}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{ij1} $\cdots$ y_{ijk} $\cdots$ y_{ijn}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{iq1} $\cdots$ y_{iqk} $\cdots$ y_{iqn}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_p	y_{p11} $\cdots$ y_{p1k} $\cdots$ y_{p1n}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{pj1} $\cdots$ y_{pjk} $\cdots$ y_{pjn}	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{pq1} $\cdots$ y_{pqk} $\cdots$ y_{pqn}

A	B	numrep	resultat
1	1	1	y_{111}
⋮	⋮	⋮	⋮
1	1	n	y_{11n}
⋮	⋮	⋮	⋮
p	1	1	y_{p11}
⋮	⋮	⋮	⋮
p	1	n	y_{p1n}
⋮	⋮	⋮	⋮
1	q	1	y_{1q1}
⋮	⋮	⋮	⋮
1	q	n	y_{1qn}
⋮	⋮	⋮	⋮
p	q	1	y_{pq1}
⋮	⋮	⋮	⋮
p	q	n	y_{pqn}

Modèle statistique

- y_{ijk} pour le couple de modalité (i,j) suit une distribution $\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$.
  - μ_{ij} moyenne inconnue pour la *i-ième* modalité du facteur *A* et pour la *j-ième* modalité du facteur *B*.
  - σ_{ij}^2 variance inconnue pour la *i-ième* modalité du facteur *A* et pour la *j-ième* modalité du facteur *B*.
  - Dans toute la suite, on supposera : $\forall i, j\ \sigma_{ij}^2 = \sigma^2$ où σ^2 est la variance commune inconnue.
- Le modèle statistique peut alors s'écrire comme suit.

$$\boldsymbol{y_{ijk} = \mu_{ij} + \varepsilon_{ijk}}$$

- ε_{ijk} est l'erreur aléatoire (ou résidu) qui contient aussi les erreurs de mesures, les variations aléatoires dues à l'individus tiré.
- $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$

Les effets

Avec ce modèle on étudie deux types d'effets.

- Les **effets principaux**.
  - **effet additif principal** dû à la modalité *i* du facteur *A* : $a_i = \mu_{i.} - \mu_{..}$
  - **effet additif principal** dû à la modalité *j* du facteur *B* : $b_j = \mu_{.j} - \mu_{..}$
    Avec

La moyenne de tous les résultats possibles sous l'effet de la modalité *i* du facteur *A* :
$\mu_{i.} = \frac{1}{q}\sum_{j=1}^{q} \mu_{ij}$
La moyenne de tous les résultats possibles sous l'effet de la modalité *j* du facteur *B* :
$\mu_{.j} = \frac{1}{p}\sum_{i=1}^{p} \mu_{ij}$
La moyenne de tous les résultats possibles $\mu_{..} = \frac{1}{pq}\sum_{j=1}^{q}\sum_{i=1}^{p} \mu_{ij}$

- L'**interaction**.

L'**interaction** due à l'effet conjugué de la modalité *i* du facteur *A* et de la modalité *j* du facteur *B* : $(ab)_{ij} = (\mu_{ij} - \mu_{.j}) - (\mu_{i.} - \mu_{..}) = (\mu_{ij} - \mu_{i.}) - (\mu_{.j} - \mu_{..})$

Avec

L'effet additif dû à la modalité *i* du facteur *A* quand le facteur *B* a la modalité *j* : $\mu_{ij} - \mu_{.j}$

L'effet additif dû à la modalité *j* du facteur *B* quand le facteur *A* a la modalité *i* : $\mu_{ij} - \mu_{i.}$

Le **modèle** peut donc aussi s'écrire comme suit :

$$y_{ijk} = \mu_{..} + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$

toujours avec

$\varepsilon_{ijk} \sim\sim \mathcal{N}(0, \sigma^2)$; $\sum_{i=1}^{p} a_i = 0$; $\sum_{j=1}^{q} b_j = 0$; $\sum_{i=1}^{p}(ab)_{ij} = 0$; $\sum_{j=1}^{q}(ab)_{ij} = 0$;

Dans R cette formule est écrite ainsi : `y ~ A*B` ou encore `y ~ A + B + A:B`

Interprétation de la formule : le résultat observé est une moyenne générale + un effet principal dû à *A* + un effet principal dû à *B* + un effet supplémentaire *i.e.* interaction due à l'effet conjugué des facteurs *A* et *B* + un terme aléatoire non explicable par les facteurs *A* et *B*.

- Propositions importantes :
  - Si $\forall i\ a_i = 0$ alors **il n'y a pas d'effet principal du facteur *A*** sur la moyenne des valeurs de la variable.

- Si $\forall j\ b_j = 0$ alors **il n'y a pas d'effet principal du facteur *B*** sur la moyenne des valeurs de la variable.

- Si $\forall i, j\ (ab)_{ij} = 0$ alors $(\mu_{ij} - \mu_{.j}) - (\mu_{i.} - \mu_{..}) = 0$ => $\mu_{ij} - \mu_{.j} = \mu_{i.} - \mu_{..}$: effet additif dû à la modalité *i* du facteur *A* quand le facteur *B* a la modalité *j* est égal à l'effet additif principal dû à la modalité *i* du facteur *A* c.a.d. que l'effet du facteur *A* ne dépend pas des modalités de *B*.

- Si $\forall i, j\ (ab)_{ij} = 0$ alors $(\mu_{ij} - \mu_{i.}) - (\mu_{.j} - \mu_{..}) = 0$ => $\mu_{ij} - \mu_{i.} = \mu_{.j} - \mu_{..}$: effet additif dû à la modalité *j* du facteur *B* quand le facteur *A* a la modalité *i* est égal à l'effet additif principal dû à la modalité *j* du facteur *B* c.a.d. que l'effet du facteur *B* ne dépend pas des modalités de *A*.

- Dans les deux cas, on dit qu'il n'y a pas d'interaction.

Les tests à effectuer

- **Absence d'interaction** contre présence d'interaction c.a.d
  $H_0(\forall i, j\ (ab)_{ij} = 0)$ contre H_1(il existe un $(ab)_{ij} \neq 0$)
- **Absence d'effet principal du facteur *A*** contre la présence d'effet principal du facteur *A* c.a.d. $H_0(\forall i\ a_i = 0)$ contre H_1(il existe un $a_i \neq 0$)
- **Absence d'effet principal du facteur *B*** contre la présence d'effet principal du facteur *B* c.a.d. $H_0(\forall j\ b_j = 0)$ contre H_1(il existe un $b_j \neq 0$)

Le rejet des tests arrivera dans les situations suivantes :

- Rejet de l'**Absence d'interaction** si :
  - la variabilité due à l'interaction est "trop" supérieure à la variabilité résiduelle ;
  - le rapport entre la variabilité due à l'interaction et la variabilité résiduelle est "trop" supérieur à 1 ;
  - pour tenir compte du nombre de modalités des facteurs et du nombre total d'expérience, on pondère avec des degrés de libertés.

- Rejet de l'**Absence d'effet principal du facteur** si :
  - la variabilité due à l'effet principal du facteur est "trop" supérieure à la variabilité résiduelle ;
  - le rapport entre la variabilité due à l'effet principal du facteur et la variabilité résiduelle est "trop" supérieur à 1 ;
  - pour tenir compte du nombre de modalités des facteurs et du nombre total d'expérience, on pondère avec des degrés de libertés.

Les calculs

- moyenne observée pour le couple de modalités *(i,j)* : $\overline{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}$
- moyenne générale observée pour la modalité *i* du facteur *A* : $\overline{y}_{i..} = \frac{1}{q}\sum_{j=1}^{q} y_{ij.}$
- moyenne générale observée pour la modalité *j* du facteur *B* : $\overline{y}_{.j.} = \frac{1}{p}\sum_{i=1}^{p} y_{ij.}$
- moyenne générale observée : $\overline{y}_{...} = \frac{1}{p}\sum_{i=1}^{p} \overline{y}_{i..} = \frac{1}{q}\sum_{j=1}^{q} \overline{y}_{.j.}$
- effet principal estimé pour la modalité *i* du facteur *A* : $\hat{a}_i = \overline{y}_{i..} - \overline{y}_{...}$
- effet principal estimé pour la modalité *j* du facteur *B* : $\hat{b}_j = \overline{y}_{.j.} - \overline{y}_{...}$
- interaction estimée due au couple *(i,j)* :

$$\widehat{(ab)}_{ij} = (\overline{y}_{ij.} - \overline{y}_{.j.}) - (\overline{y}_{i..} - \overline{y}_{...}) = \overline{y}_{ij.} - (\overline{y}_{...} + \hat{a}_i + \hat{b}_j)$$

- résidu estimé pour l'observation *(i,j,k)* : $\widehat{\varepsilon_{ijk}} = y_{ijk} - \overline{y}_{ij.}$
- la somme des carrés des écarts totale :

$$SCE_T = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2$$

- la somme des carrés des écarts du facteur *A* :

$$SCE_A = nq\sum_{i=1}^{p}\widehat{a_i}^2$$

- la somme des carrés des écarts du facteur *B* :

$$SCE_B = np\sum_{j=1}^{q}\widehat{b_j}^2$$

- la somme des carrés des écarts de l'interaction :

$$SCE_{AB} = n\sum_{i=1}^{p}\sum_{j=1}^{q}\widehat{(ab)_{ij}}^2$$

- la somme des carrés des écarts :
  - sommes des carrés due au couple *(i,j)* : $SCE_{ij} = \sum_{k=1}^{n}\widehat{\varepsilon_{ijk}}^2$
  - somme des carrés des écarts résiduelle : $SCE_R = \sum_{i=1}^{p}\sum_{j=1}^{q} SCE_{ij}$

- équation d'analyse de la variance :
  - $SCE_T = SCE_A + SCE_B + SCE_{AB} + SCE_R$
  - la variabilité totale est décomposée en une variabilité due au facteur A, une variabilité due au facteur B, une variabilité due à l'interaction AB, et une variabilité dite résiduelle.

- les degrés de liberté :
  - $d.d.l._T = npq - 1$
  - $d.d.l._A = p - 1$
  - $d.d.l._B = q - 1$
  - $d.d.l._{AB} = (p-1)(q-1)$
  - $d.d.l._R = (n-1)pq$
  - On a : $d.d.l._T = d.d.l._A + d.d.l._B + d.d.l._{AB} + d.d.l._R$
- les carrés moyens CM :
  - $CM_A = \frac{SCE_A}{d.d.l._A}$
  - $CM_B = \frac{SCE_B}{d.d.l._B}$
  - $CM_{AB} = \frac{SCE_{AB}}{d.d.l._{AB}}$
  - $CM_R = \frac{SCE_R}{d.d.l._R}$
- les F_{obs} ou Test F :
  - $F_A = \frac{CM_A}{CM_R}$
  - $F_B = \frac{CM_B}{CM_R}$
  - $F_{AB} = \frac{CM_{AB}}{CM_R}$
- les p-valeurs $Pr\{F \geq F_{obs}\}$
  - $p_A = Pr\{F \geq F_A\}$ où $F \sim \mathcal{F}(p-1, (n-1)pq)$
  - $p_B = Pr\{F \geq F_B\}$ où $F \sim \mathcal{F}(q-1, (n-1)pq)$
  - $p_{AB} = Pr\{F \geq F_{AB}\}$ où $F \sim \mathcal{F}((p-1)(q-1), (n-1)pq)$

Ce qui nous permet de compléter le **tableau ANOVA-II** :

	SCE	$d.d.l.$	CM	F_{obs}	p-valeur
Tot.	SCE_T	$d.d.l._T$			
Fact. A	SCE_A	$d.d.l._A$	CM_A	F_A	p_A
Fact. B	SCE_B	$d.d.l._B$	CM_B	F_B	p_B
Inter. AB	SCE_{AB}	$d.d.l._{AB}$	CM_{AB}	F_{AB}	p_{AB}
Res.	SCE_R	$d.d.l_R$	CM_R		

Réalisation du test et conclusion

- Tester l'intéraction :

  - $H_0(\forall i,j\ (ab)_{ij} = 0)$ contre H_1 ("il existe un $(ab)_{ij} \neq 0$")
  - $F_{th} = F_{1-\alpha}(d.d.l._{AB}\ , d.d.l._R\ ) = F_{1-\alpha}((p-1)(q-1), (n-1)pq)$
  - Si $F_{AB} \leq F_{th}$ ou si $p_{AB} \geq \alpha$
    - non rejet de H_0 ;
    - les données ne permettent pas de conclure à la présence d'une **intéraction**.

- Tester l'effet principal du facteur A

  - $H_0(\forall i\ a_i = 0)$ contre H_1 ("il existe un $a_i \neq 0$")
  - $F_{th} = F_{1-\alpha}(d.d.l._A\ , d.d.l._R\ ) = F_{1-\alpha}(p-1, (n-1)pq)$
  - Si $F_A > F_{th}$ ou si $p_A < \alpha$
    - rejet de H_0 ;
    - il y a un effet **principal** du facteur A.
  - Si $F_A \leq F_{th}$ ou si $p_A \geq \alpha$
    - non rejet de H_0 ;
    - les données ne permettent pas de conclure à un effet **principal** du facteur A.

- Tester l'effet principal du facteur B

  - $H_0(\forall j\ b_j = 0)$ contre H_1 ("il existe un $b_j \neq 0$")
  - $F_{th} = F_{1-\alpha}(d.d.l._B\ , d.d.l._R\ ) = F_{1-\alpha}(q-1, (n-1)pq)$
  - Si $F_B > F_{th}$ ou si $p_B < \alpha$
    - rejet de H_0 ;
    - il y a un effet **principal** du facteur B.
  - Si $F_B \leq F_{th}$ ou si $p_B \geq \alpha$
    - non rejet de H_0 ;
    - les données ne permettent pas de conclure à un effet **principal** du facteur B.

Travaux pratiques : Étude disparité entre des fromages

Un producteur de producteur de fromages s'intéresse à la teneur en pH au coeur du fromage. Ces fromages sont fabriqués à partir de 3 lignes de production L_1, L_2, L_3 et à partir du lait provenant de 5 sortes de citernes C_1, C_2, C_3, C_4, C_5, chaque citerne pouvant alimenter n'importe laquelle des lignes de productions. Il s'aperçoit d'une disparité de pH entre les fromages et il ne sait pas s'il doit mettre cette disparité sur le compte d'un effet "type de ligne de production" ou sur compte d'un effet "type de citerne" ou au compte des deux effets ?

On va réaliser une analyse de variance **à deux facteurs** pour tenter de répondre à la question.

Lire les données

Pour chaque couple (L_i, C_j), on préléve 2 fromages et en mesure le pH.

	C_1	C_2	C_3	C_4	C_5
L_1	5.5	6.2	5.4	5.6	6.2
	5.3	6.2	5.2	5.4	6.0
L_2	5.5	6.4	5.4	5.4	6.0
	5.3	6.2	5.4	5.4	6.0
L_3	5.6	6.0	5.3	5.6	6.3
	5.2	6.2	5.1	5.5	6.1

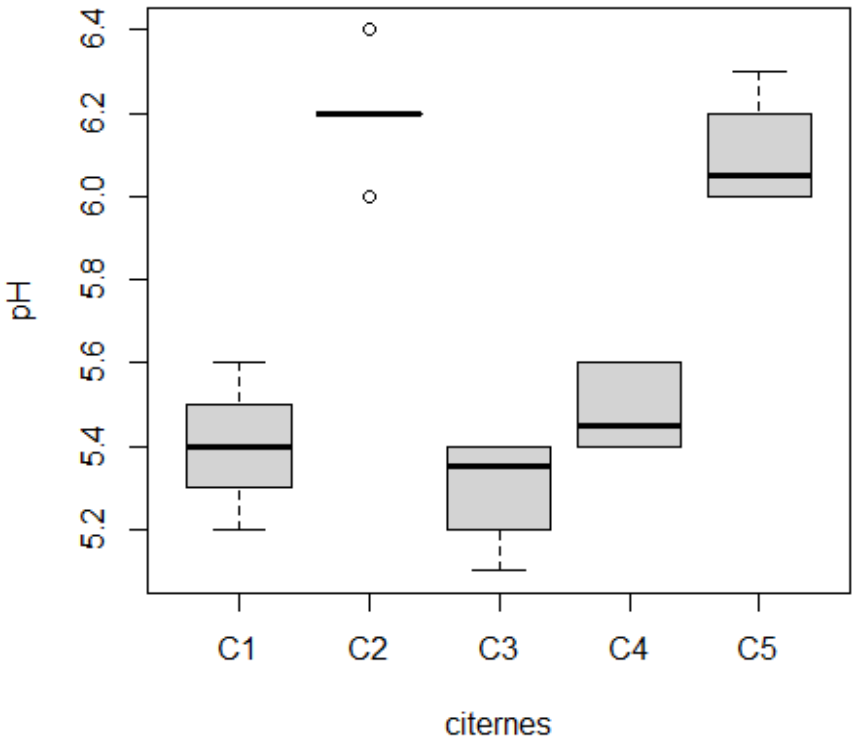
- Construire le `data.frame` pour l'analyse de variance (avec les bons formats).

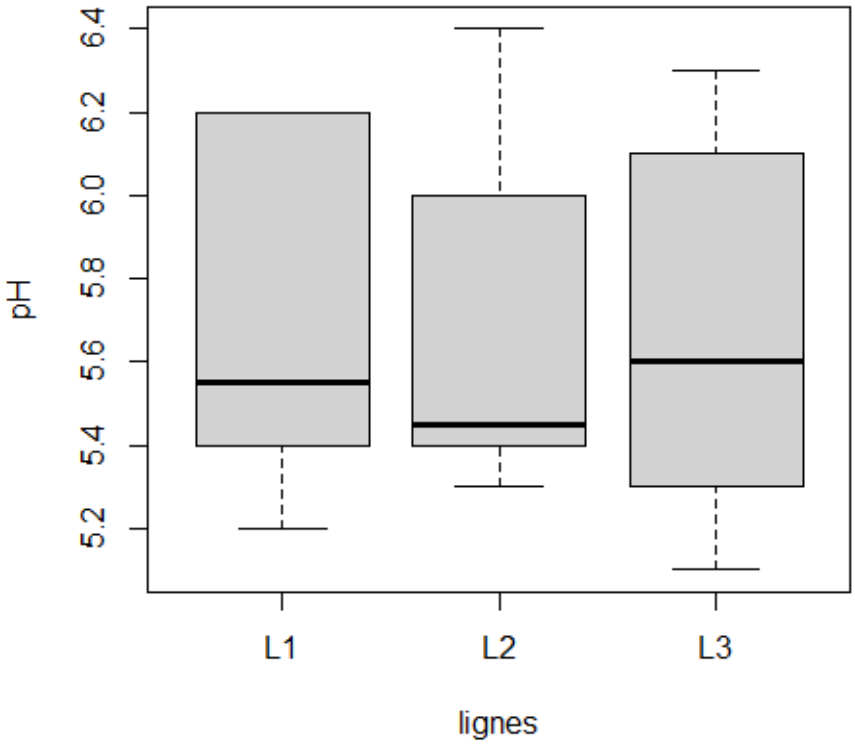
```
fromages <- data.frame(
  "pH" = c(5.5,6.2,5.4,5.6,6.2,5.3,6.2,5.2,5.4,6,5.5,6.4,5.4,5.4,6,5.3,6.2,5.
4,5.4,6,5.6,6,5.3,5.6,6.3,5.2,6.2,5.1,5.5,6.1),
  "lignes" = as.factor(rep(c("L1","L2","L3"),c(10,10,10))),
  "citernes" = factor(x = rep(paste0("C", 1:5),6), levels = paste0("C", 1:5))
)
```

Visualiser et résumer les données

- Visualiser les données.

```
## base / stats
par(mfrow = c(2, 1))
plot(pH ~ citernes*lignes, data = fromages)
```


pH
6.4
6.2
6.0
5.8
5.6
5.4
5.2
C1
C2
C3
C4
C5
citernes

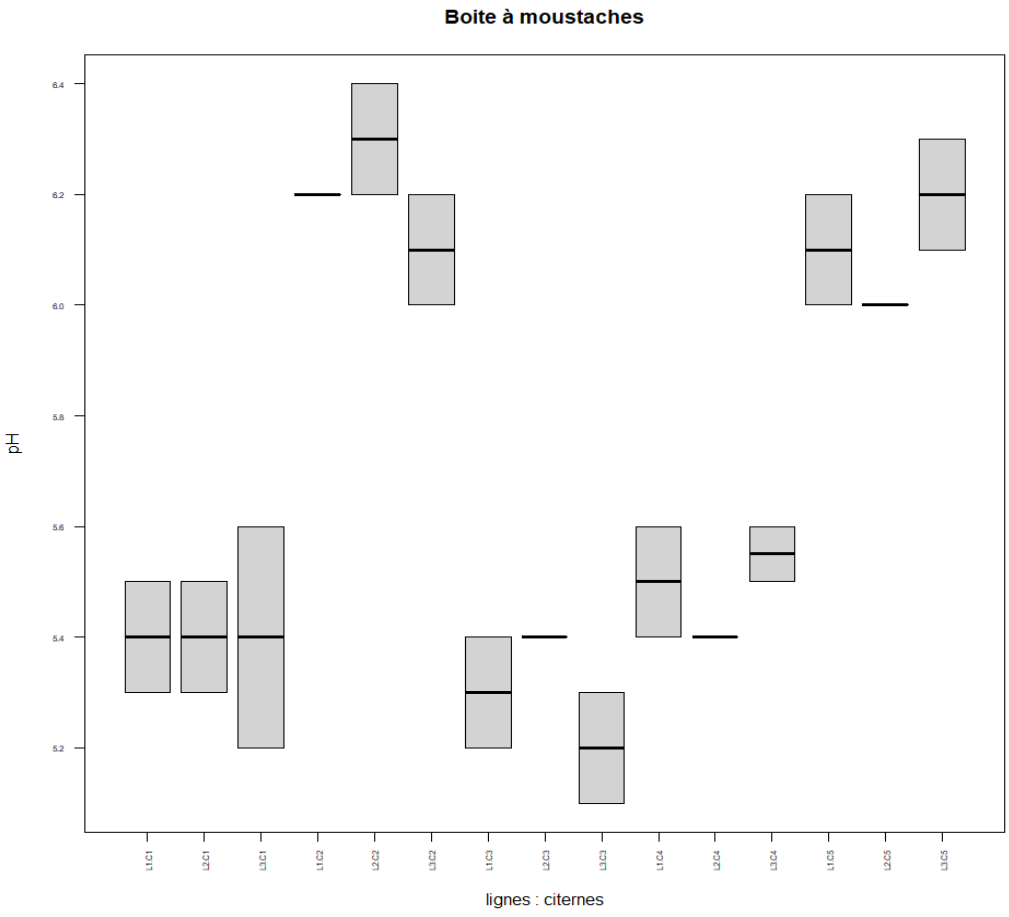

pH
6.4
6.2
6.0
5.8
5.6
5.4
5.2
L1
L2
L3
lignes

`plot(pH ~ citernes*lignes)` construit :

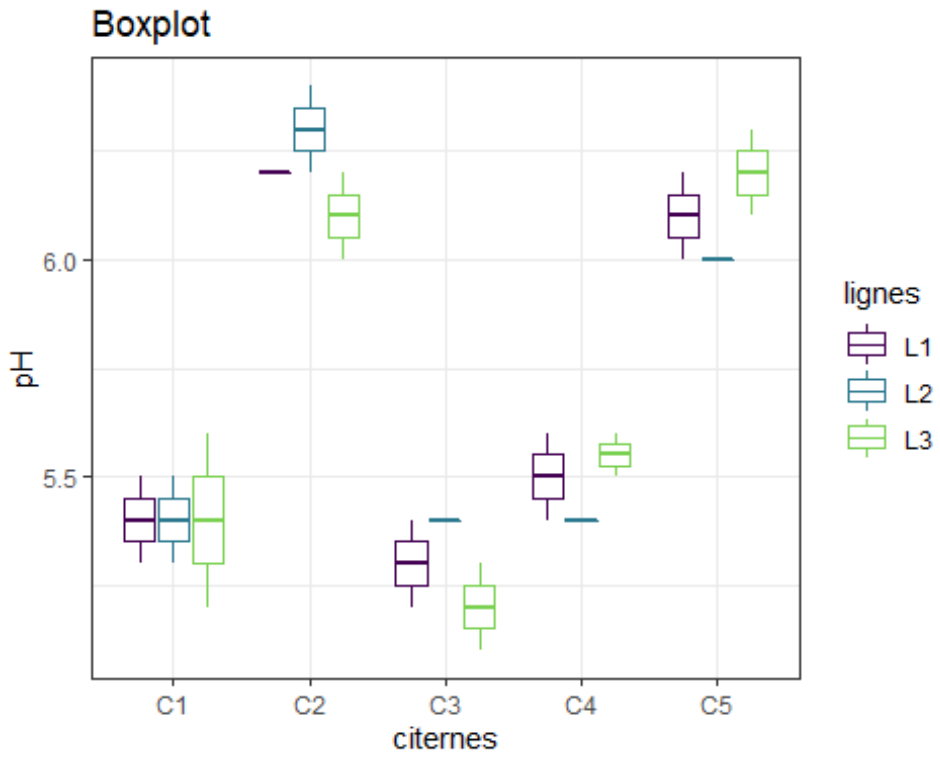
- pour chaque niveau du facteur de variation citernes, les boxplot des observations correspondant à ce niveau du facteur ; si les boxplots sont décalés, on peut soupçonner un effet principal du facteur ; ici les boxplot sont décalés, on soupçonne un effet du facteur citernes ;
- pour chaque niveau du facteur de variation lignes, les boxplot des observations correspondant à ce niveau du facteur ; ici les boxplot ne sont pas trop décalés, on soupçonne l'absence d'effet principal du facteur lignes.

On peut aussi afficher directement les deux facteurs sur un même graphique comme suit :

```
## base / stats
par(mfrow = c(1, 1))
boxplot(
  formula = pH ~ lignes*citernes,
  data = fromages,
  cex.axis = 0.5, # label size
  las = 2, # rotate label to read them properly
  main = "Boite à moustaches"
)
```



```r
## avec ggplot2
ggplot(
  data = fromages,
  mapping = aes(x = citernes, y = pH, color = lignes)
) +
  geom_boxplot() +
  scale_color_viridis_d(end = 0.8) +
  theme_bw() +
  labs(title = "Boxplot")
```



- Résumer les données pour chaque groupe :

```r
## base / stat
n_group <- aggregate(formula = pH ~ citernes*lignes, data = fromages, FUN = l
ength)
names(n_group)[3] <- "n"
mean_group <- aggregate(formula = pH ~ citernes*lignes, data = fromages, FUN
= mean)
names(mean_group)[3] <- "mean"
sd_group <- aggregate(formula = pH ~ citernes*lignes, data = fromages, FUN =
sd)
names(sd_group)[3] <- "sd"
na_group <- aggregate(formula = pH ~ citernes*lignes, data = fromages, FUN =
function(x) sum(is.na(x)))
names(na_group)[3] <- "n_NA"
```

```r
datalist <- list(n_group, mean_group, sd_group, na_group)
my_tab_summar <- Reduce(
  f = function(x,y) {
    merge(x, y, by = c("citernes", "lignes"))
  },
  x = datalist
)

## ou avec dplyr
fromages %>%
  dplyr::group_by(lignes, citernes) %>%
  dplyr::summarise(
    n = n(),
    mean = mean(pH), # /!\ mean(, na.rm = TRUE)
    sd = sd(pH), # /!\ sd(, na.rm = TRUE)
    NA_rendement = sum(is.na(pH))
  ) %>%
  dplyr::ungroup()

## `summarise()` regrouping output by 'lignes' (override with `.groups` argum
ent)

## # A tibble: 15 x 6
##    lignes citernes     n  mean     sd NA_rendement
##    <fct>  <fct>    <int> <dbl>  <dbl>        <int>
##  1 L1     C1           2  5.4  0.141            0
##  2 L1     C2           2  6.2  0                0
##  3 L1     C3           2  5.3  0.141            0
##  4 L1     C4           2  5.5  0.141            0
##  5 L1     C5           2  6.1  0.141            0
##  6 L2     C1           2  5.4  0.141            0
##  7 L2     C2           2  6.3  0.141            0
##  8 L2     C3           2  5.4  0                0
##  9 L2     C4           2  5.4  0                0
## 10 L2     C5           2  6    0                0
## 11 L3     C1           2  5.4  0.283            0
## 12 L3     C2           2  6.1  0.141            0
## 13 L3     C3           2  5.20 0.141            0
## 14 L3     C4           2  5.55 0.0707           0
## 15 L3     C5           2  6.20 0.141            0
```

Test ANOVA

La valeur pH_{ijk} est la valeur du pH du k-ième échantillon quand il est soumis à la i-ième modalité du facteur "lignes" et la j-ième modalité du facteur "citernes".

- μ est la moyenne générale inconnue
- a_i est l'effet principal de la i-ième modalité du facteur citernes
- b_j est l'effet principal de la j-ième modalité du facteur lignes
- $(ab)_{ij}$ est l'effet de l'interaction entre la j-ième modalité du facteur citernes et la j-ième modalité du facteur lignes.

On a alors modèle

$pH_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$ avec $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$

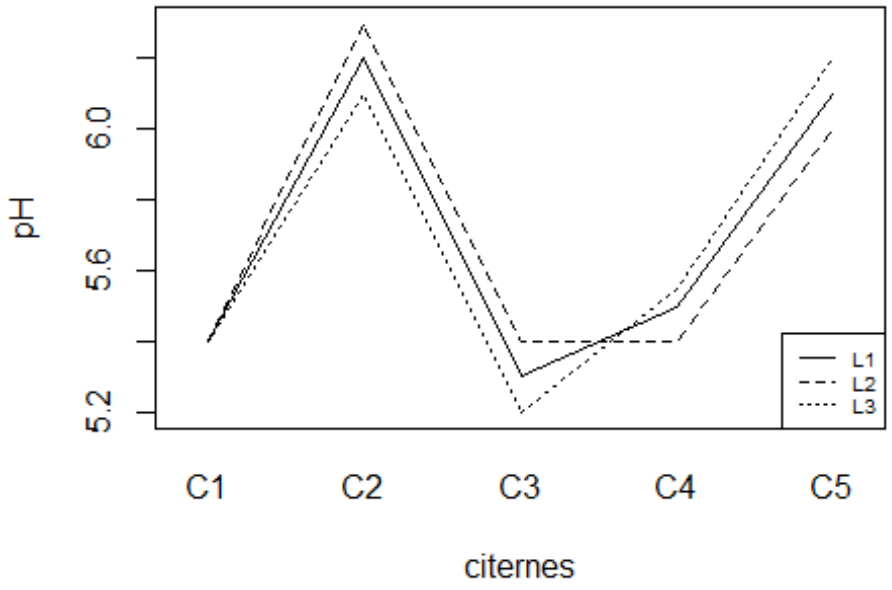
Dans R la formule est : `pH ~ citernes + lignes + citernes:lignes` *i.e.* `pH ~ citernes * lignes`.
Si on ne s'intéresse pas à l'interaction, on peut simplement écrire : `pH ~ citernes + lignes`.

Recherche de l'interaction

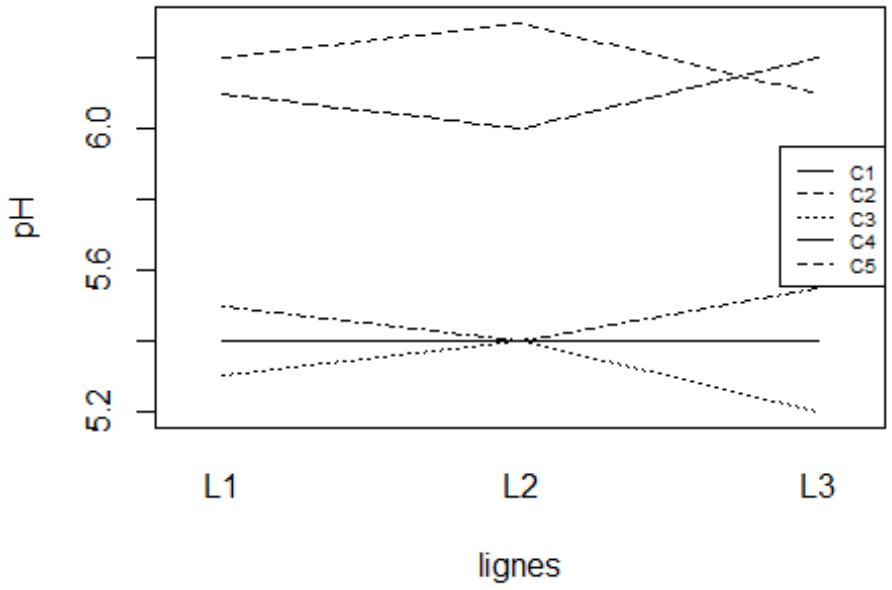
Pour l'utiliser, il faut qu'il y ait au moins 2 facteurs de variation A et B dont on veut étudier l'interaction sur la moyenne d'une variable à expliquer y (ici A = "citernes", "B" = "lignes" et y = "pH").

Si le facteur A possède p modalités et le facteur B possède q modalités, le résultat est une suite de q lignes brisées. À chaque modalité j du facteur B est associé une ligne brisée qui relie, les points $(i, y_{ij.})$. Si les lignes brisées sont "presque" parallèles, on peut soupçonner une absence d'interaction. Il faudra le tester.

```r
interaction.plot(
  x.factor = fromages$citernes,
  trace.factor = fromages$lignes,
  response = fromages$pH,
  trace.label = "lignes", xlab = "citernes", ylab = "pH",
  lty = 1:3,
  legend = FALSE
)
legend("bottomright", legend = c("L1", "L2", "L3"), lty = 1:3, cex = 0.6)
```




```
interaction.plot(
  x.factor = fromages$lignes,
  trace.factor = fromages$citernes,
  response = fromages$pH,
  trace.label = "citernes", xlab = "lignes", ylab = "pH",
  lty = 1:5,
  legend = FALSE
)
legend("right", legend = levels(fromages$citernes), lty = 1:3, cex = 0.6)
```



Le “presque” parallélisme des courbes nous laisse soupçonner l’absence d’interaction, mais cela reste à tester.

Réalisation de l'analyse de la variance deux facteurs

```
fromage.aov <- aov(formula = pH ~ citernes*lignes, data = fromages)
fromage.aov

## Call:
##    aov(formula = pH ~ citernes * lignes, data = fromages)
##
## Terms:
##                 citernes   lignes citernes:lignes Residuals
## Sum of Squares  4.241333 0.000667        0.142667  0.265000
## Deg. of Freedom        4        2               8        15
##
## Residual standard error: 0.132916
## Estimated effects may be unbalanced
```

Pour faire le lien avec le tableau ANOVA-II, l'objet "fromage.aov" affiche un résultat de la forme :

	Facteur A	Facteur B	Interaction A:B	Residuelle
SCE	SCE_A	SCE_B	SCE_{AB}	SCE_R
d.d.l.	ddl_A	ddl_B	ddl_{AB}	ddl_R

Et l'estimation de l'écart-type résiduel : $\hat{\sigma} = \sqrt{CM_R}$.

Pour aller plus loin dans le tableau d'analyse de la variance, il suffit d'afficher le résumé statistique comme suit :

```
summary(fromage.aov)

##                 Df Sum Sq Mean Sq F value   Pr(>F)
## citernes         4  4.241  1.0603  60.019 4.75e-09 ***
## lignes           2  0.001  0.0003   0.019    0.981
## citernes:lignes  8  0.143  0.0178   1.009    0.469
## Residuals       15  0.265  0.0177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ici les informations sont sous la forme suivante :

	d.d.l	SCE	CM	Statistique F	p.value
Facteur A	ddl_A	SCE_A	CM_A	$F_{obs}A$	$P(F_A > F_{obs}A)$
Facteur B	ddl_B	SCE_B	CM_B	$F_{obs}B$	$P(F_B > F_{obs}B)$
Interaction AB	ddl_{AB}	SCE_{AB}	CM_{AB}	$F_{obs}AB$	$P(F_{AB} > F_{obs}AB)$
Résiduelle	ddl_R	SCE_R	CM_R		

avec bien sûr $F_A \sim \mathcal{F}(ddl_A, ddl_R)$, $F_B \sim \mathcal{F}(ddl_B, ddl_R)$, $F_{AB} \sim \mathcal{F}(ddl_{AB}, ddl_R)$

Lorsqu'on s'est fixé un risque α a priori,

- on rejette H_0 : "Absence d'interaction" si la $p - valeur_{AB} < \alpha$;
- on rejette H_0 : "Absence d'effet principal du facteur A" si la $p - valeur_A < \alpha$;
- on rejette H_0 : "Absence d'effet principal du facteur B" si la $p - valeur_B < \alpha$;

Attention : Il faut d'abord examiner l'interaction.

- Si on ne rejette pas l'hypothèse H0 : "absence d'interaction", on regarde ensuite chacun des facteurs. L'effet (ou non) d'un facteur sera le même quelque soit la modalité prise par l'autre facteur.
- Si on rejette H0 : "absence d'interaction", il faut faire très attention sur la conclusion. L'absence d'effet principal d'un facteur ne signifie pas du tout absence d'effet de ce facteur, car cet effet ne peut apparaître que lorsqu'une ou plusieurs modalités de l'autre facteur sont présentes.

Ici, nos conclusions sont les suivantes :

- Pour citernes, $p - value = 4.75e - 09 < 0.05$, donc on rejette $H_0: \forall i \ a_i = 0$ c.a.d. H_0 : "Absence d'effet principal citernes". On dira qu'il y a un effet principal citernes.
- Pour lignes, $p - value = 0.981 > 0.05$, donc on ne rejette $H_0: \forall j \ b_j = 0$ c.a.d. H_0 : "Absence d'effet principal lignes". On dira, **par abus de langage***, qu'il n'y pas d'effet principal lignes.
- Pour ligne:citerne, $p - value = 0.469 > 0.05$, on ne rejette pas $H_0: \forall i \forall j \ (ab)_{ij} = 0$ c.a.d. H_0 : "Absence d'interaction". On dira, **par abus de langage***, qu'il n'y pas d'interaction.
- Puisqu'il y a un effet principal citernes, il y a un effet citernes. Puisqu'il n'y a pas d'interaction, cet effet est le même quelque soit le type de lignes.
- Puisqu'il n'y a pas d'effet principal lignes et qu'il n'y a pas d'interaction, il n'y a pas d'effet lignes et cet absence d'effet est la même quelque soit la citerne.

** Attention ici cet abus de langage peut de faire car notre test est bilatérale et que ses deux conclusions sont diamétralement opposées.*

Estimations

Estimation des paramètres du modele

Quand on écrit le modèle ainsi $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$;

Les paramètres inconnus sont les μ_{ij}, mais on peut s'interesser aussi $\mu_{i.}$, $\mu_{.j}$ ou encore $\mu_{..}$ la moyenne générale.

Pour obtenir ces informations et faire des comparaisons de moyennes, on peut utiliser l'instruction `model.tables` avec l'option `type = means`.

```
esti.fromage <- model.tables(
  x = fromage.aov,
  type = "means",
  se = TRUE
)
esti.fromage
```

```
## Tables of means
## Grand mean
##
## 5.696667
##
##  citernes
## citernes
##    C1    C2    C3    C4    C5
## 5.400 6.200 5.300 5.483 6.100
##
##  lignes
## lignes
##   L1   L2   L3
## 5.70 5.70 5.69
##
##  citernes:lignes
##         lignes
## citernes L1   L2   L3
##       C1 5.40 5.40 5.40
##       C2 6.20 6.30 6.10
##       C3 5.30 5.40 5.20
##       C4 5.50 5.40 5.55
##       C5 6.10 6.00 6.20
##
## Standard errors for differences of means
##         citernes  lignes citernes:lignes
##          0.07674 0.05944         0.13292
## replic.        6      10               2
```

- `esti.fromage` est une liste à plusieurs composantes

- `esti.fromage$tables` est aussi une liste avec
  - la composante `esti.fromage$tables$"Grand mean"` dont la valeur est $\hat{\mu}_{..}$
  - la composante `esti.fromage$tables$"citernes"` dont la valeur est le vecteur $(\hat{\mu}_{1.}, \hat{\mu}_{2.}, \ldots, \hat{\mu}_{p.})$
  - la composante `esti.fromage$tables$"lignes"` dont la valeur est le vecteur $(\hat{\mu}_{.1}, \hat{\mu}_{.2}, \ldots, \hat{\mu}_{.q})$
  - la composante `esti.fromage$tables$"citernes:lignes"` dont la valeur est la matrice des $\hat{\mu}_{ij}$
- `esti.fromage$"n"` est une liste donnant pour chaque paramètre le nombre d'observations qu'on a du sommer pour obtenir l'estimation
- `esti.fromage$se` est une liste donnant pour chaque facteur l'écart-type estimé de la différence des moyennes entre deux modalités du même facteur et pour l'interaction, l'écart-type estimé de la différence des moyennes entre (i, j) et (i', j')

Estimation des effets

Quand on écrit le modèle ainsi $y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$
où $\sum_{i=1}^{p} a_i = 0, \sum_{j=1}^{q} b_j = 0, \forall i \sum_{j=1}^{q} (ab)_{ij} = 0, \forall j \sum_{i=1}^{p} (ab)_{ij} = 0$;

on peut s'interesser aux effets a_i, b_j et $(ab)_{ij}$.

Pour obtenir ces informations, on peut utiliser l'instruction `model.tables` avec l'option `type = effects`.

```
estieffets.fromage <- model.tables(
  x = fromage.aov,
  type = "effects",
  se = TRUE
)
estieffets.fromage
```

```
## Tables of effects
##
##  citernes
## citernes
##      C1      C2      C3      C4      C5
## -0.2967  0.5033 -0.3967 -0.2133  0.4033
##
##  lignes
## lignes
##        L1        L2        L3
##  0.003333  0.003333 -0.006667
##
##  citernes:lignes
##          lignes
## citernes L1       L2       L3
##       C1 -0.00333 -0.00333  0.00667
##       C2 -0.00333  0.09667 -0.09333
```

```
##       C3 -0.00333  0.09667 -0.09333
##       C4  0.01333 -0.08667  0.07333
##       C5 -0.00333 -0.10333  0.10667
##
## Standard errors of effects
##         citernes  lignes citernes:lignes
##          0.05426 0.04203         0.09399
## replic.        6      10               2
```

Cette instruction donne les estimations : $\widehat{a_1}, \widehat{a_2}, \ldots, \widehat{a_p}, \widehat{b_1}, \widehat{b_2}, \ldots, \widehat{b_q}, \widehat{(ab)}_{ij}$ et les écart-types de ces estimations.

On sait que sous l'hypothèse $H_0(a_i = 0)$ alors $\frac{\hat{a}_i}{e.t.(\hat{a}_i)} \sim T(d.d.l._R)$

Ces renseignements permettent donc de tester séparément chaque a_i (idem pour b_j et pour les $(ab)_{ij}$), mais **attention** on peut très bien ne pas rejeter chacun des p test $H_0: a_i = 0$ et rejetter $H_0: a_i = 0 \forall i$!!!

Diagnostic

Diagnostiquer la normalité des résidus

```r
hist(resid(fromage.aov)) # ou
qqnorm(resid(fromage.aov))
qqline(resid(fromage.aov))
# ou encore plus rapidement
plot(fromage.aov)
```

Les tests sont assez peu sensibles à la non-normalité des résidus sauf quand elle est conjuguée avec des données très déséquilibrées c.a.d. des n_i très différents par niveau de facteur (ou n_{ij} dans le cas de deux facteurs) et des variances inconnues σ_i^2 (ou σ_{ij}^2 dans le cas de deux facteurs) très différentes.

Tester l'égalité des variances

Certains logiciels proposent le test classique de Bartlett mais, contrairement aux tests utilisés en analyse de la variance, il est très sensible à la non-normalité, ce qui est le plus souvent le cas. Le logiciel R propose le test de Levene que l'on exécute ainsi :

```r
library("car")

leveneTest(y = fromages$pH, group = fromages$citernes:fromages$lignes)
```

```
## Warning in anova.lm(lm(resp ~ group)): ANOVA F-tests on an essentially per
fect


## fit are unreliable
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df    F value    Pr(>F)
## group 14 2.5829e+28 < 2.2e-16 ***
##       15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ici la p.value est quasi nulle donc on rejette l'égalité des variances σ_{ij}^2, ce qui n'était guère étonnant vu le nombre d'ex-aequo par case.

N.B. : le package `car` propose aussi une implémentation du test anova avec sa fonction `Anova(mod, type = c("II","III", 2, 3), ...)`.

Test post-hoc

- On a détecté un effet citernes. On cherche à savoir quelles sont les citernes qui donnent les mêmes résultats moyens (s'il en existe). On peut utiliser la méthode de Bonferroni :

```
pairwise.t.test(x = fromages$pH, g = fromages$citernes, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fromages$pH and fromages$citernes
##
##    C1      C2      C3      C4
## C2 6.1e-10 -       -       -
## C3 1.0     5.1e-11 -       -
## C4 1.0     5.8e-09 0.2     -
## C5 9.2e-09 1.0     6.1e-10 1.0e-07
##
## P value adjustment method: bonferroni
```

Cette instruction fait apparaître les p-valeurs (C_i, C_j) où $p.valeur(C_i, C_j)$ est associée au test $H_0 : \mu_{C_i} = \mu_{C_j}$ contre $H_1 : \mu_{C_i} \neq \mu_{C_j}$. On rejette $H_0 : \mu_{C_i} = \mu_{C_j}$ si $p.valeur(C_i, C_j) < \alpha$. Cette valeur $p.valeur(C_i, C_j)$ apparaît au croisement de la colonne C_i et de la ligne C_j.

Ici on rejette $C1 = C2, C1 = C5, C2 = C3, C2 = C4, C3 = C5, C4 = C5$.
Par abus de langage, on "accepte" $C1 = C3, C1 = C4, C2 = C5, C3 = C4$.
On obtient deux groupes $C1, C3, C4$ et $C2, C5$. Dans les publications scientifiques on pourrait présenter les moyennes des groupes a et b ainsi :

C1	C2	C3	C4	C5
5.400000(a)	6.200000(b)	5.300000(a)	5.483333(a)	6.100000(b)

Les moyennes (estimées) pour lesquelles on a accolé la même lettre sont celles pour lesquelles on n'a pas rejeté l'hypothèse d'égalité des moyennes théoriques.

Les utilisateurs s'accordent pour dire que la méthode de Bonferronni est très "conservative" c.a.d. qu'elle ne met pas assez en évidence les différences. C'est pourquoi il lui est préféré la méthode de Holm que l'on exécute par la commande suivante :

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fromages$pH and fromages$citernes
##
##    C1      C2      C3      C4
## C2 5.5e-10 -       -       -
## C3 0.56    5.1e-11 -       -
## C4 0.56    4.0e-09 0.08    -
## C5 5.5e-09 0.56    5.5e-10 5.2e-08
##
## P value adjustment method: holm
```

Ici elle donne la même conclusion.
Pour voir la palette des méthodes possibles, faire `help("p.adjust")`.

Sources

- Le contenu de ce TP s'est basé sur un extrait du support écrit par Christophe Chesneau.
- le livre R Cookbook, 2nd Edition, James (JD) Long, Paul Teetor, 2019-09-26

Pour aller plus loin :

- Il existe aussi la fonction `Anova` provenant du package car : https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/Anova