# Analyse de la variance à un facteur

Auteur : Anne-Laure Fougères

L'objectif général de l'analyse de variance (appelée souvent Anova) est de *comparer des populations à l'aide de tests statistiques*, ou encore de *comparer différents traitements* administrés à des sujets. Elle représente en ce sens une généralisation du test de Student de comparaison de 2 moyennes (sous hypothèse de normalité). Cette méthode est fondée sur la décomposition de la variabilité d'une variable réponse en composantes associées aux différentes sources de variation.

## 1. Introduction

### 1.1. Quelques exemples

**(a)** Etude de la relation entre la "quote de popularité" (ou "statut auprès des pairs", mesuré selon 3 niveaux : populaire/moyen/rejeté) et le sentiment de solitude chez les enfants.
**(b)** Etude de la réussite scolaire, pour des enfants de différentes classes sociales (→ un facteur). Effet de la classe sociale sur la réussite chez les femmes ? ches les hommes ? (→ deux facteurs).
**(c)** Etude comparative de 4 somnifères à l'aide d'une note de satisfaction donnée par le sujet prenant le somnifère.
**(d)** Comparaison de 5 méthodes d'enseignement en terme d'apprentissage.
**(e)** Etude sur le stress au travail : le niveau de responsabilité a-t-il un impact sur l'état de stress ? Un facteur à 3 modalités : 3 catégories de personnels (technicien/cadre responsable d'une petite équipe/cadre responsable d'un département). Une variable dépendante : la mesure de stress subjectif.

### 1.2. Variabilités "intergroupe" et "intragroupe"

Reprenons l'exemple **(c)** précédent. Les degrés de satisfaction exprimés présentent des différences suivant les personnes, du fait de la variabilité des sujets (sensibilité au médicament, effets secondaires, sensation de repos, etc...). On a donc, même avec un seul somnifère, une variabilité, une variation entre les sujets. Question : y a-t-il une variabilité due au type de somnifère plus grande que la variabilité entre sujets pour un même somnifère ? une variabilité trop grande pour être due au hasard ?

On distinguera deux types de variabilité : la *variabilité intra-groupe*, qui est la variabilité à l'intérieur de chaque groupe et la *variabilité inter-groupe*, qui est la variabilité entre les différents groupes.

### 1.3. Pourquoi ne peut-on directement utiliser plusieurs tests de Student ?

Un test de Student compare les moyennes $\mu_1$ et $\mu_2$ de deux échantillons indépendants, gaussiens, de même variance $\sigma^2$ :

$$H_0 : \mu_1 = \mu_2 \text{ contre } H_1 : \mu_1 \neq \mu_2.$$

Supposons que l'on ait à comparer $k \geq 3$ échantillons indépendants, gaussiens, de même variance $\sigma^2$. On cherche donc à tester

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ contre } H_1 : \text{"au moins deux des } \mu_j \text{ sont différents"}.$$

Si on procède à une série de tests de Student, deux à deux : il faut donc faire $N = k(k-1)/2$ t-tests. Supposons que chacun de ces tests soit fait au niveau $\alpha$ fixé, de sorte que, pour tout $i$ et $j$ fixés, on ait :

$$\mathrm{P}(\text{rejeter l'hypothèse " } \mu_i = \mu_j \text{ " à tort}) = \alpha.$$

On a alors : P(accepter l'hypothèse " $\mu_i = \mu_j$ " alors qu'il y a égalité) $= 1 - \alpha$.
Et par suite :

$$\mathrm{P}(\text{accepter les } N \text{ hypothèses " } \mu_i = \mu_j \text{ " alors qu'il y a égalité}) = (1-\alpha)^N,$$

ce qui entraîne que

$$\begin{aligned}\mathrm{P}(\text{rejeter } H_0 \text{ à tort}) &= \mathrm{P}(\text{rejeter au moins une des égalités alors que } H_0 \text{ est vraie})\\ &= 1 - (1-\alpha)^N.\end{aligned}$$

Si $\alpha = 0.05$, on obtient les valeurs suivantes :

| k | N | P(rejeter $H_0$ à tort) |
|---|---|---|
| 2 | 1 | $0.05 = 1 - (1-\alpha)^1$ |
| 3 | 3 | $0.14 = 1 - (1-\alpha)^3$ |
| 4 | 6 | 0.26 |
| 10 | 45 | 0.90 |

Conclusion : on va rejeter $H_0$ à tort beaucoup trop souvent ! Les t-tests deux à deux sont donc à proscrire. L'analyse de variance est la technique correcte pour les comparaisons de plus de deux moyennes dans un cadre gaussien.

## 2. Anova à un facteur : démarche à suivre, sur un exemple

### 2.1 Présentation d'un jeu de données

Comparaison de 5 méthodes d'enseignement.
Objectif : tester si le type de méthode a un impact sur l'apprentissage.
Mesure : notes de 45 étudiants, divisés aléatoirement en 5 groupes de 9.

### 2.2 Description de l'expérience, terminologie

Il s'agit de définir tous les éléments du *plan d'expérience*. Remarquons que les techniques d'analyse de variance s'étant beaucoup développées dans le contexte agronomique et médical, la terminologie y est souvent adaptée.

**A.** Définir l'*unité expérimentale* : c'est l'unité à laquelle est appliqué un traitement. Ici, c'est l'individu qui suit le cours selon une méthode fixée.

**B.** Préciser les différents *traitements* appliqués aux unités expérimentales : ici, on a un seul *facteur* (ou *variable indépendante*), qui est appelé "méthode d'enseignement". Il y a 5 méthodes différentes, on dit que le facteur "méthode d'enseignement" comporte 5 *modalités* (ou encore 5 *traitements*, ou 5 *niveaux*) ; (Rq : on peut

distinguer 2 types de facteurs, les facteurs *fixes* et les facteurs *aléatoires*).
Par ailleurs, il y a une *variable dépendante* (ou *variable réponse*), qui est ici la note.

Si l'on a *une seule* variable indépendante, l'analyse est dite *à un facteur*, "one-way design" en anglais. S'il y a *plusieurs* variables indépendantes, on parle d'*analyse factorielle*, ou de *plan factoriel*, "factorial design" en anglais. On parle d'*analyse multivariée* (abrégé MANOVA) lorsqu'il y a plusieurs variables dépendantes.
Remarque : la (ou éventuellement les) variable dépendante est toujours une variable quantitative (continue), et la (ou les) variable indépendante toujours une variable catégorique.

**C.** Préciser comment les différentes modalités des facteurs sont assignées aux unités expérimentales : il faut s'assurer d'une certaine *randomisation*, i.e sur l'exemple, il faut veiller à ce que les individus aient tous au départ la même chance de suivre chaque méthode d'enseignement.

Par ailleurs, il vaut mieux avoir, dans la mesure du possible, un *plan balancé*, i.e tel qu'un même nombre d'unités expérimentales (ici, 9) aient été soumises à chacun des traitements (ici 5). C'est bien le cas dans l'exemple considéré.

### 2.3 Analyse graphique des données

Graphe de "note" en fonction de "méthode d'enseignement", boxplots, etc...

### 2.4 Présentation du modèle

On veut comparer les $a$ (ici $a$=5) modalités d'un facteur quant aux moyennes d'une variable réponse (ou *variable indépendante*). On se donne un modèle qui décrit l'effet d'un facteur fixe sur la variable dépendante :

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

où :
- $Y_{ij}$ = valeur de la variable dépendante (ici la note) pour l'unité expérimentale $j$ (ici un individu, appelé $j$) et pour la modalité $i$ du traitement (ici la $i$ème méthode) ;
- $\mu$ = moyenne globale de la variable dépendante ;
- $\tau_i$ = effet de la modalité $i$ sur la variable dépendante ;
- $\epsilon_{ij}$ = termes d'erreur aléatoire, tels que l'on ait indépendance des $\epsilon_{ij}$, et que $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ ;
- $i = 1, 2, \cdots, a$ et $j = 1, 2, \cdots, n_i$.

Remarque : les quantités $\mu$ et $\tau_i$ sont inconnues, ce sont les *paramètres du modèle*, elles sont à estimer à l'aide des observations.

Exemple : le modèle dans le cas de l'exemple s'écrit de façon informelle :

" Note = $\mu + \tau_i +$ erreur "

où $\mu$ est la note moyenne de l'ensemble des individus, et $\tau_i$ est l'effet de la méthode $i$ sur la note.

### 2.5 Estimation des paramètres

Soit $n = n_1 + \cdots + n_a$ le nombre total d'individus. On note

$$\bar{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \text{ et } \bar{Y}_{..} = \frac{1}{a}\sum_{i=1}^{a}\bar{Y}_{i.} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}.$$

Les estimations des paramètres $\mu$ et $\tau_i$ $(i = 1, \cdots, a)$ s'écrivent :

$$\hat{\mu} = \bar{Y}_{..} \text{ et } \hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$$

et s'obtiennent par la méthode des moindres carrés. Cette méthode consiste à prendre les valeurs $\mu$ et $\tau_i$ qui minimisent simultanément

$$S = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \tau_i)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\epsilon_{ij}^2.$$

Une fois les paramètres estimés, on associe à chacune des observations la *valeur prédite* (ou *valeur ajustée*) définie par $\hat{Y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{Y}_{i.}$. De même, à chaque observation est associé un *résidu* $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.}$. Ce résidu quantifie l'écart entre une observation et la moyenne des unités de l'échantillon pour la modalité du facteur correspondante. C'est une approximation du terme d'erreur $\epsilon_{ij}$. Les résidus servent à valider le modèle.

### 2.6 Table d'analyse de la variance.

On décompose la variabilité pour quantifier les différentes sources de cette variabilité. Il y a d'une part la variabilité inter-groupes (entre les groupes) et d'autre part la variabilité intra-groupe (à l'intérieur de chaque groupe). La variabilité due à l'ensemble des facteurs non contrôlés (erreurs, bruit) est estimée par la variabilité à l'intérieur des groupes. En effet : pourquoi les scores des sujets à l'intérieur d'un groupe, ayant suivi une même méthode, varient-ils ? Justement en raison de tous les facteurs non contrôlés, de tous les facteurs autres que la méthode, qui est la même pour tous dans un même groupe.

$$\text{Variation totale} = \text{Variation inter-groupe} + \text{Variation intra-groupe}$$

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 &= \sum_{i=1}^{a} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 \\ \text{SC totale} &= \text{SC due aux traitements} + \text{SC des erreurs,}\end{aligned}$$

où SC est une abréviation pour "Somme des Carrés". L'utilisation de logiciels pour la plupart en langue anglaise entraîne que les notations suivantes sont souvent utilisées :

**SST** = "total sum of squares" = SC totale ;
**SSE** = "sum of squares of error" = SC des erreurs ;
**SSTR** = "sum of squares of treatment" = SC due aux traitements.

L'équation d'analyse de la variance s'écrit donc avec ces notations :

$$\text{SST} = \text{SSTR} + \text{SSE}.$$

NB : la variance est obtenue comme la variation divisée par le degré de liberté.

**Carrés moyens** : Lorsque les valeurs moyennes de la variable dépendante (dans l'exemple : la note) diffèrent d'une modalité à l'autre, la quantité

$$\textbf{MSTR} = \text{carrés moyens (MC) du modèle} = \frac{1}{a-1}\sum_{i=1}^{a} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$
$$= \text{variance inter-groupe}$$

a tendance à être *grande* comparativement à la quantité

$$\textbf{MSE} = \text{carrés moyens des erreurs} = \frac{1}{n-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$$
$$= \text{variance intra-groupe.}$$

Le tableau d'analyse de variance est un résumé de toutes les quantités définies plus haut :

| Source variation | SC écarts | nb de dl | variance | F-valeur | p |
|---|---|---|---|---|---|
| Traitement | SSTR | $a-1$ | MSTR | $F_{\text{obs}} = \frac{\text{MSTR}}{\text{MSE}}$ | $P(Z > F_{\text{obs}})^{\star}$ |
| Erreur | SSE | $n-a$ | MSE | | |
| Total | SST | $n-1$ | | | |

$\star$ la variable $Z$ suit une loi de Fisher, à $a-1$ et $n-a$ degrés de liberté.

Remarque : sur Statistica, le tableau d'analyse de variance est présenté sous la forme équivalente suivante :

| Effet | dl effet | MC effet | dl erreur | MC erreur | F | niveau p |
|---|---|---|---|---|---|---|
| 1 | $a-1$ | MSTR | $n-a$ | MSE | $\frac{\text{MSTR}}{\text{MSE}}$ | $p = P(Z > F_{\text{obs}})$ |

Le test fait dans la table d'analyse de variance est un test de l'hypothèse

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0,$$

i.e : $H_0$ : "il n'y a pas d'effet dû au traitement",
contre l'hypothèse

$$H_1 : \tau_i \neq 0 \text{ pour au moins une valeur de } i,$$

i.e : $H_1$ : "il y a un effet dû au traitement."

La statistique de test est $F_0 = \text{MSTR}/\text{MSE}$. Sous $H_0$, on peut montrer que $F_0$ suit une distribution de Fisher à $a-1$ et $n-a$ degrés de liberté. En effet, sous l'hypothèse d'égalité des moyennes de groupes, i.e sous $H_0$, à la fois la variance intra-groupe ($\text{MSE}=\text{SSE}/(n-a)$) **et** la variance inter-groupe ($\text{MSTR}=\text{SSTR}/(a-1)$) sont des estimateurs de $\sigma^2$, alors que sous $H_1$, seule MSE est un estimateur de $\sigma^2$. Or un test de Fisher est un test de comparaison de variances, sous hypothèse de normalité. C'est donc lui qui est utilisé pour tester $H_0$, via MSE et MSTR.

### 2.7 Validation du modèle

Elle se fait par l'intermédiaire de l'analyse des résidus $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.}$. On vérifie les 4 points suivants :

1. *Absence de corrélation entre les observations.*
   En pratique, une randomisation appropriée de l'expérience permet d'assurer l'indépendance des observations. Mais lorsque les observations sont recueillies dans le temps, il se peut qu'il y ait une dépendance temporelle, i.e une corrélation entre des observations successives.
   Le graphe des résidus $e_{ij}$ en fonction du temps, et/ou en fonction des valeurs prédites $\hat{Y}_{ij}$, doit présenter des points répartis de manière homogène.
2. *Homogénéité de la variance des résidus.*
   Le graphe des résidus $e_{ij}$ en fonction des valeurs prédites $\hat{Y}_{ij}$ doit être relativement homogène (éviter la forme "entonnoir").
   Deux tests statistiques sont à disposition : le test de Bartlett et le test de Levene ; (celui de Levene est à recommander, il est notamment plus robuste à l'absence de normalité).
3. *Absence de données influentes.*
   Il s'agit de vérifier que les résidus standardisés $e_{ij}/\sqrt{\text{MSE}}$ sont quasiment tous (environ 95%) dans l'intervalle $[-2, 2]$, et que presque aucun d'entre eux n'est à l'extérieur de $[-3, 3]$.
4. *Normalité des résidus.*
   Droite de Henry (en anglais : "Normal probability plot") ; tests de Shapiro-Wilk et de Kolmogorov-Smirnov (de préférence celui de Shapiro-Wilk).

Remarques : **(i)** l'analyse de variance est relativement robuste par rapport à l'absence de normalité : même si cette hypothèse n'est qu'approximativement vérifiée, avec une distribution symétrique tout-de-même, on pourra utiliser l'anova ; sinon, il faut avoir recours aux procédures non paramétriques (procédure de Kruskal-Wallis), ou transformer les données (log, racine carrée, etc.) ; **(ii)** en présence d'hétérogénéité des variances, les tests d'anova sont peu affectés dans le cas d'un facteur fixe, surtout pour un plan balancé ; sinon, il existe des procédures sans hypothèse d'homogénéité : procédure de Welch, ou de Box... On s'accorde à dire (même si cette règle est bien loin d'être satisfaisante en toute généralité) que l'on peut utiliser l'anova tant que la plus grande variance n'est pas plus de quatre fois supérieure à la plus petite, et qu'aucun échantillon n'est plus de 1.5 fois plus grand qu'un autre.

### 2.8 Comparaisons a posteriori

Si l'on a validé les hypothèses du modèle d'analyse de variance, et que la table d'anova donne une différence significative (une valeur $p$ dans la colonne la plus à droite très petite - par exemple inférieure à 0.05 -, conduisant à rejeter l'hypothèse $H_0$), on peut souhaiter raffiner l'analyse, en cherchant les traitements qui sont différents. On utilise pour cela des *tests de comparaisons multiples.* De nombreuses méthodes existent. Les avis diffèrent sur "le" test à utiliser de préférence. Un test est dit *conservatif* si son risque de première espèce réel est inférieur au risque choisi, i.e s'il a tendance à rejeter $H_0$ plus difficilement que prévu théoriquement. Plus un test est conservatif, moins il est puissant. Il y a donc une balance à faire entre risque de première espèce et puissance, éternel dilemne !

Statistica propose (dans sa rubrique "Comparaisons post hoc") :

1. le test LSD ("Least Significant Difference") ou comparaison planifiée ;
2. le test de Scheffé ;
3. le test de Newman-Keuls (NK) ;
4. le test multiplicatif des étendues de Duncan ;
5. le test HSD ("High Significant Difference") de Tukey, pour des échantillons de même taille seulement ;
6. le test HSD pour $N$ différents (Spjotvoll & Stoline).

Voici leur ordonnancement en terme de conservatisme :

| Libéral | | | Conservatif |
|---|---|---|---|
| $<$ − − − − − − − − − − | − − − − − − − − − − | − − − − − − − − − − | − − − − − − − $>$ |
| Duncan / LSD | NK | HSD Tukey | Scheffé |

** On préconisera donc d'utiliser la méthode HSD de Tukey, ou Newman-Keuls. **

Remarque : dans le tableau affiché par Statistica, les différences significatives (au seuil 5% par défaut) sont en rouge, et les autres en noir. Une fois noté quelles sont les différences significatives, on peut effectuer un tracé des moyennes pour savoir si une moyenne est significativement inférieure ou supérieure à une autre. Avec Statistica, cela se fait avec la rubrique "Moyennes/Graphiques", et le choix "Graphique".

Remarque : On peut également s'intéresser aux comparaisons multiples lorsque la table d'analyse de variance ne donne pas un test de Fisher significatif. Dans une telle situation, on ne peut toutefois *pas* utiliser le test LSD.

Mise en garde : il est vivement conseillé de décider *au départ* quel type de test de comparaisons multiples on souhaite utiliser. En aucun cas on ne choisit en les regardant tous, et en conservant celui qui semble le plus "coller à ce que l'on voulait faire dire aux données" ! !

### 2.9 Comparaisons a priori, ou planifiées, ou "contrastes"

Il peut arriver que l'on ne veuille pas spécifiquement tester l'hypothèse $[H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0]$, c'est-à-dire que toutes les comparaisons ne soient pas pertinentes, mais que l'on soit plutôt intéressé par deux ou trois comparaisons plus ciblées. On peut imaginer, par exemple, dans un cas d'analyse de variance à un facteur à 4 modalités, vouloir tester si le nouveau traitement 1 est différent de la moyenne des 3 autres (anciens) traitements 2, 3 et 4. On souhaite donc dans ce cas tester $[H_0 : \tau_1 = (\tau_2 + \tau_3 + \tau_4)/3]$. Ceci est possible avec l'instruction "comparaisons planifiées".

Remarque : pour de telles comparaisons, il faut toujours avoir une somme des coefficients égale à 0, soit : $\sum_{i=1}^{a} c_i \tau_i = 0$.

Remarque : lorsque l'on souhaite faire une comparaison planifiée, il faut la faire *directement*, sans faire une analyse de variance auparavant !

Date : mars 2007.