ANOVA à 1 et 2 facteurs
Analyse de variance

Ricco Rakotomalala
Université Lumière Lyon 2

PLAN

ANOVA (ANalysis Of VAriance)
Et
Planification des expériences

Position du problème :
Exemple introductif

On veut connaître l 'effet de trois types de fertilisants sur la croissance des arbres d 'une plantation

1) Principe de l 'expérimentation

☞ *extraire 3 échantillons (groupes) d 'arbres et appliquer chaque fertilisant pour chaque échantillon : comparer ensuite les moyennes de croissance annuelle des arbres*

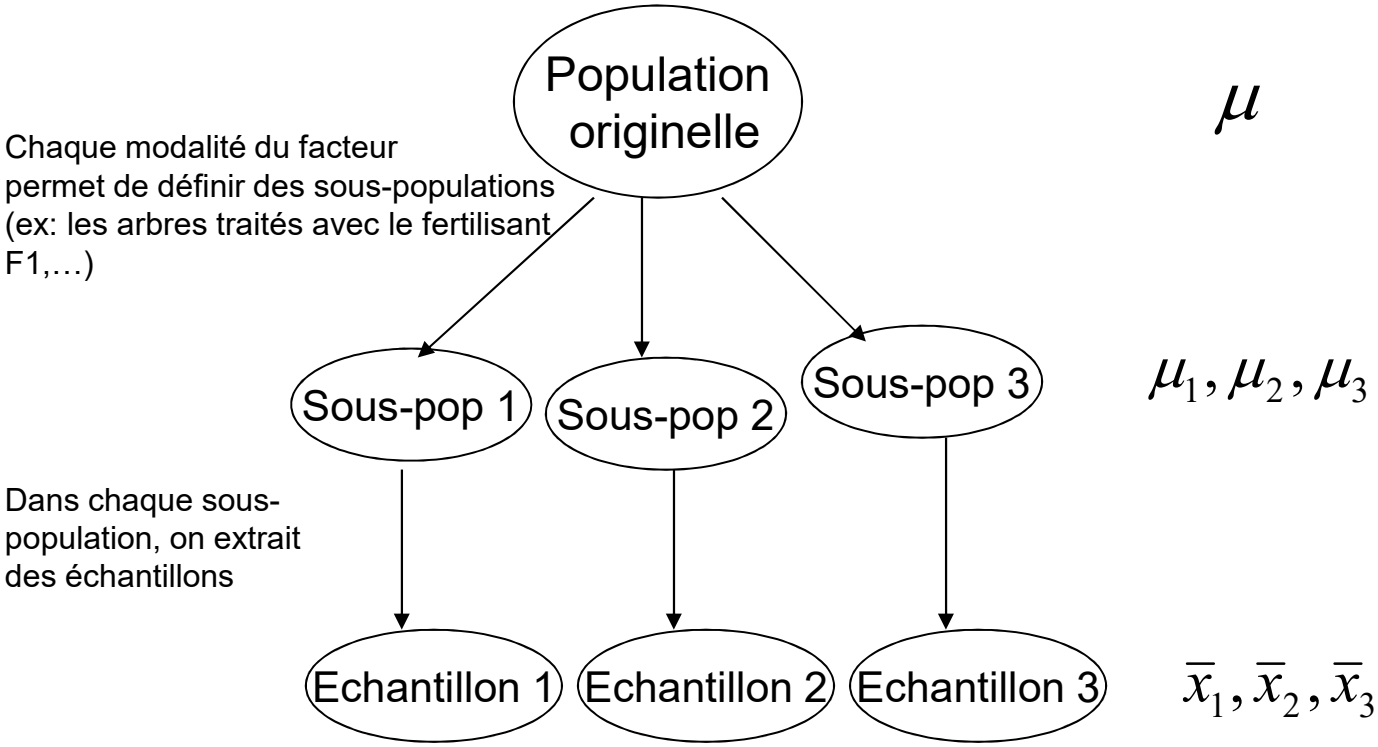
Variable d 'intérêt
(variable dépendante)
en cm/an par exemple
Autres exemples
• rendement d'un paquet d'action
• taux de virus dans le sang

Facteur
(variable indépendante)
type de fertilisant
Autres exemples
• stratégie de placement
• traitement médical

Les domaines d'études sont variés. L'ANOVA s 'applique dès que :
☞ on veut monter une expérimentation
☞ on veut vérifier l 'effet de variables qualitatives sur une variable quantitative

2) Principe statistique

Indicateur mesuré sur la variable dépendante



La problématique de l 'ANOVA consiste à utiliser les moyennes observées sur les échantillons pour conclure à des différences significatives sur les moyennes (espérance mathématique) dans les sous-populations

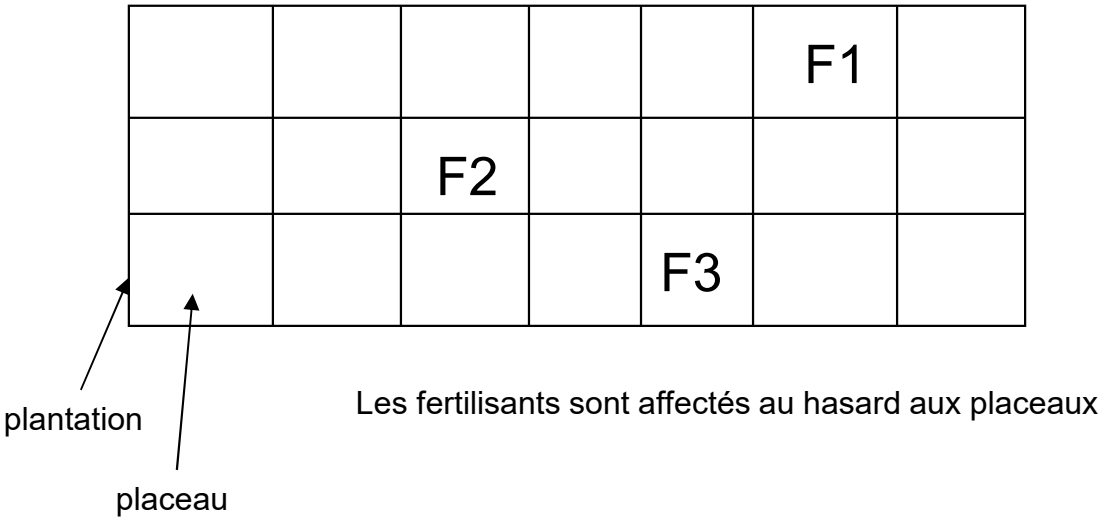
Ex: pour la plantation, tous les fertilisants sont-ils équivalents, ou bien y a-t-il un qui soit meilleur (moins bon) que les autres…?

3) Problèmes pratiques et quelques définitions

A) Plan d 'expérimentation

Dans la pratique, plusieurs problèmes peuvent corrompre les résultats...
ex: on ne peut pas donner un fertilisant à un arbre, il y a des problèmes de diffusion aux arbres voisins (pluie, vent…)

=> Une solution possible serait d 'effectuer un maillage de la plantation (on définit ainsi des placeaux), on applique un fertilisant aux arbres qui sont dans le même placeau



Quelques définitions

- ☞ *individu statistique : un arbre de la plantation*
- ☞ *population : les arbres de la plantation*
- ☞ *échantillon expérimental : les arbres dans les 3 placeaux*
- ☞ *unité expérimentale : un placeau*

Remarque : Dans certains problèmes, les unités expérimentales sont confondues avec les individus statistiques (ex: médecine, le patient est à la fois unité expérimentale et individu statistique)

B) Les facteurs non-contrôlés

Si la plantation est grande, différents facteurs peuvent perturber l' expérimentation

☞ différences climatiques, il peut y avoir des zones de micro-climat

☞ facteurs génétiques : les graines n 'ont peut-être pas tous la même provenance et des arbres voisins (sur le même placeau) ont toutes les chances de provenir du même lot

☞ le sol n 'est pas de même qualité dans toute la plantation

Facteurs non-contrôlés, on sait qu'ils peuvent perturber les résultats mais on ne sait pas les prendre en compte explicitement dans l 'analyse

Le rôle du plan d'expériences est de définir au mieux l 'expérimentation (ex: répartir les traitements sur les placeaux) de manière à annihiler le rôle des facteurs non-contrôlés.

C) Les facteurs de blocs

Certains facteurs (que l 'on connaît cette fois-ci) peuvent perturber les résultats, il appartient au statisticien de les maîtriser au mieux

ex: pour évaluer un médicament, les facteurs de blocs peuvent être l 'âge (jeune, adulte, vieux), le sexe (homme, femme)… si on doit évaluer 2 médicaments, l'un est administré uniquement aux hommes, l'autre aux femmes, la validité des résultats devient sujette à caution...

Même si ce n 'est pas l 'objectif (évaluer l 'efficacité des médicaments selon le sexe), il est important d 'en tenir compte pour évaluer les résultats (qui est de mesurer l'efficacité du médicament sans distinction de sexe)

Il est également possible de supprimer l 'effet de bloc par une plan d 'expériences approprié, par ex. en mettant le même nombre d 'hommes et de femmes dans chaque échantillon

D) Facteurs fixes et facteurs aléatoires

Dans l'exemple des fertilisants, le nombre de modalités du facteur est faible (3 fertilisants qui sont tous connus). Que faire quand il est innombrable, ou que l 'on ne veut en traiter que quelques-uns pour inférer sur les autres

Ex: on veut vérifier qu'il y a un facteur opérateur dans la réparation des roues de voitures dans un garage ayant plusieurs succursales en Europe (les modalités du facteurs sont les employés affectés à ce type de réparation dans les garages)

☞ *on ne va pas traiter tous les employés, il est plus intéressant d 'en sélectionner au* ***hasard*** *et d 'inférer sur le rôle du facteur « opérateur » sur le temps de réparation*

Facteur aléatoire

☞ on utilise un *échantillon des modalités* du facteur (ex: trois médicaments parmi tous les médicaments traitant de la grippe)

☞ *à partir du résultat, on infère sur le reste des autres modalités*

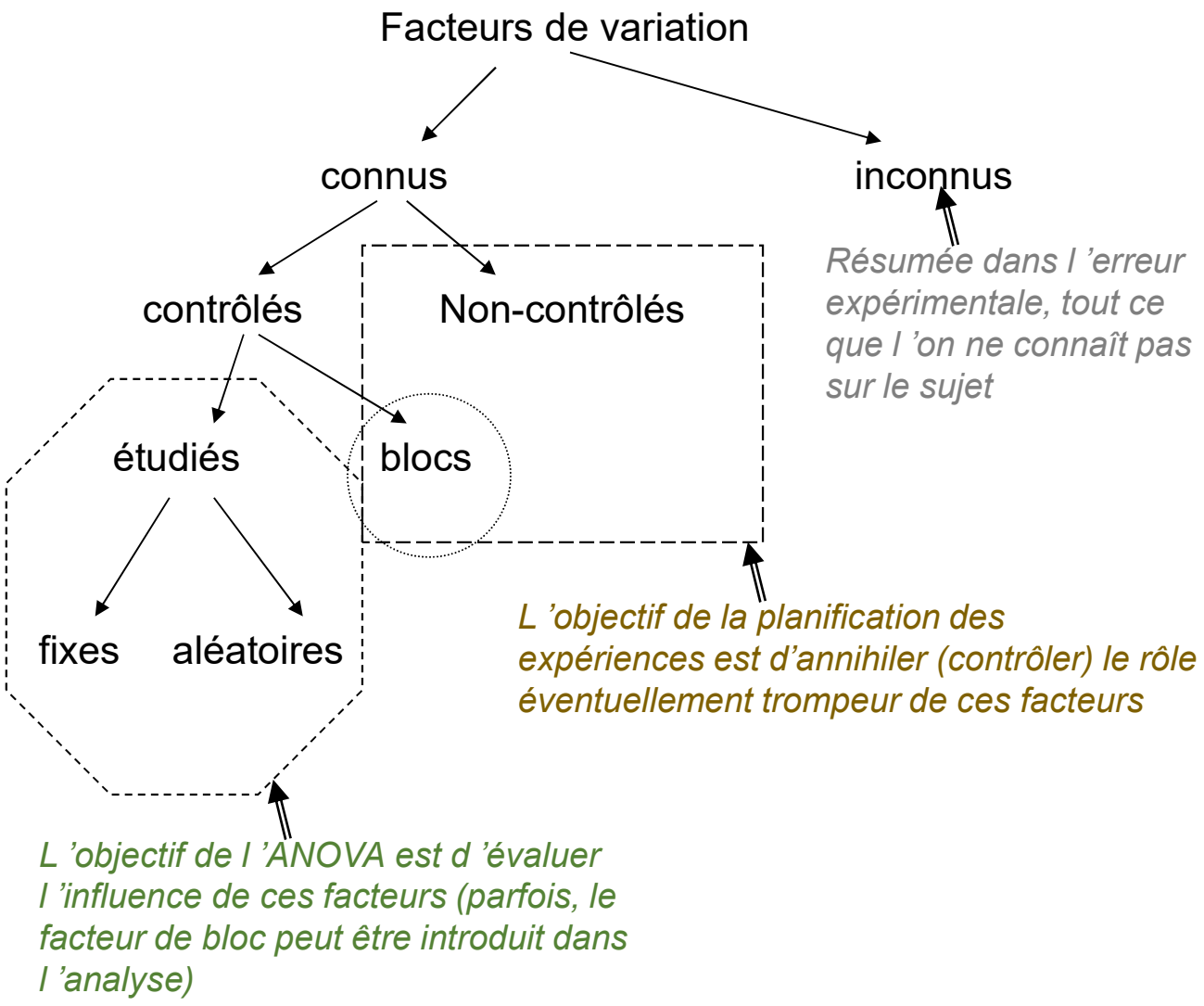
☞ *pour que le résultat soit valide, il est impératif que les modalités utilisés lors du calcul soient extraits de manière aléatoire*

Facteur fixe

☞ on utilise *toutes les modalités ou un sous-ensemble des modalités* du facteur

☞ *les résultats ne sont valables que pour ces modalités*

E) Récapitulatif sur les facteurs de variation



4) Etude prospective - étude rétrospective

Le facteur est **contrôlé - étudié**
on veut mesurer son influence sur la variable d 'intérêt

☞ on est en relation directe avec la planification des expériences, le facteur est manipulé

Le facteur est **contrôlé - bloc**
on veut vérifier son influence sur la variable d 'intérêt sans pour autant avoir à faire une expérimentation

ex: comparer les niveaux de salaire selon le sexe
ex: comparer les salaires d 'embauche selon les écoles d 'origine

5) Extensions : analyse à plusieurs facteurs

Au lieu d 'un facteur, on peut analyser le rôle de deux ou plusieurs facteurs pris conjointement

Analyse des influences individuelles

Analyse des interactions

Ex 1 : fertilisant **et** mode de diffusion des fertilisants
Ex 2 : médicament **et** sexe du patient

On peut combiner facteur étudié et facteur de bloc...

ANALYSE DE VARIANCE
à UN FACTEUR (One-way ANOVA)

1.A) Hypothèse de travail

☞ l 'unité expérimentale est confondue avec l 'individu statistique
☞ le plan est complètement randomisé

Les modalités du facteur sont affectés de manière aléatoire aux unités expérimentales

1.B) Hypothèses stochastiques

☞ les échantillons sont issus d'une population normale (gaussienne) : on parle de test paramétrique
☞ les variances conditionnelles (variances dans chaque sous-population) sont identiques : homoscédasticité
☞ les sous-échantillons sont indépendants

En toute rigueur, on devrait vérifier les deux premières hypothèses. En pratique, l'ANOVA présente une certaine robustesse. On en reparle au point n°6

2) Description des données

☞ P est la population
☞ A est le facteur à étudier (avec p modalités : $A_1, A_2, \ldots, A_p$)
☞ X est la variable d 'intérêt de moyenne μ

Ce qui induit une subdivision de la population en sous-population $P_1, P_2, \ldots, P_p$

Dans chaque sous-population, on a $\mu_1, \mu_2, \ldots, \mu_p$

☞ E est l 'échantillon total, d 'effectif n
☞ $E_1, \ldots E_p$ sous les sous-échantillons relatifs aux sous-populations, d 'effectifs $n_1 \ldots, n_p$
☞ sur la variable X, on calcule les moyennes empiriques

$$n = \sum_{j=1}^{p} n_j$$

$$\bar{x} \ et \ \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$$

Si les n_j sont constants d 'un groupe à l 'autre, on parle de plans (blocs) équilibrés (balancés)

3) Tableaux de données

Deux types de tableaux sont disponibles, les tableaux
a) adaptés pour la compréhension du problème et les calculs « à la main »
b) que l 'on retrouve sur la plupart des logiciels de statistique

a)

Etudier la puissance des véhicules selon le type de carburant utilisé

Facteur, qui prend deux modalités {essence,diesel}

« Puissance » <=> Variable d 'intérêt

essence	diesel
111	64
111	72
154	123
102	123
115	123
110	
110	
110	
140	

pour chaque modalité du facteur, on dispose des observations de la variable d 'intérêt (9 voitures à essence, 5 voitures diesel)

b)

On dispose de la liste des observations, à chaque ligne (observations) on observe la valeur prise de la variable d 'intérêt et la valeur prise par le facteur

ⓘ Cette représentation a l 'avantage de s 'adapter très facilement au cas ou on a deux ou plus de facteurs

puissance	carburant
111	essence
111	essence
154	essence
102	essence
115	essence
110	essence
110	essence
110	essence
140	essence
64	diesel
72	diesel
123	diesel
123	diesel
123	diesel

4) Test d 'hypothèse, Indicateurs statistiques et représentations graphiques

☞ *L 'ANOVA consiste à construire le test d 'hypothèse*

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = \mu \\ H_1 : \exists j, \mu_j \neq \mu \end{cases}$$

La moyenne de la variable dépendante est la même quelque soit les groupes définis par le facteur, il est égal à la moyenne global (en filigrane, le facteur n 'a aucune influence sur la variable dépendante)

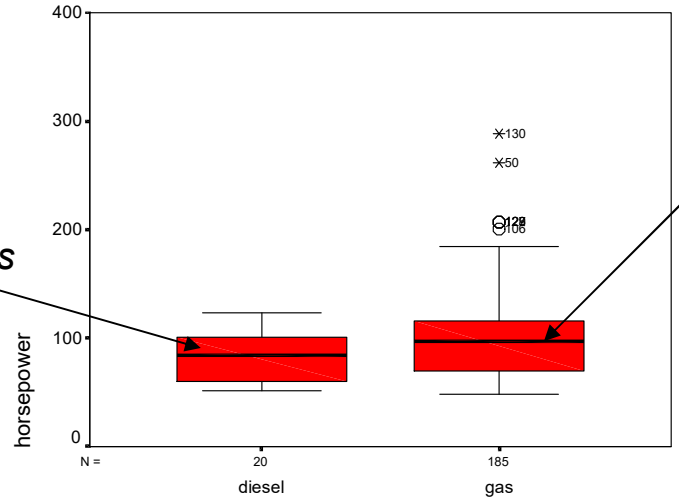
☞ *En utilisant les informations suivantes (mesurés sur l 'échantillon)*

Moyenne conditionnelle (pour chaque facteur) $\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$

Moyenne globale (tous facteurs confondus) $\bar{x} = \frac{1}{n} \sum_{j=1}^{p} n_j \times \bar{x}_j$

☞ *des représentations graphiques peuvent aider à appréhender la solution (séries de « boxplot »)*

Médiane « horsepower » pour les véhicules « diesel »



Médiane « horsepower » pour les véhicules « essence»

5) ANOVA à un facteur fixe (Modèle I)

Les écarts à la moyenne peuvent s'écrire de la manière suivante :

$$x_{ij} - \bar{x} = (\bar{x}_j - \bar{x}) + (x_{ij} - \bar{x}_j)$$

Ecart à la moyenne globale

Ecart entre les groupes (définis par les facteurs)

Ecart à l 'intérieur des groupes

En passant au carré et en faisant les sommations idoines, on obtient l'équation d 'analyse de variance

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{p}(\bar{x}_j - \bar{x})^2 + \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$$

SCT : somme des carrés totaux
Exprime la variabilité totale des observations

SCR : somme des carrés résiduels
Exprime la variabilité résiduelle, à savoir la variation que le facteur n 'arrive pas à expliquer

SCE : somme des carrés expliqués
Exprime la variabilité expliquée, à savoir la variation que le facteur explique

Si le facteur permet de mettre à jour une unité de comportement chez les individus qu'il regroupe (ex: les individus de même sexe ont tous les même taille), la variabilité résiduelle est nulle (dans chaque groupe, les individus sont tous identiques *du point de vue de la variable dépendante*) et la variabilité expliquée est égale à la variabilité totale

Calculs

Carrés moyens

$$CMT = \frac{SCT}{n-1}$$

$$CME = \frac{SCE}{p-1}$$

$$CMR = \frac{SCR}{n-p}$$

Statistique du test et loi associée sous l 'hypothèse H0

$$F = \frac{CME}{CMR} = \frac{SCE/p-1}{SCR/n-p} \equiv Fischer(p-1, n-p)$$

Loi de Fischer à p-1 et n-p degrés de liberté

=>

Pour décider l 'acceptation ou le rejet de l 'hypothèse nulle, il reste à comparer la p-value avec le risque de première espèce que l 'on s 'est choisi

☞ *p-value < α alors rejeter H0, le facteur a bien une influence sur la variable dépendante*

Tableau d 'analyse de variance

(tableau récapitulatif proposé par les logiciels)

Source de variation	Degrés de liberté	Somme des carrés	Carrés moyens	F	p-value
Expliqués	p-1	SCE	CME	CME/CMR	
Résidus	n-p	SCR	CMR		
Total	n-1	SCT			

5) ANOVA à un facteur aléatoire (Modèle II)

L ’échantillonnage est maintenant à deux degrés :
☞ choix aléatoire de certaines modalités du facteur
☞ échantillonnage aléatoire dans les sous-populations décrites par les modalités sélectionnées

=>

Dans la pratique, les calculs sont les mêmes (ceci est valable uniquement pour l ’ANOVA à un facteur !!!).

En revanche, dans l ’interprétation, il est important de noter que l ’on juge essentiellement l’effet global de la variable indépendante sur la variable dépendante ici, on ne peut pas détailler le rôle de telle ou telle modalité du facteur puisque l’on peut en changer d’une expérience à l’autre.

6) Robustesse de l ’ANOVA

Plusieurs hypothèses ont été avancées pour poser le calcul de l ’ANOVA, qu’en est-il si certains d ’entre eux ne sont pas respectés

☞ *Normalité de la distribution de X (variable dépendante)*

• problème : en toute rigueur, les lois de distribution sont invalidées
• diagnostic : la normalité fait surtout intervenir la symétrie de la distribution, on peut s’en assurer de deux manières
  ⌦ graphiquement : histogramme de fréquences
  ⌦ coefficient d ’asymétrie de Fischer (γ_1)
• solution 1 : on peut aussi faire des changements de variable (passage au logarithme, à la racine carrée…) qui symétrise la distribution (ex : distribution des salaires, très dissymétrique, en log elle devient symétrique)
• solution 2 : en fait, l’ANOVA est robuste dès que les formes de distribution sont similaires dans les sous populations (la comparaison peut être graphique)

☞ *Non-homogéneité de la variance (hétéroscédasticité)*

- problème : les groupes à forte variance « tirent » sur les résultats
- diagnostic : test de Bartlett

$$\begin{cases} H_0 : \forall j, \sigma^2{}_j = \sigma^2 \\ H_1 : \exists j, \sigma^2{}_j \neq \sigma^2 \end{cases}$$

On utilise les estimations

$$\hat{\sigma}^2 = \frac{\sum_{i,j} \left(x_{ij} - \bar{x}\right)^2}{n-p}$$

$$\hat{\sigma}_j{}^2 = \frac{\sum_i \left(x_{ij} - \bar{x}_j\right)}{n_j - 1}$$

La statistique du test s 'écrit

Loi du Chi-2 à n-1 degrés de liberté

$$\omega = \frac{(n-p)\ln\hat{\sigma}^2 - \sum_j (n_j - 1)\ln\hat{\sigma}_j{}^2}{1 + \frac{1}{3(p-1)}\left(\sum_j \frac{1}{n_j - 1} - \frac{1}{n-p}\right)} \equiv \chi^2(n-1)$$

Décision de rejet de H_0 au risque α : $\omega \geq \chi_\alpha{}^2(n-1)$ $(\textit{équivalent à } p-value < \alpha)$

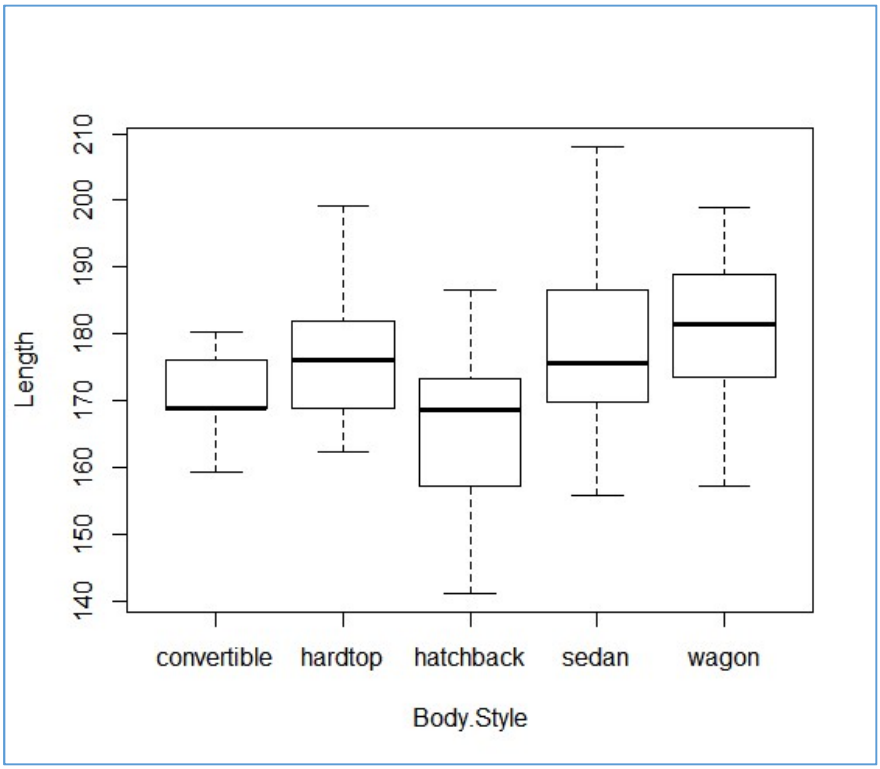
- solution : travailler sur des plans équilibrés (même effectif dans chaque sous-échantillon) atténue l'effet néfaste de l'hétérogénéité des variances

<u>Remarque :</u> Le test de Bartlett n'est pas très robuste par rapport à la non normalité. Mieux vaut utiliser le test de Levene ou de Brown-Forsythe

7.A) Exemple : longueur des voitures en fonction de leur style

Variable dépendante

Facteur



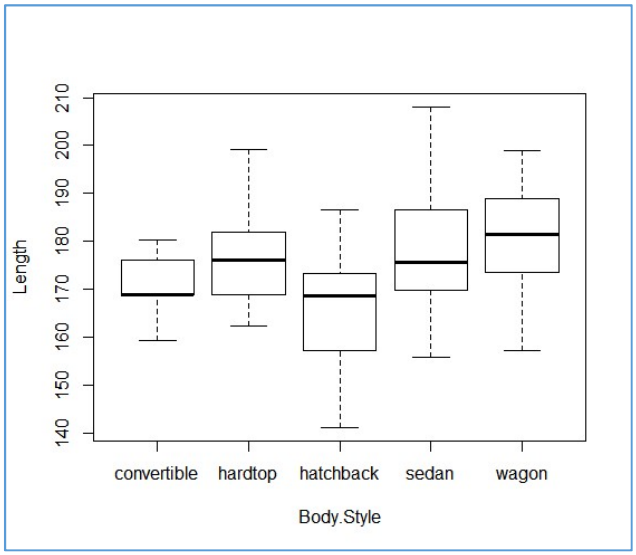
ANOVA

LENGTH

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6555.430	4	1638.857	13.381	.000
Within Groups	24495.143	200	122.476		
Total	31050.572	204			

La longueur des autos diffèrent bien selon leurs styles (au risque 5%)

7.B) ANOVA sous R




```
#charger les données
setwd(''... votre dossier ...'')
library(xlsx)
autos.1 <- read.xlsx("autos_anova.xlsx",header=T,sheetIndex=1)
print(summary(autos.1))

#boxplot conditionnel
boxplot(length ~ body.style, data = autos.1,cex=0.75,ylab="Length",xlab="Body.Style")

#anova
fit <- aov(length ~ body.style, data = autos.1)
print(summary(fit))
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
body.style    4   6555  1638.9   13.38 1.11e-09 ***
Residuals   200  24495   122.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SCE

SCR

CME

CMR

F

$\alpha' = 1.11 \times 10^{-9} < \alpha = 0.05$

COMPARAISON MULTIPLE DE MOYENNES

Position du problème :

L 'ANOVA met en évidence une influence d 'un facteur sur une variable d 'intérêt en utilisant les moyennes, il peut être intéressant de spécifier nommément sur quelles groupes porte ces différences

Ex: il y a une influence des fertilisants sur les rendements, quel est le meilleur fertilisant ?

1) Comparaison deux à deux des moyennes

1.A - Correction de Bonferroni

Pour le test d 'hypothèses suivant

$$\begin{cases} H_0 : \mu_j = \mu_{j'} \\ H_1 : \mu_j \neq \mu_{j'} \end{cases}$$

On utilisera la quantité

$$t_{jj'} = \frac{\bar{x}_j - \bar{x}_{j'}}{\sqrt{\frac{SCR_j + SCR_{j'}}{n_j + n_{j'} + 2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}} \equiv Student(n-p)$$

Qui fournit la p-value α' à comparer avec le risque de première espèce α

Problème

Dans les comparaisons deux à deux, on a p*(p-1)/2 tests à faire. Plus on multipliera les tests, plus on augmentera nos chances de conclure à tort

=> *Il est impératif que le risque global de nos comparaisons deux à deux soit équivalent au risque α pris pour l 'ANOVA*

Inégalité de Bonferroni

Nombre de tests effectués

$$\alpha_{Total} \leq \sum_{t=1}^{T} \alpha_{individuel}$$

=> *On choisit alors comme risque de première espèce pour les risques individuels (test de comparaison de deux moyennes)*

Risque consenti dans l 'ANOVA

$$\alpha_{individuel} = \frac{\alpha}{p(p-1)/2}$$

Equivalent au nombre de tests effectivement réalisés

En toute rigueur, on devrait comparer notre p-value avec cette valeur

Les logiciels de statistique fournissent directement une p-value corrigée que l 'on peut comparer avec le risque α de l 'ANOVA

p-value corrigée (bornée à 1)

$$\tilde{\alpha} = \alpha' \times \frac{p(p-1)}{2}$$

p-value fournie classiquement dans un test de comparaison de moyenne

1.B - Inégalité de Sidak

$$\alpha_{Total} \leq 1 - \prod_{t=1}^{T} \left(1 - \alpha_{individuel}\right)$$

=> *On choisit alors comme risque de première espèce pour les risques individuels (test de comparaison de deux moyennes)*

$$\alpha_{individuel} = 1 - (1 - \alpha)^{\frac{2}{p(p-1)}}$$

1.C – Traitements sous R

```
#comparaison des facteurs deux à deux
pairwise.t.test(autos.1$length,autos.1$body.style,
p.adjust="none",pool.sd=T)
#NT nombre de tests à effectuer
p = nlevels(autos$body.style)
NT = p*(p-1)/2
#alpha corrigé Bonferroni
print(0.05/NT)
#alpha corrigé Sidak
print(1-(1-0.05)^(1/NT))
```

p = 5

α_{anova} = 0.05

$\alpha_{bonferonni}$ = 0.005

α_{sidak} = 0.005116

```
        Pairwise comparisons using t tests with pooled SD

data:  autos$length and autos$body.style

          convertible hardtop hatchback sedan
hardtop   0.274       -       -         -
hatchback 0.422       0.013   -         -
sedan     0.124       0.874   1.7e-09   -
wagon     0.031       0.333   4.1e-08   0.136

P value adjustment method: none
```

Sont les seules différences considérées comme significatives.

Pas celles-ci !

2) Comparaison par rapport à un groupe témoin (test de Dunnett)

Dans l 'ANOVA, il existe souvent un groupe de référence contre lequel on veut se comparer

ex: différents médicaments face à un placebo

Il y a donc (p-1) tests à faire contre ce groupe témoin

☞ Dunnett utilise l 'hypothèse selon laquelle les variance sont homogènes dans les groupes, l 'estimateur de la variance va donc utiliser toutes les données disponibles

$$\hat{\sigma}^2 = \frac{\sum_{i,j} \left(x_{ij} - \bar{x}\right)^2}{n-p}$$

La statistique de Dunnet pour une comparaison entre un groupe (correspondant à une modalité du facteur) avec un groupe témoin (une modalité témoin du facteur) s 'écrit

$$d_{jk} = \frac{\bar{x}_j - \bar{x}_k}{\sqrt{\frac{SCR}{n-p}\left(\frac{1}{n_j} + \frac{1}{n_k}\right)}}$$

☞ Dunnet a tabulé les valeurs de d_{jk} (les logiciels font automatiquement le bon calcul et fournissent une p-value comparable avec α de l 'ANOVA), mais dans les faits elle est proche d 'un test de Student avec une correction de Bonferroni ou Sidak, mais où le facteur de correction s 'écrit

$$\alpha_{individuel} = \frac{\alpha}{p-1}$$

ANALYSE DE VARIANCE
à DEUX FACTEURS (A * B)

Position du problème :

On veut mesurer maintenant le rôle conjoint de deux facteurs A et B sur la variable dépendante

3 effets sont à mesurer
- ☞ effet de A
- ☞ effet de B
- ☞ interaction entre A et B

→ Effets principaux

Exemples :
- type de fertilisants et mode d'épandage => croissance des arbres
- type de fumeur (actif, modéré, non-fumeur) et sexe => durée de vie

1) Description des données

☞ P est la population

☞ X est la variable d 'intérêt de moyenne globale μ

☞ on étudie le rôle de deux facteurs A et B

B est le second facteur (avec q modalités : $B_1, B_2, \ldots, B_q$)

A est le premier facteur (avec p modalités : $A_1, A_2, \ldots, A_p$)

☞ A et B définissent p*q sous - population P_{ij}

☞ on note $P_{i.}$ (resp. $P_{.j}$) les individus corresp. à $A=A_i$ ($B=B_j$)

μ_{ij}

$\mu_{i.}$

$\mu_{.j}$

☞ Dans chaque sous-population Pij, on extrait un échantillon Eij (tirage indépendant et équiprobable)

Dans tout ce qui suit, on considère que le plan d 'expériences est équilibré, card(Eij)=n

☞ les résultats sont plus clairs et plus facilement explicités

=> X_{ijr} est l 'observation numéro r dans l 'échantillon Eij ceci nour permet de définir les moyennes croisées et marginales

$$\bar{x}_{ij} = \frac{1}{n}\sum_{r=1}^{n} x_{ijr}$$

$$\bar{x}_{i.} = \frac{1}{q}\sum_{j=1}^{q} \bar{x}_{ij}$$

$$\bar{x}_{.j} = \frac{1}{p}\sum_{i=1}^{p} \bar{x}_{ij}$$

2) Tableau de données

Facile à lire mais encombrant
Pratique pour les calculs manuels

aspiration / carburant	atmo	turbo
diesel	52 56 58	68 65 67
essence	48 49 67	102 145 130

$$x_{2,2,2} = puissance_{essence,turbo,2}$$

Tables de données usuelles
Utilisées par les logiciels

carburant	aspiration	puissance
essence	atmo	110
essence	atmo	69
essence	atmo	112
essence	turbo	142
essence	atmo	152
essence	atmo	94
diesel	turbo	106

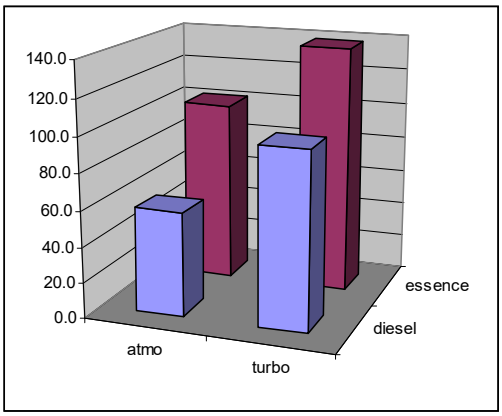
Représentation des moyennes

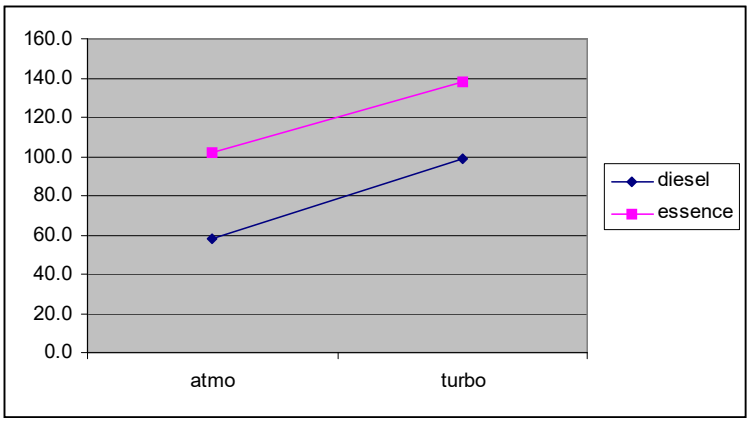
Tableau croisé dynamique
d 'Excel par exemple

$\bar{x}_{1,1}$ $\bar{x}_{1.}$

Moyenne puissance	aspiration		
carburant	atmo	turbo	Total
diesel	58.1	98.6	84.5
essence	101.6	138.4	106.4
Total	99.8	124.4	104.3

☞ *Graphiques associés*





Celui-ci est plus intéressant car il permet de distinguer les interactions (lorsque les lignes se croisent)

3) Hypothèses statistiques

Ce sont les mêmes que pour l 'ANOVA à 1 facteur (normalité de la variable dépendante, effets additifs, variance homogène dans les groupes)

☞ encore une fois, en passant par un plan équilibré, on améliore la robustesse du test...

4.A) ANOVA à deux facteurs

Hypothèses soumises au test (il y en a 3 en tout)

$$\begin{cases} H_0 : \mu_{i.} = \mu, \forall i \\ H_0 : \mu_{.j} = \mu, \forall j \\ H_0 : \mu_{ij} = \mu, \forall i, j \end{cases}$$

Décomposition de la moyenne

$$x_{ijr} - \overline{x} = \left(\overline{x}_{i.} - \overline{x}\right) + \left(\overline{x}_{.j} - \overline{x}\right) + \left(\overline{x}_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}\right) + \left(x_{ijr} - \overline{x}_{ij}\right)$$

Effets des facteurs principaux

Effet de l 'interaction

Erreur résiduelle (erreur expérimentale)

A partir de laquelle on extrait l 'équation d 'ANOVA

$$SCT = SCE_A + SCE_B + SCE_{AB} + SCR$$

Variabilité totale

Variabilité expliquée

Variabilité résiduelle

Carrés moyens

$$CMT = SCT / pqn-1$$

$$CME_A = SCE_A / p-1$$

$$CME_B = SCE_B / q-1$$

$$CME_{AB} = SCE_{AB} / (p-1)(q-1)$$

$$CMR = SCR / pq(n-1)$$

Quels sont les rapports de carrés moyens à prendre pour mettre à jour les effets (principaux et interactions)

=> **La réponse dépend du type de facteur considéré** (fixe ou aléatoire)

Pour mettre à jour les effets principaux

Pour mettre à jour l 'interaction

Modèles	F_A	F_B	F_{AB}
(I) A et B fixes	$\frac{CME_A}{CMR}$	$\frac{CME_B}{CMR}$	$\frac{CME_{AB}}{CMR}$
(III) A fixe et B aléat.	$\frac{CME_A}{CME_{AB}}$	$\frac{CME_B}{CMR}$	$\frac{CME_{AB}}{CMR}$
(II) A et B aléat.	$\frac{CME_A}{CME_{AB}}$	$\frac{CME_B}{CME_{AB}}$	$\frac{CME_{AB}}{CMR}$

☞ *Ces quantités suivent une loi de Fischer, les degrés de libertés sont lus dans les dénominateurs des carrés moyens associés*

4.B) ANOVA à deux facteurs sous R

Puissance des véhicules en fonction du type de carburant (fuel-type) et le mode d'alimentation (aspiration)
(tests à 5%)

```
diesel     gas
84.450 106.396
```

```
     std     turbo
99.81105 124.43243
```

```
#Données pour ANOVA à 2 facteurs
autos.2 <- read.xlsx("autos_anova.xlsx",header=T,sheetIndex=2)
print(summary(autos.2))

#moyennes conditionnelles
#vs. fuel.type
print(tapply(autos.2$horsepower,list(autos.2$fuel.type),mean))

#vs. aspiration
print(tapply(autos.2$horsepower,list(autos.2$aspiration),mean))

#vs. fuel.type * aspiration
print(tapply(autos.2$horsepower,list(autos.2$fuel.type,autos.2$aspirati
on),mean))

#ANOVA à 2 facteurs
fit2 <- aov(horsepower ~ fuel.type + aspiration + fuel.type*aspiration,
data = autos.2)
print(summary(fit2))
```

```
            std     turbo
diesel  58.14286  98.61538
gas    101.62271 138.41667
```

ANOVA à 2 facteurs

```
                     Df Sum Sq Mean Sq F value   Pr(>F)
fuel.type             1   8693    8693   6.373   0.0124 *
aspiration            1  35678   35678  26.156 7.32e-07 ***
fuel.type:aspiration  1     51      51   0.037   0.8475
Residuals           201 274179    1364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

« Fuel-type » et « aspiration » influent sur la puissance, pas leur interaction.

5) Cas particulier de l ’ANOVA à deux facteurs :
ANOVA à 1 facteur avec mesures répétées

Exemple : évaluer l ’efficacité de 3 méthodes anti-tabac (A1,A2,A3)

☞ de manière classique, on construirait 3 échantillons (E1,E2,E3) sur lesquelles on applique respectivement A1,A2 et A3

Cette procédure est complètement équivalente avec le schéma de tirage d ’urne de la randomisation

Ne peut-on pas imaginer un mode d ’expérimentation plus judicieux (qui nous mettrait par exemple à l ’abri des facteurs de blocs)

On peut utiliser le procédé suivant
☞ appliquer la méthode A1 à l ’individu i
☞ une semaine plus tard, appliquer A2 au même individu i
☞ une semaine plus tard, appliquer A3 au même individu i

=> **C ’est un plan d ’expériences à mesures répétées**
☞ la variable d ’intérêt est mesurée plusieurs fois sur le même individu

Tableau de données

facteur / individu	A1		Aj		Ap
1					
i			Xij		
n					

Facteur 1 = Facteur individu (forcément aléatoire)
Facteur 2 = Facteur A (fixe ou aléatoire)

L 'ensemble des modalités est en fait la population ici

Particularités

☞ on utilise moins d 'individus que dans un plan complètement randomisé

☞ la « sensibilité » des résultats est meilleure dans le sens où on détecte mieux les effets

Dans le jugement de l 'effet du facteur A, on enlève l 'incertitude liée au fait que l 'on utilise des individus différents dans les groupes - les facteurs de bloc sont annihilés

☞ danger : ce plan peut être impraticable s'il y a des phénomènes d'accoutumance (test des posologies en médecine), d'apprentissage (évaluation des tests de QI en psychologie) ou d'accumulation (fertilisants successifs sur le même arbre)

Décomposition de la moyenne

$$x_{ij} - \bar{x} = \left(\bar{x}_{i.} - \bar{x}\right) + \left(\bar{x}_{.j} - \bar{x}\right) + \left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}\right)$$

Facteur individu — *Facteur A* — *Joue le rôle de facteur résiduel ici*

=> L 'interaction entre le facteur et l'individu n 'existe pas

Carrés moyens et F calculés

$$CMT = SCT / np-1 = \sum_i \sum_j (x_{ij} - \bar{x})^2 \Big/ np-1$$

$$CME_I = SCE_I / n-1 = n \sum_j (\bar{x}_{i.} - \bar{x})^2 \Big/ n-1$$

$$CME_A = SCE_A / p-1 = p \sum_i (\bar{x}_{.j} - \bar{x})^2 \Big/ p-1$$

$$CME_{IA} = SCE_{IA} / (p-1)(n-1) = SCT - SCE_I - SCE_A / (p-1)(n-1)$$

=> L 'effet du facteur A est alors transcrit par

$$F_A = \frac{CME_A}{CME_{IA}} \equiv Fischer\left(p-1, (p-1)(n-1)\right)$$

On utilise tout simplement la p-value pour prendre une décision

Bibliographie

Abdi H., « Introduction au traitement statistique des données expérimentales », PUG, 1987.

Dagnelie P., « Statistique théorique et appliquée – Tome 2. Inférence statistique à une et à deux dimensions », De Boeck, 2011.

Guenther W., « Analysis of variance », Prentice-Hall, 1964.

Scherrer B., « Biostatistique », Vol. 1, 2ème Edition, Gaëtan Morin Editeur, 2007.

Sheskin D.J., « Handbook of Parametric and Nonparametric Statistical Procedures », Chapman & Hall, 2007.

… et les très nombreux supports de cours sur Internet.
Ex. Arnold S., « STAT 502 : Analysis of Variance and Design of Experiments », PennState – Eberly College of Science, 2008 ;
https://onlinecourses.science.psu.edu/stat502/