Université Paris Ouest - Nanterre La Défense
U.F.R. SPSE - Master1
PMPSTA21 Méthodes Statistiques pour l'analyse de données en psychologie

Chapitre 4 - Comparaison de plusieurs moyennes pour des échantillons indépendants

1 Motivation

Supposons que l'on souhaite évaluer l'effet de cinq traitements différents sur le comportement des patients dépressifs âgés de 18 à 50 ans. On mesure le niveau de dépression (donné par un score). Cinq échantillons, de 9 patients chacun, ont été considérés.

Nous avons ici une variable quantitative (score) et une variable qualitative (traitement) dont les modalités sont : traitement 1, traitement 2, ..., traitement 5. On se pose la question de savoir si ces 5 traitements diffèrent. Nous souhaitons comparer l'effet des traitements et voir s'il y a un lien entre la variable quantitative et la variable qualitative.

En utilisant des tests $t-$ Student de comparaison de deux moyennes pour deux échantillons indépendants, nous devrions comparer le traitement 1 avec le traitement 2, le traitement 1 avec le traitement 3, ... le traitement 1 avec le traitement 5. Il faudrait alors faire 10 tests de comparaisons de deux moyennes, ce qui implique un nombre considérable de calculs. Le problème ici est que le test t-Student devient impraticable. On a alors recours à l'analyse de variance (appelée souvent ANOVA) développée par Fisher (sous hypothèse de normalité).

L'ANOVA est un test statistique qui généralise le test $t-$ Student au cadre de comparaisons de plusieurs moyennes. On l'applique dès lors que l'on étudie les effets d'une ou plusieurs variables qualitatives sur une variable quantitative.

2 ANOVA à un facteur

On utilise l'analyse de la variance à un facteur quand on dispose :

- d'une variable quantitative Y (variable dépendante, VD) ;
- d'une variable qualitative X à k modalités (variable indépendante, VI, appelée facteur à k niveaux) ;
- de k échantillons indépendants $(E_1, \ldots, E_k)$ de taille $n_1, \ldots, n_k$, respectivement.

On cherche un lien entre la VI et la VD. Plus précisément, on veut étudier l'influence des différentes modalités de la VI sur la VD.

Exemples.

a) Etude sur le stress du personnel enseignant supérieur : le niveau de responsabilité a-t-il un impact sur l'état de stress ? Un facteur, variable indépendante VI à 4 niveaux : 4 catégories de personnels (professeurs, maîtres de conferences, ATER et autres). Une variable dépendante VD : la mesure de stress.

b) L'etude de la réussite scolaire, pour d'élèves de troisième de différentes Pays. Un facteur, VI à 3 niveaux : Pays 1, Pays 2, Pays 3. Une VD : performance à l'examen.

c) Le taux de cholestérol en fonction de la CSP. On se donne 5 CSP : Retraités, étudiants, agriculteurs, cadres, ouvriers. Une VD : taux de cholestérol.

Remarque 1 : Il existe différentes types d'ANOVA qui se distinguent par le nombre de facteurs étudiés. Si l'on a une seule variable indépendante, l'analyse est dite à un facteur. S'il y a plusieurs variables indépendantes, on parle d'analyse factorielle, ou de plan factoriel.

Remarque 2 : Nous supposons ici que les k échantillons E_j $(j = 1 \dots k)$ ont tous la même taille $(n_1 = n_2 = \dots = n_k)$. Il est possible de généraliser l'ANOVA aux cas d'échantillons ayant des tailles différentes.

Soit r la taille commune des échantillons, Y_j $(j = 1, \dots, k)$ la variable correspondant aux valeurs de la VD Y observées sur l'echantillon E_j et $\bar{Y}_j$ la moyenne de Y_j. Avant toute analyse, il est intéressant de représenter les données. Par exemple, les valeurs de Y observées sur E_1 sont : $y_{11}, y_{21}, y_{31}, \dots, y_{r1}$ et la moyenne observée est $\bar{y}_1 = (y_{11} + y_{21} + y_{31} + \dots + y_{r1})/r$. Pour calculer la moyenne observée de Y, notée $\bar{y}$, sur toute la population, on peut réutiliser les moyennes sur les k échantillons. En effet, comme on est dans le cas où tous les échantillons ont la même taille :

$$\bar{y} = \frac{1}{k \times r} \sum_{j=1}^{k} \sum_{i=1}^{r} y_{ij} = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{r} \sum_{i=1}^{r} y_{ij} = \frac{1}{k} \sum_{j=1}^{k} \bar{y}_j.$$

Le tableau ci-dessous contient les valeurs de la VD Y observées sur l'ensemble des k échantillons.

Individu	Niveau 1	Niveau 2	...	Niveau k
1	y_{11}	y_{12}	$\dots$	y_{1k}
2	y_{21}	y_{22}	$\dots$	y_{2k}
3	y_{31}	y_{32}	$\dots$	y_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	y_{r1}	y_{r2}	$\dots$	y_{rk}
Moyenne	$\bar{y}_1$	$\bar{y}_2$	$\dots$	$\bar{y}_k$

L'ANOVA nous indique si les différents échantillons proviennent ou non de la même population $\mathcal{P}$.

Conditions d'application. Pour pouvoir appliquer l'ANOVA, il est indispensable que les 2 propriétés soient vérifiées :

1. les tirages effectués pour constituer les échantillons sont aléatoires et indépendants ;
2. la distribution de chaque variable Y_j est normale, de moyenne μ_j et de variance σ^2 (même variance pour chaque population : homogénéité des variances ou homoscédasticité).

Remarque 3 : Une manière plus formelle de représenter notre cadre consiste à introduire la notation

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1, ..., r, \quad j = 1, ..., k,$$

où r est la taille commune des échantillons, et ε_{ij} (erreurs correspondent aux fluctuations expérimentales pour chaque valeur de Y_{ij} mesurée) est une variable normale de moyenne nulle et variance σ^2. Cette notation indique que la i-ème observation associée à l'échantillon j est égale à la somme de sa moyenne μ_j et d'une 'erreur'. Une autre formulation du problème précèdent est

$$Y_{ij} = \mu + a_j + \varepsilon_{ij}, \quad i = 1, ..., r, \quad j = 1, ..., k,$$

avec $\mu = \frac{1}{k}(\mu_1 + \mu_2 + \ldots + \mu_k)$ la moyenne globale ou effet moyen de la VD et a_j l'effet principal du niveau j du facteur sur la VD. On remarque que dans notre notation $\mu_j = \mu + a_j$. Dans ce chapitre, on écrira le modèle avec μ_j ou avec $\mu + a_j$.

Remarque 4 : Les quantités, μ, μ_j $(j = 1, \ldots, k)$ et a_j $(j = 1, \ldots, k)$ étant inconnues, ce sont des paramètres à estimer à l'aide des observations.

Reprenons l'Exemple 1 (b). Nous avons trois échantillons d'élèves de troisième qui font leurs études dans trois pays, Pays 1, Pays 2, Pays 3. Chaque échantillon est composé de 5 élèves aléatoirement choisis parmi la population des élèves du pays. On fait passer le même test de logique (noté sur 100) aux trois échantillons d'élèves.

— **Population** : élèves de troisième qui font leurs études dans trois pays, Pays 1, Pays 2, Pays 3.

— **VI (facteur)** : le pays. Variable qualitative à trois modalités (niveaux) (Pays 1, Pays 2, Pays 3).

— **VD** : performance à l'examen de logique. Variable quantitative.

Nous voulons déterminer si les élèves des trois pays ont des performances différentes ou non.

L'analyse de variance (ANOVA) va répondre à la question suivante : *Y a-t-il une influence du Pays sur la performance à l'examen de logique ?*.

Cela revient à tester :

H_0 : La performance en logique est la même pour les élèves des trois pays.
H_1 : La performance en logique est différente dans au moins deux pays.

or

$H_0 : \mu_1 = \mu_2 = \mu_3$.
H_1 : il existe au moins deux moyennes μ_j différentes.

or

$H_0 : a_j = 0$ pour tout $j = 1, \ldots, k$.
H_1 : il existe au moins deux a_j non nuls.

Nous avons rempli le tableau suivant avec des données obtenues pou 15 élèves répartis sur 3 échantillons indépendants. Comme on le voit ici, l'appartenance à un pays plutôt qu'un autre semble avoir un effet important.

Eleve	Pays 1	Pays 2	Pays 3
1	30	40	50
2	35	45	55
3	40	50	60
4	45	55	65
5	50	60	70
Moyenne	40	50	60

Pour les données du tableau, il y a 3 échantillons ($k = 3$) et 5 observations dans chaque échantillon ($r = 5$). Le nombre total d'observations est $n = r \times k = 3 \times 5 = 15$. Dans cet exemple on suppose que les 2 conditions sont vérifiées.

Dans la Figure 1 on a tracé les valeurs de la VD (Performance) en fonction de la VI (Pays) à trois modalités (1, 2 et 3). En regardant les observations (pour chaque modalité) on remarque que les variances observées sont égales dans les trois échantillons.

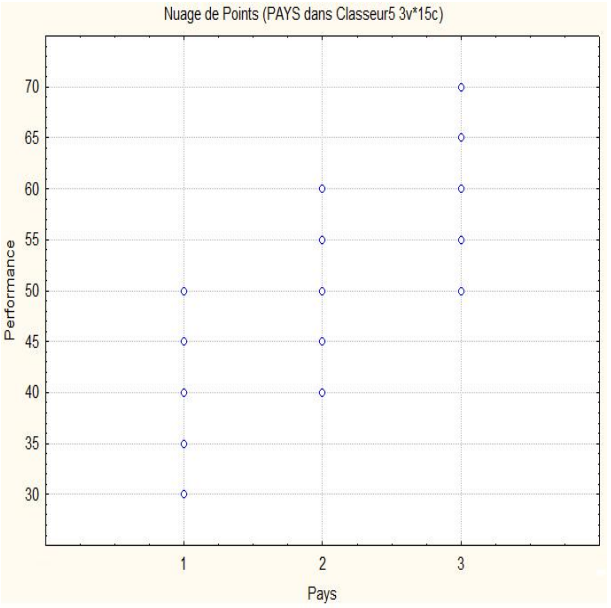

Figure 1

2.1 Etude descriptive des données

— **Moyennes des k échantillons et moyenne globale.**

On calcule les moyennes des k échantillons, notées $\bar{Y}_j$ $(j = 1 \dots k)$, ainsi que la moyenne des moyennes $\bar{Y} = \frac{1}{k}\sum_{j=1}^{k} \bar{Y}_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{j=1}^{r} Y_{ij}$ qui est la moyenne de toutes les données individuelles.

Remarque 5 : Les estimations des paramètres μ_j $(j = 1, \dots, k)$, μ et a_j $(j = 1, \dots, k)$ sont données par $\bar{Y}_j$, $\bar{Y}$ et $\bar{Y}_j - \bar{Y}$ $(j = 1, \dots, k)$, respectivement. Une fois les paramètres estimés, on associe à chacune des observations la valeur prédite (ou valeur ajustée) définie par $\hat{Y}_{ij} = \hat{\mu} + \hat{a}_j = \bar{Y}_j$. De même, à chaque observation est associé un résidu $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_j$.

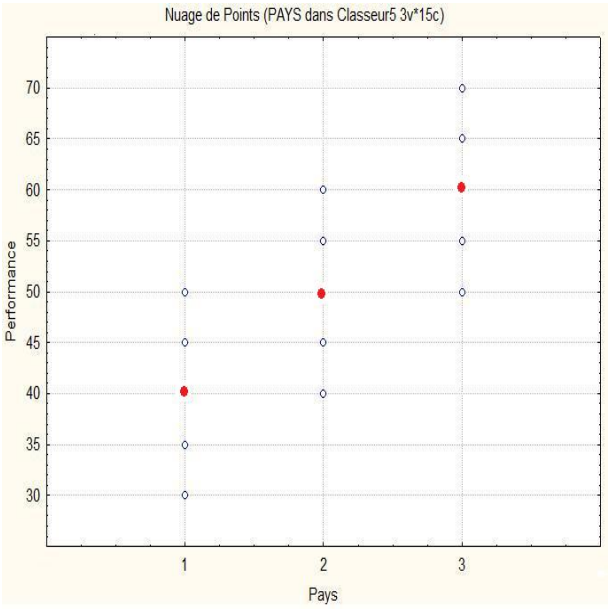

Figure 2

Dans cet exemple la moyenne de chaque échantillon regroupe 5 observations. La valeur de la moyenne obtenue dans chaque échantillon est :

$$\bar{y}_1 = 40, \quad \bar{y}_2 = 50, \quad \bar{y}_3 = 60$$

comme on peut le visualiser sur la Figure 2 (voir les cercles pleins). La valeur observée de la moyenne globale (moyenne des moyennes) est

$$\bar{y} = (40 + 50 + 60)/3 = 150/3 = 50.$$

— **Variabilité intergroupe : la somme des carrés moyens intergroupe.**

Les trois moyennes correspondant à l'Exemple 1 (b) ne sont pas identiques, il existe une variabilité due à la difference entre les moyennes $\mu_j, j = 1, \dots, 3$. Nous allons quantifier cette variabilité, que nous appelons *variabilité intergroupe* (variabilité entre les différents groupes), à l'aide d'une statistique dite "carré moyen intergroupe", notée CM_{inter}. Cette statistique se calcule en utilisant la formule suivante

$$CM_{\text{inter}} = \frac{SC_{\text{inter}}}{k-1},$$

avec SC_{inter} la somme des carrés (SC) des écarts intergroupe (entre les moyennes des groupes et la moyenne globale)

$$SC_{\text{inter}} = r \sum_{j=1}^{k} (\bar{Y}_j - \bar{Y})^2.$$

avec r le nombre d'individus dans chaque échantillon. Dans la première formule on divise SC_{inter} par les degrés de liberté $k-1$.

Ici et dans tout le chapitre on note la valeur observées d'une variable quelconque X par x. Pour les données du tableau, la valeur observée de la somme des écarts intergroupe (notée sc_{inter}) est :

$$sc_{\text{inter}} = 5 \times [(40-50)^2 + (50-50)^2 + (60-50)^2] = 1000.$$

Remarque 6 : Plus la valeur observée de la somme des écarts intergroupe, sc_{inter}, est proche de zéro, plus les moyennes μ_j sont proches les unes des autres.

La valeur observée du carré moyen intergroupe est

$$cm_{\text{inter}} = 1000/(3-1) = 1000/2 = 500.$$

— **Variabilité intragroupe : la somme des carrés moyens intragroupe.**

Il faut remarquer qu'il est possible que la variabilité entre les élèves d'un même pays soit aussi grande que celle existant entre les élèves des différents pays. Dans l'Exemple 1 (voir tableau), nous observons que les élèves d'un même pays n'obtiennent pas tous le même résultat. Il est possible de quantifier cette variabilité, que l'on appelle *variabilité intragroupe*(qui est la variabilité à l'intérieur de chaque groupe) , à l'aide d'une statistique dite "moyenne des carrés intragroupe", ou plus simplement "carré moyen intragroupe". Cette statistique se calcule en utilisant la formule suivante

$$CM_{\text{intra}} = \frac{SC_{\text{intra}}}{n-k}$$

avec SC_{intra} le carré moyen intragroupe. Ici $n-k$ est le degré de liberté de CM_{intra}, et

$$SC_{\text{intra}} = \sum_{j=1}^{k} \sum_{i=1}^{r} (Y_{ij} - \bar{Y}_j)^2$$

où Y_{ij} est le score du sujet i $(i = 1, \ldots, 5)$ dans le échantillon j $(j = 1, \ldots, 3)$.

Pour nos données (voir tableau), la valeur observée de la somme des carrés intragroupe pour le pays 1 est

$$(30-40)^2 + (35-40)^2 + (40-40)^2 + (45-40)^2 + (50-40)^2 = 250,$$

pour le pays 2 c'est

$$(40-50)^2+(45-50)^2+(50-50)^2+(55-50)^2+(60-50)^2=250,$$

et pour le pays 3

$$(50-60)^2+(55-60)^2+(60-60)^2+(65-60)^2+(70-60)^2=250.$$

Ce qui donne une valeur observée du carré moyen intragroupe

$$cm_{\text{intra}}=(250+250+250)/(15-3)=750/12=62,5.$$

Remarque 7 : nous avons aussi la décomposition de la somme des carrés totale

$$SC_{\text{totale}}=\sum_{i=1}^{r}\sum_{j=1}^{k}\left(Y_{ij}-\overline{Y}\right)^2=SC_{\text{inter}}+SC_{\text{intra}}. \tag{1}$$

On appelle (1) la "relation fondamentale" de l'ANOVA.

Remarque 8 : La relation fondamentale de l'ANOVA ne s'applique pas aux variabilités. C'est à dire,

$$CM_{\text{totale}}\neq CM_{\text{inter}}+CM_{\text{intra}}.$$

2.2 Test

— **Hypothèses et niveau du test**

L'hypothèse nulle suppose toujours l'égalité des moyennes des k populations, (les échantillons proviennent tous d'une population unique $\mathcal{P}$). Plus précisément, on suppose que les k moyennes sont égales a une même moyenne μ.

$$\text{TEST}:\begin{cases}H_0:\mu_1=\mu_2=\ldots=\mu_k=\mu\\ H_1:\mu_l\neq\mu_j \text{ pour au moins un couple } (l,j)\\ \text{Niveau}:\alpha\end{cases}$$

Remarque 9 : l'hypothèse nulle H_0 correspond à l'absence d'influence du facteur sur la VD ($a_j=0, j=1,\ldots,k$). Alors sous H_0

$$Y_{ij}=\mu+\varepsilon_{ij}.$$

Par contre sous H_1

$$Y_{ij}=\mu+a_j+\varepsilon_{ij},\qquad i=1,\ldots,r,\qquad j=1,\ldots,k,$$

avec μ la moyenne globale de la variable dépendante et a_j l'effet de la modalité j du facteur sur la VD.

Dans l'Exemple 1 (b) nous avons 3 populations. Nous écrivons

$$\text{TEST}:\begin{cases}H_0:\mu_1=\mu_2=\mu_3=\mu\\ H_1:\mu_l\neq\mu_j \text{ pour au moins un couple } (l,j)\\ \text{Niveau}:\alpha=5\%\end{cases}$$

— **Statistique du test**

La statistique de test, notée F, est définie par le rapport entre le carré moyen intergroupe, CM_{inter}, et le carré moyen intragroupe, CM_{intra}

$$F = \frac{CM_{\text{inter}}}{CM_{\text{intra}}}.$$

Sous H_0 on peut montrer que la statistique F suit la loi de Fisher à $(k-1, n-k)$ degrés de liberté, que l'on note $F(k-1, n-k)$.

Remarque 10 : sous l'hypothèse d'égalité des moyennes de groupes, i.e sous H_0, à la fois la variance intra-groupe ($CM_{\text{intra}} = SC_{\text{intra}}/(n-k)$) et la variance inter-groupe ($CM_{\text{inter}} = SC_{\text{inter}}/(k-1)$) sont des estimateurs sans bias de σ^2. En revanche sous H_1, seule CM_{intra} est un estimateur de σ^2.

Soit f_{obs} la valeur observée de la statistique F.

Dans l'Exemple 1 (b), la valeur observée de F est $f_{\text{obs}} = 500/62,5 = 8$.
La valeur de la statistique obtenue indique que la variabilité intergroupe est 8 fois plus grande que la variabilité intragroupe.

— **Critère de décision**

On définit le critère de décision à l'aide de la p-valeur

$$\alpha_{\text{obs}} = P_{H_0}(F \geq f_{\text{obs}}).$$

Au risque α, on rejette H_0 si $\alpha_{\text{obs}} < \alpha$.

Soit $\alpha = 5\%$. Dans notre exemple le logiciel STATISTICA nous donne une p-valeur de $0,006$. Alors,

$$\alpha_{\text{obs}} = P_{H_0}(F \geq 8) = 0,006.$$

Comme $\alpha_{\text{obs}} < 5\%$, on rejette H_0 au risque $\alpha = 5\%$. Au risque d'erreur de 5% il est peu probable d'obtenir une telle variabilité entre les élèves des différents pays si la performance en logique dans le pays est en réalité la même. Les trois moyennes sont globalement différentes au risque $\alpha = 5\%$.

Les valeurs obtenues par STATISTICA sont résumées dans le tableau ANOVA ci-après.

Remarque 11 : Le rejet de l'égalité des moyennes ne permet pas de savoir quelles sont les moyennes significativement différentes. Pour cela, la méthode des contrastes ou méthode de Scheffé associe à l'analyse de variances permet de répondre à cette question.

— **La statistique R^2**

La statistique R^2, connue sous le nom de *rapport de corrélation* est définie par le rapport entre la variabilité intergroupe et la variabilité totale. Plus précisément :

$$R^2 = \frac{SC_{\text{inter}}}{SC_{\text{total}}}.$$

Cette statistique prend des valeurs entre 0 et 1.

Dans l'Exemple 1, la valeur observée de R^2 est $r^2 = 1000/(1000 + 750) = 0,57$. Le modèle de l'ANOVA explique 57% de la variation totale.

Ci-dessous les résultats fournis par STATISTICA.

	Test de SC Modèle Complet vs. SC Résidus (PAYS dans Classeur5)										
Dépendnt Variable	**Multiple R**	Multiple R²	Ajusté R²	SC Modèle	dl Modèle	MC Modèle	SC Résidus	dl Résidus	MC Résidus	F	p
Performance	0,755929	0,571429	0,500000	1000,000	2	500,0000	750,0000	12	62,50000	8,000000	0,006196

Figure 3 : Résultats du test données par STATISTICA

2.3 Validation du modèle

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

1. **Homoscédasticité** . En pratique on trace le graphe des résidus e_{ij} en fonction des valeurs prédites $\hat{Y}_{ij}$. Ce graphique doit présenter des points répartis de manière homogène. Si une structure (tendance) apparaît la homocédasticité ne se vérifie pas.

2. **Absence de données influentes**. Il s'agit de vérifier que les résidus standardisés $e_{ij}/\sqrt{CM_{\text{intra}}}$ 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].

3. **Normalité des résidus**. Tests de Shapiro-Wilk, le test de Lillieforset et le test de Kolmogorov-Smirnov. Il est important de remarquer que en pratique on regarde aussi la normalité à l'aide d'un graphique comparant les quantiles des résidus estimés aux quantiles sous l'hypothèse de normalité. Ce type de graphique est appelé droite de Henry. Nous verrons ce type de graphique en TP avec Statistica dans la séance de Anova et Régression linéaire.