

Module 9 - Analyse de la variance à un facteur

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Principes d'expérimentation

Principes d'expérimentation

Pour qu'une expérience mène à une découverte réelle,

1. l'expérience doit être bien planifiée,
2. les résultats doivent être adéquatement analysés.

Pourquoi planifier une expérience ?

- se protéger contre les sources de biais potentielles,
- contrôler la variabilité due à certains facteurs.

Une expérience sur la durabilité des batteries

But : Déterminer quel type de batterie (Lithium-ion, Lithium-fer-phosphate ou Nickel-hydrure métallique) conserve le mieux sa capacité après 500 cycles de recharge.

Plusieurs variables influentes :

- Le chargeur utilisé peut légèrement varier d'une station à l'autre.
- La température ambiante dans la salle d'essai n'est pas parfaitement stable.
- Les capteurs de mesure de capacité ont de légers biais entre les bancs de test.

Comment concevoir l'expérience pour isoler l'effet du type de batterie malgré les variations causées par les chargeurs, les conditions ambiantes et les instruments de mesure ?

Principes d'expérimentation

Trois principes essentiels :

- identification des sources de variabilité (pour les contrôler).
- répétition (pour estimer la variabilité)
- randomisation (pour éliminer les biais)

L'analyse de la variance

L'analyse de la variance compare l'effet des sources de variabilité. Utile pour :

- déterminer si les différences observées sont bien dues aux facteurs d'intérêt et non à la chance
- construire un modèle de prévision
- déterminer la combinaison optimale des facteurs

On dit **anova**, pour ANalysis Of VAriance.

Plan d'expérience à un facteur

But d'une expérience à un facteur

Comparer 2 moyennes :

2 populations normales
variances égales
 $H_0 : \mu_1 = \mu_2$

} test t à 2 échantillons indépendants

Comparer plusieurs moyennes :

a populations normales
variances égales

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

$H_1 : \mu_k \neq \mu_l$ pour au moins
une paire (k, l)

} analyse de la variance
à 1 facteur
(a échantillons indépendants)

On veut tester l'effet d'un facteur sur la moyenne d'une variable.

Vocabulaire

- **Variable réponse** : La variable pour laquelle nous désirons tester l'égalité des moyennes. Appelée aussi *variable dépendante*, notée Y .
- **Facteur** : Ce qui différencie les populations ou les traitements. Chaque traitement ou population correspond à un *niveau* du facteur, ou à une *modalité* du facteur.
- **Unités expérimentales** : Les individus ou objets sur lesquels le traitement est appliqué.

En anova à un facteur, **les échantillons doivent être indépendants** (une seule mesure par unité expérimentale).

QUIZ!

Dans l'expérience sur les batteries, identifiez les quantités suivantes :

- La variable réponse
- Le facteur et le nombre de niveaux
- L'unité expérimentale

Devis de l'expérience à un facteur

Deux scénarios d'échantillonnage possibles :

1. Choisir $n \times a$ unités expérimentales. Déterminer aléatoirement n unités qui recevront le traitement 1, n unités qui recevront le traitement 2, ..., n unités qui recevront le traitement a ;
2. échantillonner aléatoirement n unités dans la population 1, échantillonner aléatoirement n unités dans la population 2, ..., échantillonner aléatoirement n unités dans la population a .

Dans les deux cas, on obtient a échantillons indépendants de taille n .

⇒ on obtient les observations

$$Y_{ij}, \quad \text{où} \quad \begin{matrix} i = 1, \dots, a, \\ j = 1, \dots, n. \end{matrix}$$

Notation pour les sommes et moyennes

Le symbole $\bullet$ signifie qu'une moyenne ou une somme est calculée pour **toutes les valeurs de l'indice que $\bullet$ remplace**.

Supposons y_{ij} pour $i = 1, \dots, 6$, $j = 1, \dots, 8$. On aurait, par exemple,

$$y_{\bullet j} = \sum_{i=1}^6 y_{ij}$$

$$\bar{y}_{i\bullet} = \frac{1}{8} \sum_{j=1}^8 y_{ij}$$

$$y_{\bullet\bullet} = \sum_{i=1}^6 \sum_{j=1}^8 y_{ij}$$

$$\bar{y}_{\bullet\bullet} = \frac{1}{6 \times 8} \sum_{i=1}^6 \sum_{j=1}^8 y_{ij}$$

Exemple 1 : la force du textile

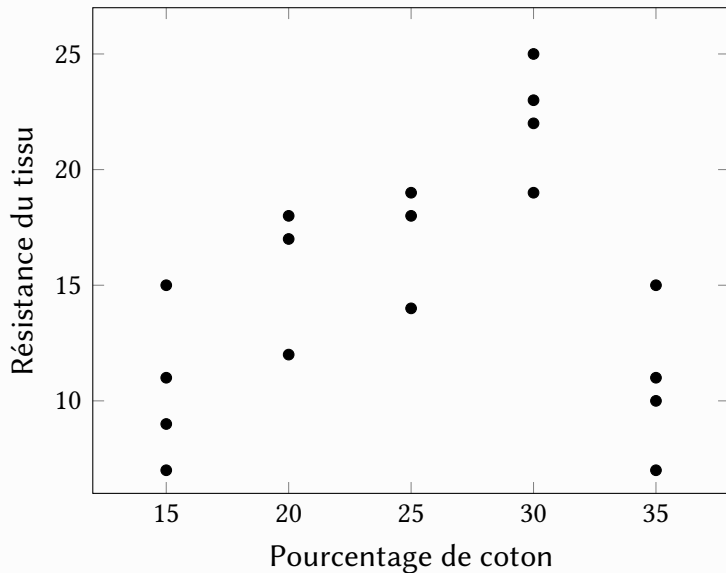
Un ingénieur veut savoir si la force de résistance d'une fibre textile est différente selon la proportion de coton dans la fibre.

Son expérience : mesurer la résistance pour 25 unités de tissu fabriquées avec différentes proportions de coton.

Exemple 1 : les données (en *psi*, lbs/pouce carré)

Pourcentage de coton i	Observations j					Somme	Moy.	É.-type
	1	2	3	4	5	$y_{i\cdot}$	$\bar{y}_{i\cdot}$	s_i
15	7	7	15	11	9	49	9,8	3,35
20	12	17	12	18	18	77	15,4	3,13
25	14	18	18	19	19	88	17,6	2,07
30	19	25	22	19	23	108	21,6	2,61
35	7	10	11	15	11	54	10,8	2,86
						$\bar{y}_{..} = 15,04$		

Exemple 1 : représentation graphique



Modèle et interprétation

$$\begin{aligned} Y_{ij} &= \mu_i + E_{ij}, \quad \text{où } i = 1, \dots, a, \quad j = 1, \dots, n, \\ &= \mu + \tau_i + E_{ij}, \end{aligned}$$

où

Y_{ij} : valeur de la variable réponse pour l'observation j de l'échantillon/du traitement i ;

μ_i : valeur moyenne de Y dans la population i ;

μ : moyenne globale de Y (toutes pop. confondues);

τ_i : différence entre la moyenne du traitement i et la moyenne globale ($\tau_i = \mu_i - \mu$). On a la contrainte $\sum_{i=1}^a \tau_i = 0$;

E_{ij} : terme d'erreur pour la j^e observation du traitement i ;

Illustration des paramètres

Postulats du modèle

Le modèle se base sur les postulats suivants :

(P1) Les a échantillons sont indépendants.

(P2) La variance des Y_{ij} est homogène entre les traitements.

(P3) Les Y_{ij} sont des variables indépendantes de loi normale.

Nous pouvons résumer ces postulats ainsi :

$$Y_{ij} = \mu_i + E_{ij}, \quad \text{où } E_{ij} \text{ sont des v.a. i.i.d. } N(0, \sigma^2).$$

Ces postulats devront être vérifiés avant de tirer des conclusions basées sur des tests d'hypothèses.

Estimation et test d'hypothèses

Estimation ponctuelle des paramètres

- Paramètres inconnus du modèle : $\mu, \tau_1, \dots, \tau_a$ et σ^2 .
- L'estimateur de μ est $\hat{\mu} = \bar{y}_{..}$, la moyenne de toutes les observations.
- On estime μ_i par la moyenne des observations du traitement i , $\hat{\mu}_i = \bar{y}_{i.}$.
- Puisque $\tau_i = \mu_i - \mu$, on a que $\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$.
- La variance étant considérée égale dans tous les traitements, on estime σ^2 en combinant les variances des a échantillons : $\hat{\sigma}^2 = \frac{S_1^2 + \dots + S_a^2}{a} = MS_E$.

Exemple 1

Quelles sont les estimations des paramètres pour l'exemple 1? [Illustration](#)

Les hypothèses testées

La question la plus importante dans ce type d'étude est

Le facteur a-t-il un effet sur la moyenne de la variable réponse?

Si le facteur n'a aucun effet, la variable réponse aura la même loi pour tous les traitements,

$$H_0 : \mu_1 = \dots = \mu_a = \mu \quad \text{c'est-à-dire} \quad H_0 : \tau_1 = \dots = \tau_a = 0.$$

Si le facteur a un effet, **au moins deux moyennes diffèrent**

$$H_1 : \exists(k, l) \text{ tel que } \mu_k \neq \mu_l \quad \text{c'est-à-dire} \quad H_1 : \text{au moins un des } \tau_i \neq 0.$$

Loi de Y_{ij} sous H_0 et sous H_1

L'analyse de la variance

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

Tous les Y_{ij} ne sont pas égaux à μ , parce que

- (i) le traitement leur ajoute potentiellement une déviation τ_i et
- (ii) la fluctuation entre les unités expérimentales (l'erreur aléatoire) leur ajoute une erreur E_{ij} .

Ceci suggère que l'on peut diviser la variabilité observée dans les Y_{ij} selon deux sources : l'effet du facteur et l'erreur aléatoire.

Décomposition de la variabilité

On peut montrer que

$$\underbrace{\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2}_{SS_T} = \underbrace{\sum_{i=1}^a \sum_{j=1}^n (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{SS_{trt}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2}_{SS_E}$$

Variabilité totale = Variabilité due aux traitements + Variabilité résiduelle.

Si H_0 est vraie $\Rightarrow$ Les τ_i sont nuls.

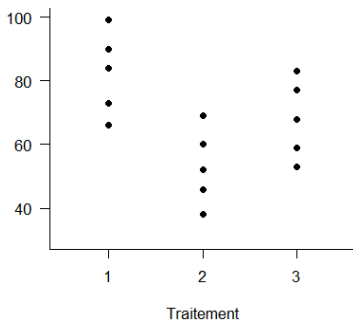
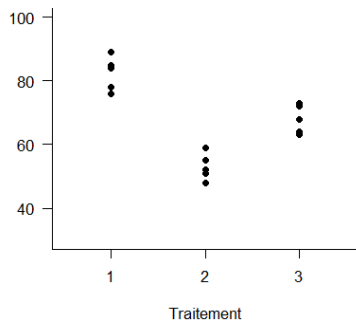
$\Rightarrow$ SS_{trt} devrait prendre une faible valeur.

$\Rightarrow$ une faible proportion de la variabilité totale entre les observations sera expliquée par le facteur et une grande proportion sera expliquée par l'erreur.

Illustration

QUIZ!

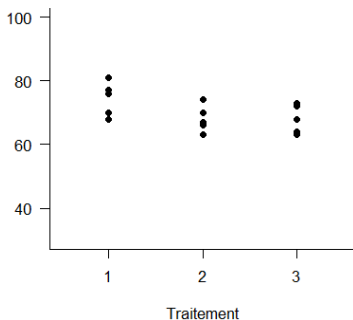
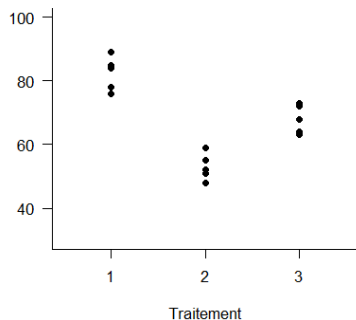
2 expériences ont obtenu les mêmes moyennes échantillonnales :



Quelle expérience a donné lieu au plus grand SS_{trt} ?
au plus grand SS_E ? au plus grand SS_T ?

QUIZ!

2 expériences ont obtenu les mêmes variances échantillonnales :



Quelle expérience a donné lieu au plus grand SS_{trt} ?
au plus grand SS_E ? au plus grand SS_T ?

Statistique de test

Pour déterminer si SS_{trt} est suffisamment élevée par rapport à SS_T pour rejeter H_0 , on utilise l'approche suivante.

On peut montrer que

- $\frac{SS_{trt}}{\sigma^2} \sim \chi^2_{a-1}$ si H_0 est vraie
- $\frac{SS_E}{\sigma^2} \sim \chi^2_{an-a}$
- $\frac{SS_{trt}}{\sigma^2}$ et $\frac{SS_E}{\sigma^2}$ sont indépendantes

On peut en déduire que si H_0 est vraie,

$$F_0 = \frac{SS_{trt}/(a-1)}{SS_E/(an-a)} \sim F_{a-1, an-a}.$$

Tableau d'analyse de la variance

Les logiciels résument habituellement cette décomposition de la variabilité dans un **tableau d'analyse de la variance** (table ANOVA) :

Source de variabilité	Somme des carrés	Degrés de liberté	Carré moyen	F_0	Valeur p
Traitement	SS_{trt}	$a - 1$	$MS_{trt} = \frac{SS_{trt}}{a-1}$	$F_0 = \frac{MS_{trt}}{MS_E}$	$P(F_{a-1, an-a} > F_0)$
Erreur	SS_E	$an - a$	$MS_E = \frac{SS_E}{an-a}$		
Totale	SS_T	$an - 1$			

Espérance des carrés moyens

On peut montrer que

- $E(MS_{trt}) = \sigma^2 + \frac{n \sum_{i=1}^a \tau_i^2}{a - 1}$
- $E(MS_E) = \sigma^2 \rightarrow$ on se souvient que MS_E est notre estimateur de σ^2
- Si H_0 est vraie,
- Si H_1 est vraie,

Exemple 2 : inférence sur les résistances moyennes

La force moyenne de résistance de la fibre textile est-elle différente selon la proportion de coton, au seuil de 5% ?

Pourcentage de coton i	Observations j					Somme	Moy.	É.-type
	1	2	3	4	5	$y_{i\cdot}$	$\bar{y}_{i\cdot}$	s_i
15	7	7	15	11	9	49	9,8	3,35
20	12	17	12	18	18	77	15,4	3,13
25	14	18	18	19	19	88	17,6	2,07
30	19	25	22	19	23	108	21,6	2,61
35	7	10	11	15	11	54	10,8	2,86
						$\bar{y}_{..} = 15,04$		

Source : *Design and Analysis of Experiments*, D.C. Montgomery (2001).

Exemple 2 : pré-calculs pour la table d'anova

Exemple 2 : table d'anova et seuil observé

Source de variabilité	Somme des carrés	Degrés de liberté	Carré moyen	F_0	Valeur p
Traitement					
Erreur					
Totale					

Exemple 2 : Décision et conclusion

Décision : ☐ On rejette H_0 ☐ On ne rejette pas H_0

Conclusion :

Le pourcentage de coton $\left\{ \begin{array}{l} \text{a un effet significatif} \\ \text{n'a pas d'effet significatif} \end{array} \right\}$
sur la force moyenne de la fibre textile, au seuil de 5%.

QUIZ!

Lorsque H_0 est rejetée, lesquelles des conclusions suivantes sont correctes ?

- ☐ Toutes les moyennes sont égales.
- ☐ Toutes les moyennes sont différentes.
- ☐ Au moins deux moyennes sont différentes.
- ☐ La variable réponse a un effet sur le facteur.
- ☐ Le facteur a un effet sur la variable réponse.

Validation des postulats

Postulats à valider

L'inférence (test F) est basée sur trois postulats :

(P1) l'indépendance des observations ;

(P2) l'homogénéité des variances d'un traitement à l'autre ;

(P3) la normalité des E_{ij} .

(P2) et (P3) peuvent être validés graphiquement.

Les erreurs E_{ij} sont inconnues. On les estime avec les *résidus*.

Résidus

On estime les $E_{ij} = Y_{ij} - \mu - \tau_i$ par les **résidus** e_{ij} :

$$\begin{aligned} e_{ij} &= Y_{ij} - \hat{\mu} - \hat{\tau}_i \\ &= Y_{ij} - \bar{Y}_{..} - (\bar{Y}_{i.} - \bar{Y}_{..}) \\ &= Y_{ij} - \bar{Y}_{i.} \end{aligned}$$

Propriétés des résidus

Pour valider les deux postulats, on peut regarder les graphiques suivants :

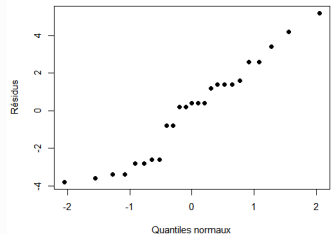
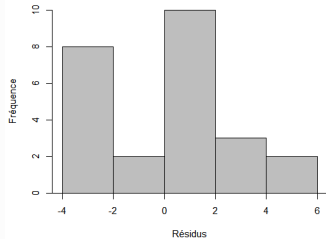
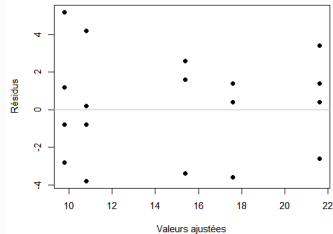
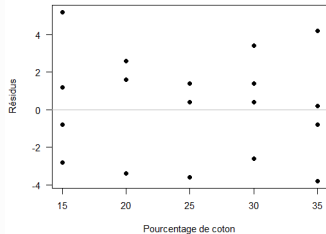
(P2) Homogénéité des variances :

- juxtaposer les a diagrammes en boîte des résidus, ou tracer les résidus pour chaque traitement
(la dispersion devrait être similaire);
- graphique des e_{ij} en fonction des valeurs ajustées $\bar{y}_i$.
(la distribution des e_{ij} ne devrait pas changer selon les $\bar{y}_i$).

(P3) Normalité des erreurs :

- histogramme des an résidus
(on devrait observer une forme de cloche symétrique);
- diagramme quantile-quantile des an résidus.
(on devrait voir une ligne assez droite).

Exemple 3 : Analyse des résidus



Oui, mais...

Que faire si les postulats ne semblent pas respectés, d'après l'analyse des résidus?

- Tests de normalité (tests d'hypothèses non abordés dans le cours)
- Transformation des Y_{ij}
- Anova non paramétrique (test de Kruskal-Wallis, non abordé dans le cours)

Tests de comparaison des moyennes 2 à 2

Test F significatif : la suite

1. Si le test F d'égalité des moyennes n'est pas significatif (H_0 non rejetée), l'analyse est terminée.
2. Si le test global du tableau ANOVA nous mène à rejeter H_0 , on veut savoir quelles moyennes diffèrent les unes des autres.

Plusieurs procédures permettent de faire les comparaisons de moyennes 2 à 2 (souvent appelées *tests de comparaisons multiples*).

L'approche de la *plus petite différence significative* (PPDS) de Fisher offre une réponse à cette question.

Approche “PPDS” de Fisher

- On se base sur l'IC($\mu_i - \mu_k$) de niveau $1 - \alpha$.
- Si la différence entre les moyennes est supérieure à la marge d'erreur de cet intervalle (en valeur absolue), on les déclarera significativement différentes au seuil α .

Approche PPDS de Fisher

$$\text{PPDS} = t_{\alpha/2; an-a} \sqrt{\frac{2\text{MS}_E}{n}}.$$

On déclare **significativement différentes au seuil α** toute paire de moyennes qui diffèrent (en val. abs.) par au moins PPDS.

Exemple 4 : calcul de la PPDS

Nous avons conclu que les moyennes des 5 traitements n'étaient pas toutes égales. Mais quelles paires de moyennes sont significativement différentes au seuil de 5% ?

Exemple 4 : comparaisons 2 à 2

$\bar{y}_{1\cdot}$	$\bar{y}_{5\cdot}$	$\bar{y}_{2\cdot}$	$\bar{y}_{3\cdot}$	$\bar{y}_{4\cdot}$
9,8	10,8	15,4	17,6	21,6
15%	35%	20%	25%	30%

Réponses aux exemples

1. $\hat{\mu} = \bar{y}_{..} = 15,04$; $\hat{\tau}_1 = -5,24$; $\hat{\tau}_2 = 0,36$; $\hat{\tau}_3 = 2,56$; $\hat{\tau}_4 = 6,56$; $\hat{\tau}_5 = -4,24$; $\hat{\sigma}^2 = 8,06$

2.

Source de variabilité	Somme des carrés	Degrés de liberté	Carré moyen	F_0	Valeur p
Traitement	475,8	4	118,94	14,76	$9,13 \times 10^{-6}$
Erreur	161,2	20	8,06		
Totale	637,0	24			

Effet significatif du pourcentage de coton.

3. Variances très similaires, normalité questionnable (bimodalité).

4. PPDS = 3,75

$\bar{y}_{1\cdot}$	$\bar{y}_{5\cdot}$	$\bar{y}_{2\cdot}$	$\bar{y}_{3\cdot}$	$\bar{y}_{4\cdot}$
9,8	10,8	15,4	17,6	21,6

Adaptation de la méthode

Comme plusieurs parmi vous le verrez dans un cours de troisième ou quatrième année, le modèle d'analyse de la variance à un facteur peut être adapté/généralisé pour faire face à des situations plus complexes. Par exemple :

- **Contrôler la variabilité** : On peut généraliser le test t pour échantillons appariés au cas où l'on a plus de 2 échantillons : plan d'expérience à blocs aléatoires (voir section 11.2 du manuel du cours).
- **Étudier l'effet de plus d'un facteur** : On peut planifier des expériences servant à étudier l'effet de plusieurs facteurs simultanément et en analyser les résultats à l'aide d'une généralisation du principe de décomposition de la variance (voir sections 11.3-11.4 du manuel du cours).