ANOVA (ANAlysis Of VAriance)

Masoumeh Sajedi

Biostatisticienne, Unité de recherche clinique appliquée (URCA)

Novembre, 2022

Introduction

- ANOVA (ANAlysis Of VAriance): déterminer si deux ou plusieurs moyennes de population sont différentes
- l'ANOVA vs le test t de Student
- Différentes versions de l'ANOVA: à un facteur, à deux facteurs, mixte, à mesures répétées, etc.
    - L'ANOVA à un facteur: les moyennes se rapportent aux différentes modalités d'une seule variable indépendante catégorielle (ou facteur). Celle à deux facteurs est utilisée pour tester l'effet combiné de deux facteurs sur une même variable dépendante.

Test statistique

- Compare la variabilité observée entre les moyennes des différents groupes (la variance *inter-groupe*) et la variance au sein de chaque groupe (la variance *intra-groupe*)
- Rapport:

$$F = \frac{variance\ intergroup}{variance\ intragroup}$$

- Seuil de la distribution de probabilité de Fisher (un seuil basé sur un niveau de signification spécifique, généralement 5 %)

Table des Matières

Objectif et l'hypothèse nulle/alternative

- Déterminer si les mesures sont similaires à travers différentes modalités d'une variable catégorielle.
- Comparer l'impact de différents modalités d'une variable catégorielle sur une variable quantitative continue.

Pour un facteur avec a modalités, les hypothèses nulle et alternative d'une ANOVA sont:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

$$H_1 : \mu_i \neq \mu_j \quad \text{pour au moins une paire}(i, j)$$

Base de données

Durocher, Jill, 2019, "A double-blind, randomized controlled trial on oxytocin route for prevention of postpartum hemorrhage in Argentina", https://doi.org/10.7910/DVN/MDZRKU, Harvard Dataverse, V1, UNF:6:cFZwNd9CHxjepreGPi0ZFg== [fileUNF]

- 543 patients et 148 variables
- Variables: Perte de sang totale à l'arrêt du saignement actif(*perdida_sangre_total*), âge de la femme (*edad*), âge gestationnel (*edadgest*), et groupe d'étude (1:IV infusion et 2:IM Injection)
- Sous-ensemble de données: les 480 participants randomisées.
- Ajoutons une variable catégorielle pour âge de la femme (moins de 19 ans, entre 20 et 30 ans et plus de 30 ans), et âge gestationnel (moins de 38 semaines, entre 38 et 40 semaines et plus de 40 semaines)

Questions

1. La perte de sang totale est-elle différente entre les 3 groupes d'âge de femme?
2. La perte de sang totale est-elle différente entre les 3 groupes d'âge de femme et cela dépend-il de leur groupe d'étude?

Importation et préparation des données

Lire les données, modifier les noms des variables et ajouter une variable catégorielle pour âge de la femme.

```
library(haven)
library(tidyverse)

dataset = read_sav("Gynuity_Argentina_oxytocin_IVIM_PPH_Prevention DB.sav")

dat=dataset%>%
  select(edad,studygrp_AR,perdida_sangre_total,edadgest)%>%
  filter(studygrp_AR %in% c("1","2"))

#modification des noms des variables
colnames(dat)=c("age_femme","groupe_étude","perte_sang_totale","age_gestationnel")

# l'ajout d'une variable catégorielle pour âge de la femme
attach(dat)
age_femme_grp= case_when(age_femme>= 30~ '>=30',
            age_femme > 19  & age_femme < 35 ~ '20-30',
            age_femme <= 19 ~ '<=19')

age_gestationnel_grp= case_when(age_gestationnel>= 40~ '>=40',
            age_gestationnel > 38  & age_gestationnel < 40 ~ '38-40',
            age_gestationnel <= 38 ~ '<=38')

dat$age_femme_grp = factor(age_femme_grp,
                  levels=c("<=19","20-30",">=30"))
dat$age_gestationnel_grp=factor(age_gestationnel_grp,
                                levels=c("<=38","38-40",">=40"))
dat$groupe_étude=as.factor(dat$groupe_étude)
```

Statistiques descriptives

```
group_by(dat, groupe_étude) %>% summarise(
     count = n(),
     mean = mean(perte_sang_totale, na.rm = TRUE),
     sd = sd(perte_sang_totale, na.rm = TRUE))
```

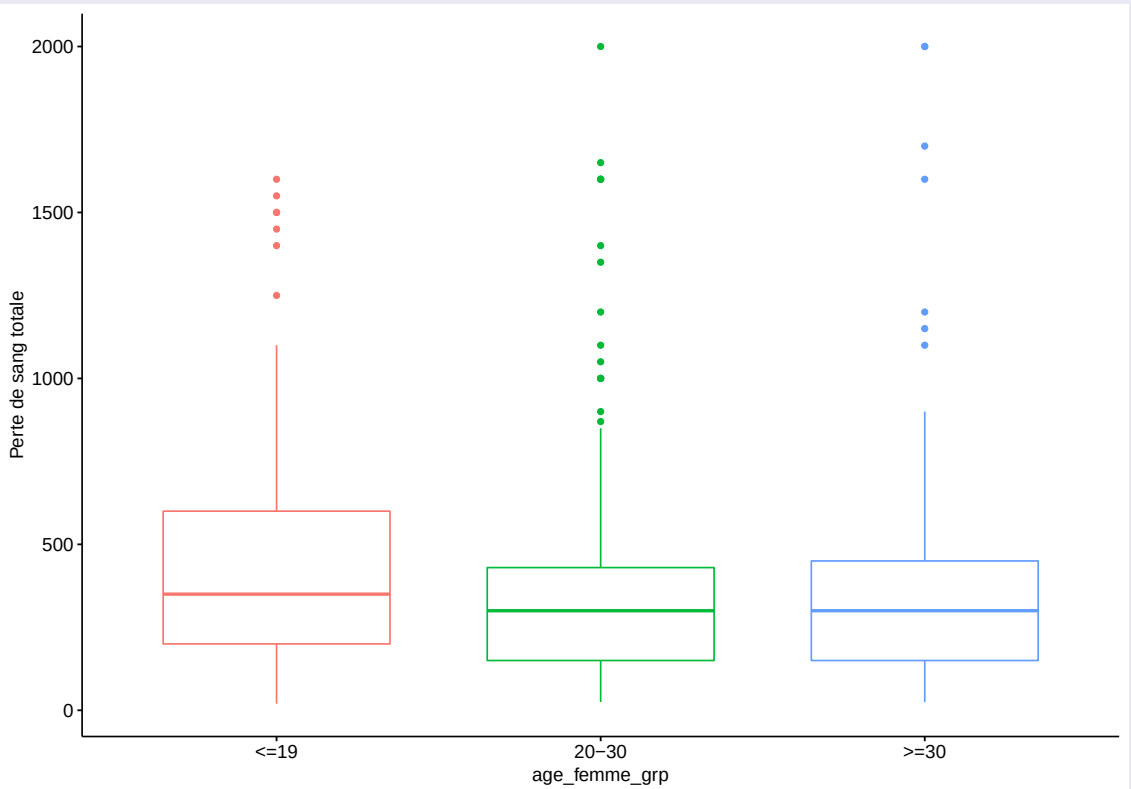
```
## # A tibble: 2 x 4
##   groupe_étude count  mean    sd
##   <fct>        <int> <dbl> <dbl>
## 1 1              239  363.  322.
## 2 2              241  406.  344.
```

```
group_by(dat, age_femme_grp) %>% summarise(
    count = n(),
    mean = mean(perte_sang_totale, na.rm = TRUE),
    sd = sd(perte_sang_totale, na.rm = TRUE))
```

```
## # A tibble: 3 x 4
##   age_femme_grp count  mean    sd
##   <fct>         <int> <dbl> <dbl>
## 1 <=19            106  458.  361.
## 2 20-30           292  352.  294.
## 3 >=30             82  407.  411.
```

Visualisation des données

```
library("ggpubr")
ggboxplot(dat,x="age_femme_grp", y="perte_sang_totale",colo ="age_femme_grp",
          ylab = "Perte de sang totale",xlab="age_femme_grp")+theme(legend.position = "none")
```



Les femmes de moin de 19 ans semblent avoir *Perte de sange totale* le plus éleve, et les femmes entre 20 et 30 ans ont le *Perte de sange totale* le moins élevé.

ANOVA à un facteur

Déterminer s'il existe une différence significative entre les perte de sange totale moyens des femme dans les 3 groupes d'âge.

- Le facteur: *age_femme* qui contient 3 modalités ou groupes

Effectuer l'ANOVA et interpréter les résultats

La fonction R *aov()* peut être utilisée pour ajuster un modèle d'analyse de la variance. La fonction summary.aov() est utilisée pour résumer ce modèle.

```
res_aov <- aov(perte_sang_totale ~ age_femme_grp, data = dat)
summary(res_aov)
```

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## age_femme_grp   2   911524  455762   4.146 0.0164 *
## Residuals     474 52104892  109926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

Interprétation

- La valeur de p est inférieure à 0,05 (p=0.016), nous rejetons l'hypothèse nulle que toutes les moyennes sont égales.
- Au moins une des categories est différente des autres.
- Les résultats d'une ANOVA, ne nous disent pas quel(s) groupe(s) est(sont) différent(s) des autres. Pour tester cela, nous devons utiliser d'autres types de tests, appelés tests post-hoc ou des tests de comparaison par paires multiples.
- Dans le cas où la valeur de p est superieur au niveau de signification de 0,05, nous ne pouvons pas rejeter l'hypothèse nulle que toutes les moyennes sont égales.

Tests Post-Hoc

- D'effectuer une comparaison multiple par paires entre les moyennes des groupes à fin de déterminer si la différence moyenne entre des paires spécifiques de groupes est statistiquement significative
- Deux des tests post hoc les plus courants: le test HSD de Tukey (Tukey Honest Significant Differences) et le test t par paires

Test HSD de Tukey

TukeyHSD(): Comparer tous les groupes entre eux (donc toutes les comparaisons possibles de 2 groupes).

```
TukeyHSD(res_aov)
```

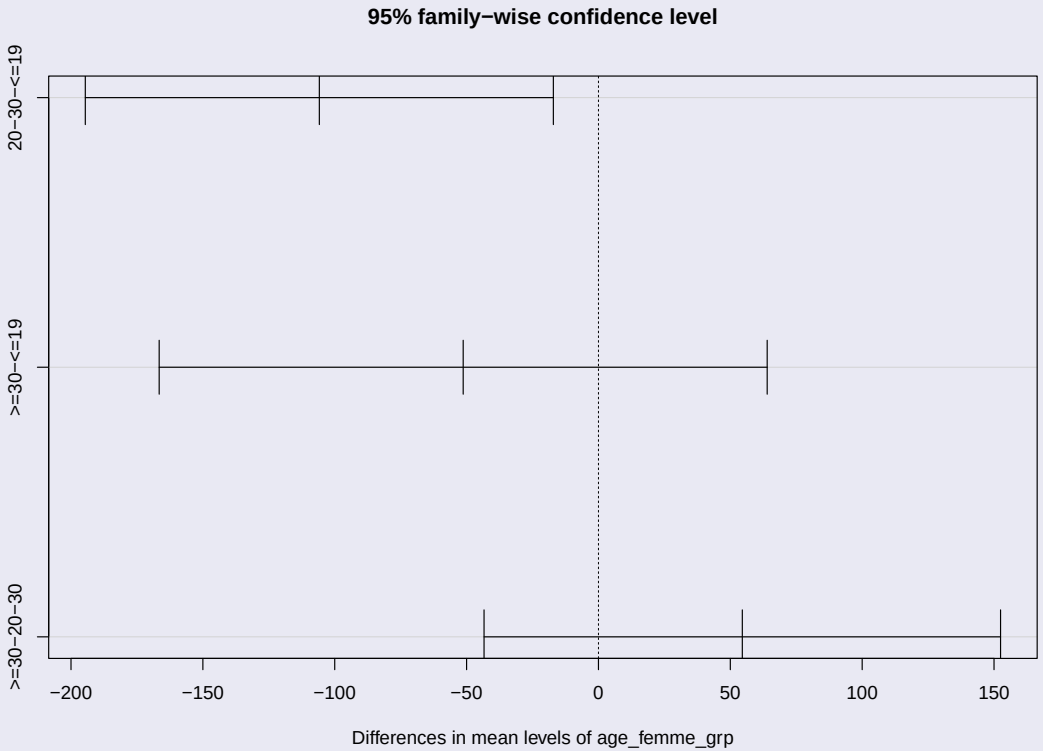
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = perte_sang_totale ~ age_femme_grp, data = dat)
##
## $age_femme_grp
##                    diff        lwr       upr     p adj
## 20-30-<=19 -105.84045 -194.58205 -17.09886 0.0145065
## >=30-<=19   -51.29101 -166.56732  63.98531 0.5481602
## >=30-20-30   54.54945  -43.37773 152.47662 0.3904644
```

- *diff*: signifie la différence entre les moyennes des deux groupes.
- *lwr*, *upr*: signifie la limite inférieure et la limite supérieure de l'intervalle de confiance à 95% (par défaut)
- *p adj*: la valeur de p après ajustement pour les comparaisons multiples.

Il ressort de la sortie que seule la différence entre *20-30 ans* et *<=19 ans* est significative avec une valeur de p ajustée de 0.014.

Visualisation des résultats du test post-hoc

```
plot(TukeyHSD(res_aov))
```



L'intervalle de confiance correspondant à la comparaison des modalites *20-30 ans* et *<=19 ans* ne franchit pas la ligne zéro, ce qui indique que ces deux modalités sont significativement différents.

Test t par paires

pairwise.t.test(): calculer des comparaisons par paires entre les modalités avec des corrections pour les tests multiples

```
pairwise.t.test(dat$perte_sang_totale, dat$age_femme_grp, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  dat$perte_sang_totale and dat$age_femme_grp
##
##       <=19  20-30
## 20-30 0.016 -
## >=30  0.888 0.573
##
## P value adjustment method: bonferroni
```

Les valeurs P sont ajustées à l'aide de la méthode de correction des tests multiples de Bonferroni.

Hypothèses de validité

- L'indépendance: les données doivent être indépendantes entre les groupes et au sein de chaque groupe.
- La normalité des résidus: Dans le cas de petits échantillons, les résidus doivent suivre approximativement une distribution normale. La normalité n'est pas requise lorsque le nombre d'observations dans chaque groupe est important (généralement n>=30).
- Homogénéité de la variance: les variances des différents groupes doivent être égales. Si l'hypothèse d'égalité des variances est rejetée, l'ANOVA de Welch (oneway.test dans R) peut être utilisée.

Note: lorsque les hypothèses ANOVA ne sont pas satisfaites, le test de somme des rangs de Kruskal-Wallis (kruskal.test dans R) est une alternative non paramétrique à l'ANOVA.

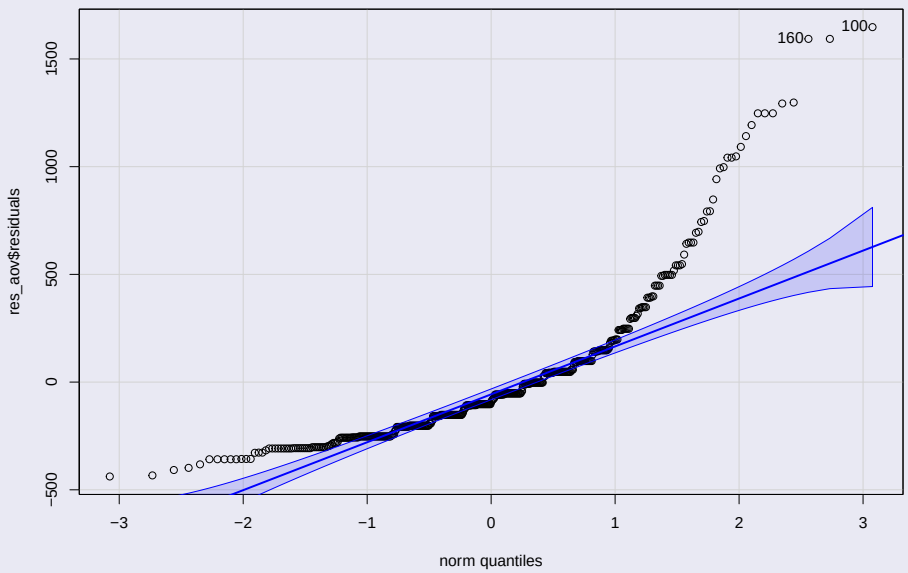
Indépendance

Les mesures sont indépendantes entre les groupes d'âges et nous ne sommes pas en présence de mesures répétées pour chaque sujet.

Normalité

- Visuellement (un histogramme, un QQ-plot)

```
#plot(res_aov, which = 2)
#autre option
library(car)
qqPlot(res_aov$residuals)
```




```
## 100 160
## 100 159
```

Normalité

- Un test de normalité tel que le test de Shapiro-Wilk ou de Kolmogorov-Smirnov

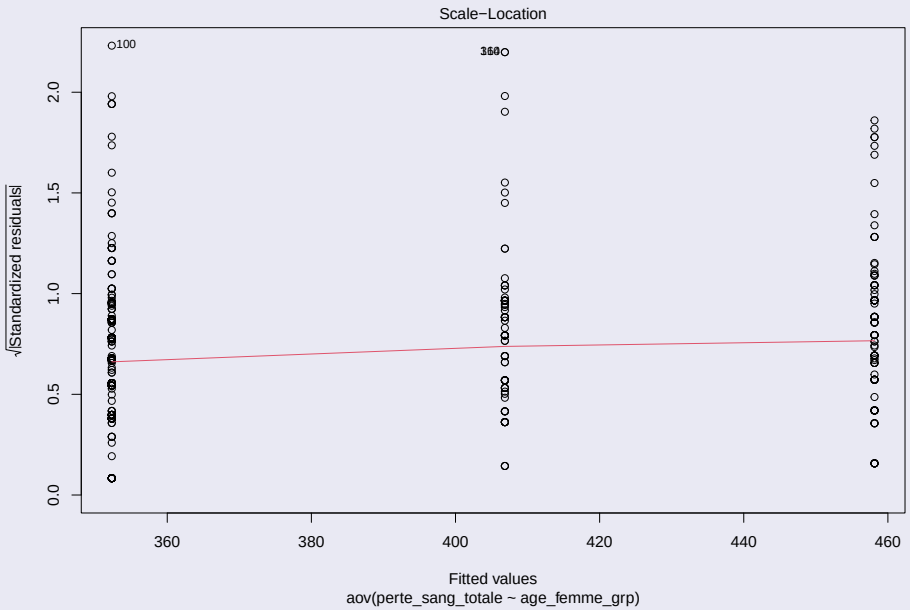
```
# test Shapiro-Wilk test
shapiro.test(res_aov$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov$residuals
## W = 0.78786, p-value < 2.2e-16
```

Homogénéité de la variance

- Le diagramme des valeurs résiduelles en fonction des valeurs ajustées

```
plot(res_aov, which = 3)
```



Il n'y a pas de relations évidentes entre les valeurs résiduelles et les valeurs ajustées (la moyenne de chaque groupe). On peut donc supposer l'homogénéité des variances.

Homogénéité de la variance

- Le test de Levene

```
library(car)
leveneTest(perte_sang_totale ~ age_femme_grp, data = dat)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   2  2.5734 0.07734 .
##       474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La valeur de p étant supérieure au seuil de signification de 0,05, nous ne rejetons pas l'hypothèse nulle, nous ne pouvons donc pas rejeter l'hypothèse que les variances entre les groupes d'âge sont égale.

ANOVA à deux facteur

Évaluer simultanément l'effet de deux variables sur une variable de réponse

Les hypothèses nulles sont:

- Il n'y a pas de différence dans les moyennes des deux facteurs
- Il n'y a pas d'interaction entre les facteurs

Dans notre example, les facteurs sont *groupe_étude* et *age_femme_grp* et ils contiennent respectivement 2 et 3 modalités.

Statistiques descriptives

```
#Génération de tables de fréquences:
table(dat$groupe_étude,dat$age_femme_grp)
```

```
##
##     <=19 20-30 >=30
##   1   50   149   40
##   2   56   143   42
```

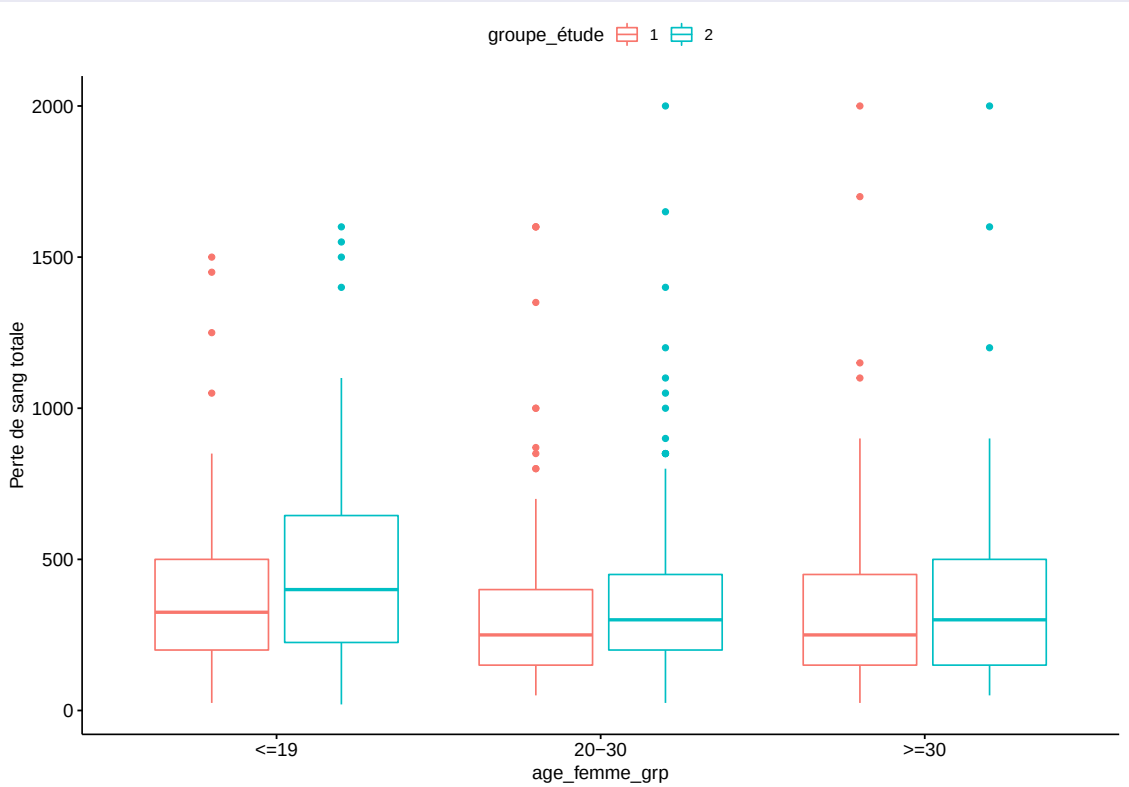
Le nombre de sujets dans chaque groupe n'est pas égale. Nous somme en presence d'un plan déséquilibré.

```
# Calcul de la moyenne et l'écart-type par groupes
group_by(dat, groupe_étude, age_femme_grp) %>% summarise(
    count = n(),
    mean = mean(perte_sang_totale, na.rm = TRUE),
    sd = sd(perte_sang_totale, na.rm = TRUE))
```

```
## # A tibble: 6 x 5
## # Groups:   groupe_étude [2]
##   groupe_étude age_femme_grp count  mean    sd
##   <fct>        <fct>         <int> <dbl> <dbl>
## 1 1            <=19             50  415.  329.
## 2 1            20-30           149  338.  286.
## 3 1            >=30             40  393.  426.
## 4 2            <=19             56  498.  386.
## 5 2            20-30           143  367.  303.
## 6 2            >=30             42  420.  400.
```

Visualisation des données

```
library("ggpubr")
ggboxplot(dat,x = "age_femme_grp", y = "perte_sang_totale",
          color = "groupe_étude",
          ylab = "Perte de sang totale",
          xlab="age_femme_grp")
```



Effectuer l'ANOVA à deux facteurs

- Méthodes: sommes de carrés de type I, de type II et de type III.
- Les trois méthodes donnent le même résultat lorsque le plan est équilibré.
- Résumer l'analyse du modèle de variance:
  - Fonction summary.aov(): pour les plans équilibrés
  - Fonction Anova(): pour les plans déséquilibrés

Modèle additif

Le modèle additif: Il fait l'hypothèse que les deux facteurs sont indépendantes

```
res_aov2 <- aov(perte_sang_totale ~ age_femme_grp + groupe_étude, data = dat)
Anova(res_aov2, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: perte_sang_totale
##                 Sum Sq  Df  F value  Pr(>F)
## (Intercept)   16140698   1 147.0758 < 2e-16 ***
## age_femme_grp   888690   2   4.0489 0.01805 *
## groupe_étude    195928   1   1.7853 0.18214
## Residuals     51908964 473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La groupe d'âge est statistiquement significatif.

Modèle avec effet d'interaction

```
res_aov3 <- aov(perte_sang_totale ~ age_femme_grp * groupe_étude,data = dat)
Anova(res_aov3, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: perte_sang_totale
##                              Sum Sq  Df F value Pr(>F)
## (Intercept)                 8598805   1 78.1128 <2e-16 ***
## age_femme_grp                262107   2  1.1905 0.3050
## groupe_étude                 180150   1  1.6365 0.2014
## age_femme_grp:groupe_étude    60371   2  0.2742 0.7603
## Residuals                 51848593 471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous observerons que l'interaction n'est pas significative, dans ce cas il faut utiliser le modèle additif.

Une interaction significative implique que la valeur moyenne de la variable réponse pour chaque modalité d'un facteur change selon la modalité de l'autre facteur.

Test Post-Hoc

Déterminer si la différence moyenne entre des paires spécifiques des groupes est statistiquement significative

```
TukeyHSD(res_aov2, which = "age_femme_grp")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = perte_sang_totale ~ age_femme_grp + groupe_étude, data = dat)
##
## $age_femme_grp
##                   diff        lwr       upr     p adj
## 20-30-<=19 -105.84045 -194.50921 -17.17169 0.0144085
## >=30-<=19   -51.29101 -166.47271  63.89070 0.5476173
## >=30-20-30   54.54945  -43.29736 152.39626 0.3898614
```

Récapitulatif

- Statistiques descriptives et Visualisation des données: summarise(), ggboxplot()
- Effectuer l'ANOVA: aov(), summary.aov(), Anova
- Hypothèses de validité:
  - Indépendance
  - Normalité: shapiro.test()
  - Homogénéité de la variance: leveneTest()
- Tests Post-Hoc: TukeyHSD(), pairwise.t.test()

Exercice

En utilisant le test ANOVA, répondez à la question suivante:

- La perte de sang totale est-elle différente entre les 3 groupes d'âge gestationnel?

Références

1. **Base de données**: Durocher, Jill, 2019, "A double-blind, randomized controlled trial on oxytocin route for prevention of postpartum hemorrhage in Argentina", Harvard Dataverse, V1, https://doi.org/10.7910/DVN/MDZRKU, UNF:6:cFZwNd9CHxjepreGPi0ZFg==
2. **Packages tidyverse**: https://cran.r-project.org/web/packages/tidyverse/index.html
3. **Package car**: Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/
4. **Package ggplot2**: https://cran.r-project.org/web/packages/ggpubr/index.html

Références

5. **aov RDocumentation**: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aov
6. **TukeyHSD RDocumentation**: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/TukeyHSD