ANALYSE DE LA VARIANCE A UN FACTEUR

Quelques idées qui m'ont parues intéressantes et exploitables
pour l'initiation à la statistique

Roland Chiavassa

1 L'analyse de la variance à un facteur

1.1 Introduction

L'analyse de la variance (ANOVA pour ANalysis Of VAriance) à un facteur est une technique statistique permettant de comparer les moyennes d'une variable statistique dans (au moins) trois échantillons aléatoires afin de déterminer s'il existe une différence significative entre les populations dont sont issus les échantillons.

Exemple : Un ensemble de pièces mécaniques a été divisé en 3 lots (3 populations). Chacun de ces lots a été soumis à un traitement thermique spécifique (le facteur à 3 niveaux). La question que l'on se pose : le traitement thermique a-t-il une influence sur une caractéristique mécanique de la pièce (la résistance mécanique par exemple).

1.2 Intérêt de la mise en oeuvre de l'analyse de la variance

Les échantillons doivent vérifier 3 propriétés : les tirages pour constituer ces échantillons sont aléatoires, la distribution de la variable statistique doit être normale dans chacun d'eux, les variances de cette variable statistique (calculées pour chacun des échantillons) doivent être égales. La vérification de chacune de ces propriétés met en oeuvre des aspects intéressants du cours de statistique :

- tirage aléatoire dans chacune des populations par génération de nombres aléatoires.
- vérification de la répartition normale de la variable mesurée dans chaque échantillon à l'aide du test du χ^2 par exemple.
- vérification de l'homogénéïté des variances. Deux tests (parmi plusieurs autres) sont possibles :
  * le test de Barlett qui se ramène à un test du χ^2.
  * le test du Log-Anova qui utilise à son tour les techniques de l'analyse de la variance.

1.3 Principe

On tire un échantillon dans chaque population. On compare les moyennes mesurées sur chaque échantillon de la résistance mécanique. Il s'agit donc d'un test portant sur l'égalité des moyennes. L'hypothèse nulle de ce test est :

$$\mathbf{H}_0 : \text{les moyennes sont égales}$$

ce qui peut se traduire dans l'exemple ci-dessus par « le traitement thermique n'a pas d'influence sur la résistance mécanique ». L'hypothèse alternative est :

$$\mathbf{H}_1 : \text{au moins 2 moyennes sont différentes}$$

1.3.1 *Aspect graphique*

Voir les graphiques ci-dessous dans lesquels m_1, m_2, m_3 sont respectivement les moyennes empiriques des échantillons 1 , 2 , 3.

Dans le cas n ° 1, les valeurs des échantillons recouvrent de larges intervalles communs : il y a de fortes chances que les moyennes empiriques des 3 échantillons soient des estimations d'une même moyenne μ.

Dans le cas n ° 2, le regroupement des valeurs pour chaque échantillon, conduit naturellement à penser que les 3 moyennes empiriques sont des estimations de valeurs réellement différentes.

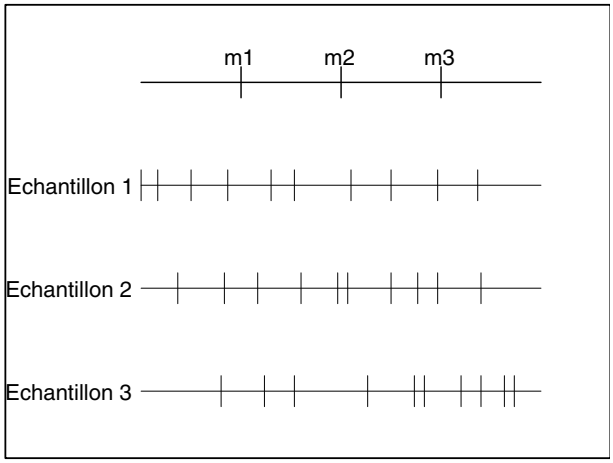

FIG. 1 – *Cas n ° 1*

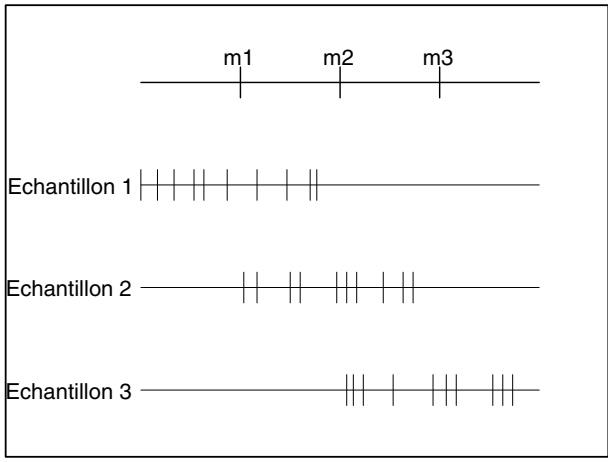

FIG. 2 – *Cas n ° 2*

1.3.2 *Conclusion*

« L'éloignement » des valeurs m_1, m_2, m_3 étant mesuré, ainsi que la dispersion des valeurs mesurées autour de ces moyennes, il faudra alors comparer ces deux dispersions pour décider si l'on est dans le cas n ° 1 ou dans le cas n ° 2.

1.4 *Mise en oeuvre mathématique*

La population totale a été divisée en p sous populations, on dit aussi que *le facteur controlé* présente p traitements. Pour i variant de 1 à p, on tire dans la i^{ieme} sous population un échantillon (ou groupe) de n_i individus. On mesure pour chacun des individus, la variable étudiée : $x_{i,j}$ représente la valeur mesurée sur

le i^{ieme} individu du j^{ieme} échantillon. On définit aussi la moyenne $\bar{x}_j$ du j^{ieme} échantillon et la moyenne globale $\bar{x}$ par :

$$\bar{x}_j = \frac{1}{n_j} \sum_{k=1}^{k=n_j} x_{k,j} \qquad \bar{x} = \frac{1}{n} \sum_{j=1}^{j=p} n_j \bar{x}_j$$

Les calculs qui vont suivre ne sont valables que si les 3 conditions suivantes sont respectées :

1. les échantillons ont été choisis aléatoirement et sont indépendants.
2. les distributions des sous populations sont normales (ou gaussiennes).
3. ces distributions possèdent la même variance σ^2.

On démontre alors que :

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

En notant SCE pour « Somme des Carrés des Ecarts », on a :

$$SCE_{tot} = \sum_{j=1}^{j=p} \sum_{i=1}^{i=n_j} (x_{i,j} - \bar{x})^2 : SCE \text{ « totale »}$$

$$SCE_{ent} = \sum_{j=1}^{j=p} n_j (\bar{x}_j - \bar{x})^2 : SCE \text{ « entre les groupes »}$$

$$SCE_{int} = \sum_{j=1}^{j=p} \sum_{i=1}^{i=n_j} (x_{i,j} - \bar{x}_j)^2 : SCE \text{ « à l'intérieur des groupes »}$$

La relation (1) se traduit alors par la relation (2) dite *équation de la variance* :

$$SCE_{tot} = SCE_{ent} + SCE_{int} \quad (2)$$

Il s'agit maintenant de déterminer le « poids » respectif de chacun de deux termes du membre de droite de (2). Pour cela, on fait intervenir :

– la *variance entre les groupes*

$$s^2_{ent} = \frac{SCE_{ent}}{p-1}$$

où $p-1$ est le nombre de degrés de liberté.

– la *variance à l'intérieur des groupes*

$$s^2_{int} = \frac{SCE_{int}}{n-p}$$

où $n-p$ est le nombre de degrés de liberté.

En admettant les résultats suivants :
s^2_{int} est une estimation de σ^2 qui *ne dépend pas de l'hypothèse nulle*
s^2_{ent} est une estimation de σ^2 *sous l'hypothèse nulle*
le rapport $F = \frac{s^2_{ent}}{s^2_{int}}$ traduit « l'écart à l'hypothèse nulle ». On rejettera l'hypothèse nulle, avec un seuil de risque α fixé, si F est trop grand.

1.4.1 Loi théorique de F

Sous l'hypothèse nulle $\mathbf{H}_0$, F suit une loi de Fisher à $\nu_1 = p-1$ et $\nu_2 = n-p$ degrés de liberté. Cette loi est tabulée pour différentes valeurs des d.d.l. ν_1 et ν_2.

Définition : La variable aléatoire réelle X suit une loi de Fisher à ν_1 et ν_2 d.d.l. si sa fonction de densité est de la forme :

$$f(x) = 0 \qquad \text{pour} \qquad x \leqslant 0$$

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{x^{\frac{\nu_1-2}{2}}}{\left(1+\frac{\nu_1}{\nu_2}x\right)^{\frac{\nu_1+\nu_2}{2}}} \qquad \text{pour} \qquad x > 0$$

où $\Gamma(\nu)$ est la fonction Gamma définie par :

$$\Gamma(\nu) = \int_0^{+\infty} t^{\nu-1} e^{-t}\, dt \qquad \text{pour} \qquad \nu > 0$$

Exemple : densité de la loi de Fisher pour $\nu_1 = 9$ et $\nu_2 = 40$

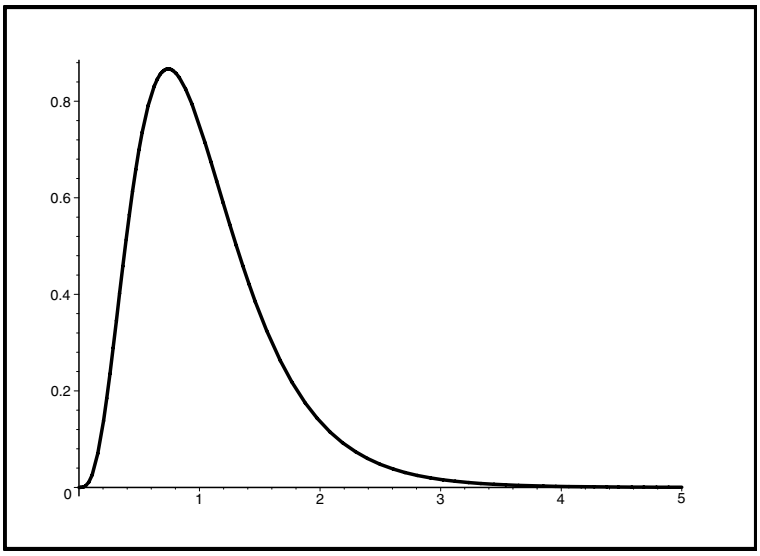

FIG. 3 – *Densité de probabilité de la loi de Fisher à 9 et 40 d.d.l*

1.5 Tableau d'analyse de la variance ou tableau ANOVA

Les calculs sont disposés dans le tableau suivant :

Source de variation	Somme des carrés	d.d.l.	Variances	F
entre les groupes	SCE_{ent}	$p-1$	s^2_{ent}	$\frac{s^2_{ent}}{s^2_{int}}$
à l'intérieur des groupes	SCE_{int}	$n-p$	s^2_{int}	
totale	SCE_{tot}	$n-1$		

Pour α fixé, (en général $\alpha = 0.05$), n et p étant connus (par exemple $n = 50$ et $p = 10$), la table de la loi de Fisher donne, pour $\nu_1 = p-1 = 9$ et $\nu_2 = n-p = 40$, $F_{th} = 2.12$. Si le F trouvé dans le tableau est supérieur à F_{th} on rejette l'hypothèse nulle $\mathbf{H}_0$.

2 Généralisation

L'analyse de la variance à un facteur présentée ci-dessus peut se généraliser. Il existe ainsi une analyse de la variance à deux facteurs contrôlés. En reprenant l'exemple initial : le facteur A est le traitement thermique avec a niveaux, le facteur B est la composition de l'alliage avec b niveaux. Le raisonnement est calqué sur le précédent, il permet de prendre en compte l'interaction des facteurs A et B.

D'autres généralisations sont possibles : plans d'expérience, plans en carrés latins, analyse de la covariance ... Voir les ouvrages ci-dessous.

Bibliographie

Quelques ouvrages de base en Statistique :

1. Statistique Dictionnaire encyclopédique. Y. Dodge. DUNOD
2. Introduction à la statistique. Amzallag, Piccioli. HERMANN
3. Guide de statistique appliquée. Manoukian. HERMANN
4. Aide–mémoire Statistique. CISIA.CERESTA
5. Statistique. H. Wonnacott, R. Wonnacott. ECONOMICA
6. Probabilités et Statistique tome 1. Dacunha-Castelle, Duflo. DUNOD
7. Exercices de probabilités et Statistique tome 1. Dacunha-Castelle, Duflo. DUNOD

3 Exemples traités

3.1 Exemple n ° 1

Les températures moyennes journalières ont été mesurées 4 fois en 3 lieux différents d'un même pays. Peut-on dire, au vu de ces données, que ces 3 lieux ont une même température moyenne ?

Dans cet exemple académique, on supposera que les 2 hypothèses de validité de l'ANOVA sont respectées : répartition normale des températures en chacun des lieux, égalité des variances.

Tableau des données :

Lieux	A	B	C
Températures	22	28	26
	24	31	27
	27	27	28
	25	27	27
Moyennes	24.5	28.25	27

Tableau ANOVA :

Source de variation	Somme des carrés	d.d.l.	Variances	F
entre les groupes	29.17	2	14.58	5.09
à l'intérieur des groupes	25.75	9	2.86	
totale	54.92	11		

Pour $\alpha = 0.05 \quad \nu_1 = 2 \quad \nu_2 = 9$, la table de la loi de Fisher donne $F_{th} = 4.26$. La valeur trouvée pour F dans le tableau est supérieure à la valeur théorique, On rejette l'hypothèse nulle de l'égalité des moyennes. Ces 3 lieux ont donc des températures moyennes différentes.

3.2 Exemple n ° 2

Les données sont extraites d'une étude réalisée par Fisher en 1938. L'échantillon comporte 120 iris de 3 espèces différentes : *iris setosa, iris versicolor, iris virginica* à raison de 40 individus par espèce.Les mesures effectuées sont :

1. la longueur du sépale
2. la largeur du sépale
3. la longueur du pétale
4. la largeur du pétale

Dans cette étude le facteur possède 3 niveaux (les trois espèces). On désire savoir si l'une des dimensions mesurées ci-dessus (l'un des critères) permet de distinguer les 3 espèces d'iris. Choisissons par exemple le critère n ° 2 : la largeur du sépale.

Test de normalité

La première étape consiste à vérifier les hypothèses de validité de l'ANOVA. Il faut s'assurer que la répartition du critère est normale pour chacune des espèces. Nous choisissons pour cela un test du χ^2. Les intervalles sont choisis de façon à ce que les effectifs observés ne soient pas trop faibles.

Pour la variété *setosa* :

intervalles	$[2.9, 3.15]$	$[3.15, 3.45]$	$[3.45, 3.65]$	$[3.65, 3.95]$	$[3.95, 4.40]$
effectifs observés n_i	10	13	7	6	4
effectifs théoriques n_i^*	8.04	11.85	8.43	8.32	3.36

Sur cet échantillon, la valeur observée du χ^2 est :

$$\chi^2_{obs} = \sum_{i=1}^{i=5} \frac{(n_i - n_i^*)^2}{n_i^*} = 1.60$$

Pour le risque $\alpha = 0.05$ et un nombre de degrés de liberté égal à : $\nu = n - 1 - r = 2$, le seuil théorique est lu dans une table : $\chi^2_{th} = 5.99$. On a donc $\chi^2_{obs} < \chi^2_{th}$ par conséquent, on *accepte* l'hypothèse de normalité pour la distribution de la largeur du sépale des iris sétosa.

Pour la variété *versicolor* :

intervalles	$[2, 2.35]$	$[2.35, 2.55]$	$[2.55, 2.85]$	$[2.85, 3.05]$	$[3.05, 3.40]$
effectifs observés n_i	5	6	10	11	8
effectifs théoriques n_i^*	3.93	5.86	13.54	8.32	3.35

Sur cet échantillon, la valeur observée du χ^2 est :

$$\chi^2_{obs} = \sum_{i=1}^{i=5} \frac{(n_i - n_i^*)^2}{n_i^*} = 2.09$$

Pour le risque $\alpha = 0.05$ et un nombre de degrés de liberté égal à : $\nu = n - 1 - r = 2$, le seuil théorique est lu dans une table : $\chi^2_{th} = 5.99$. On a donc $\chi^2_{obs} < \chi^2_{th}$ par conséquent, on *accepte* l'hypothèse de normalité pour la distribution de la largeur du sépale des iris versicolor.

Pour la variété *virginica* :

intervalles	$[2.2, 2.65]$	$[2.65, 2.85]$	$[2.85, 3.05]$	$[3.05, 3.25]$	$[3.25, 3.80]$
effectifs observés n_i	6	11	11	6	6
effectifs théoriques n_i^*	7.14	7.74	9.34	8.01	7.78

Sur cet échantillon, la valeur observée du χ^2 est :

$$\chi^2_{obs} = \sum_{i=1}^{i=5} \frac{(n_i - n_i^*)^2}{n_i^*} = 2.76$$

Pour le risque $\alpha = 0.05$ et un nombre de degrés de liberté égal à : $\nu = n - 1 - r = 2$, le seuil théorique est lu dans une table : $\chi^2_{th} = 5.99$. On a donc $\chi^2_{obs} < \chi^2_{th}$ par conséquent, on *accepte* l'hypothèse de normalité pour la distribution de la largeur du sépale des iris virginica.

Test d'égalité des 3 variances

L'autre hypothèse à vérifier est l'égalité des 3 variances. L'hypothèse nulle est dans ce cas : *les trois variances sont égales* Nous effectuons pour cela le test de Barlett.

Test de Barlett

C'est un test du χ^2. On pose pour $i = 1, \cdots, p$ où p est le nombre de traitements du facteur contrôlé (ici $p = 3$) : $\nu_i = n_i - 1$ (n_i est la taille du i^{ieme} échantillon) ainsi que : S_i^2 variance empirique du i^{ieme} échantillon. On a aussi : $\nu = \sum_{i=1}^{i=p} \nu_i$ et $S^2 = \frac{1}{\nu} \sum \nu_i \, S_i^2$. On démontre que, sous l'hypothèse nulle, la variable :

$$\nu \ \ln(S^2) - \sum_{i=1}^{i=p} \nu_i \ln(S_i^2)$$

suit une loi du χ^2 à $p - 1$ degrés de liberté. Pour l'exemple étudié, on obtient :

S_1^2	S_2^2	S_3^2	S^2
0.1303	0.1109	0.1132	0.1181

$\chi^2_{obs} = \nu \ \ln(S^2) - \sum_{i=1}^{i=p} \nu_i \ln(S_i^2) = 0.3035$. Pour le risque $\alpha = 0.05$ et un nombre de degrés de liberté égal à : $p - 1 = 2$, le seuil théorique est lu dans une table : $\chi^2_{th} = 5.99$. On a donc $\chi^2_{obs} < \chi^2_{th}$ par conséquent, on *accepte* l'hypothèse d'égalité des variances.

Analyse de la variance

Le tableau d'analyse de la variance :

Source de variation	Somme des carrés	d.d.l.	Variances	F
entre les groupes	9.696	2	4.848	41.045
à l'intérieur des groupes	13.820	117	0.118	
totale	23.52	119		

Pour $\alpha = 0.05 \quad \nu_1 = 2 \quad \nu_2 = 117$, la table de la loi de Fisher donne $F_{th} = 3.07$. La valeur trouvée pour F dans le tableau est supérieure à la valeur théorique, On rejette l'hypothèse nulle de l'égalité des moyennes. Pour le critère : largeur sépale, les moyennes des trois groupes sont différentes.