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

Document - Chapitre 4 : ANOVA à un facteur

On cherche un lien entre une variable quantitative (variable dépendante : VD) et une variable qualitative (variable indépendante : VI) à k modalités. On utilise l'analyse de la variance (ANOVA) à un facteur quand on dispose : d'une VD Y, d'une VI X à k modalités (facteur à k niveaux) et de k échantillons indépendants $(E_1, \ldots, E_k)$ de taille $n_1, \ldots, n_k$, respectivement. Nous supposons ici que les k échantillons E_j ont tous la même taille $(n_1 = n_2 = \ldots = n_k)$.

Soit r la taille commune des échantillons, Y_j $(j = 1, \ldots, k)$ la variable correspondant aux valeurs de la VD Y observées sur l'echantillon E_j et $\bar{Y}_j$ la moyenne de Y_j. 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}	...	y_{1k}
2	y_{21}	y_{22}	...	y_{2k}
3	y_{31}	y_{32}	...	y_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	y_{r1}	y_{r2}	...	y_{rk}
Moyenne	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_k$

Conditions d'application :

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).

Modèle pour l'ANOVA à un facteur :

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

ou

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

Ici ε_{ij} est une variable normale de moyenne nulle et variance σ^2, μ est la moyenne globale ou effet moyen de la VD et a_j l'effet principal du niveau j du facteur sur la VD.

Exemple : Etude de la réussite scolaire, pour d'élèves de troisième de différentes Pays : Pays 1, Pays 2, Pays 3. Nous avons trois échantillons d'élèves de troisième. 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. *Y a-t-il une influence du Pays sur la performance à l'examen de logique ?*

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

Dans cet exemple on suppose que les 2 conditions (en haut) sont vérifiées.

- **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.

Pour répondre à la questions nous formulons les hypothèses :

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.

Dans cet example, $k = 3$ et $r = 5$. Le nombre total d'observations est $n = r \times k = 3 \times 5 = 15$.

- **Moyennes des k échantillons ($\bar{Y}_j$) et moyenne de moyennes ($\bar{Y}$).**

$$\bar{Y}_j = \frac{1}{r}\sum_{i=1}^{r} Y_{ij}; \qquad \bar{Y} = \frac{1}{k}\sum_{j=1}^{k} \bar{Y}_j = \frac{1}{k \times r}\sum_{j=1}^{k}\sum_{j=1}^{r} Y_{ij}$$

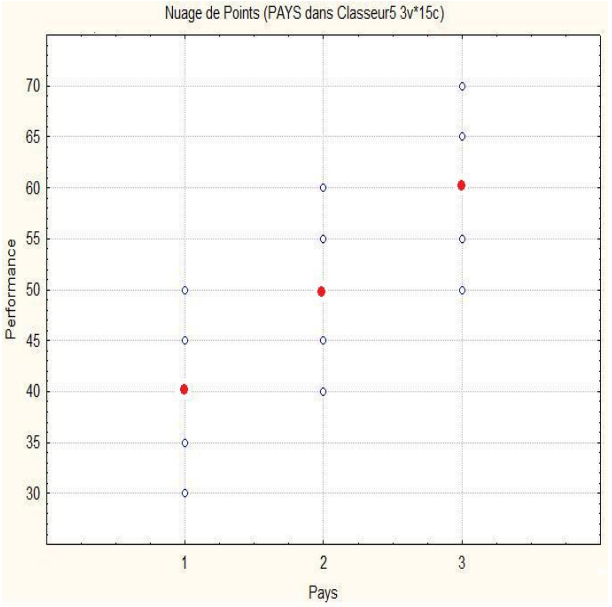

Fig. Valeurs de la VD (Performance) en fonction de la VI (Pays)

Dans cet exemple

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

et La valeur observée de la moyenne globale (moyenne des moyennes) est

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

— **Variabilité intergroupe (variabilité entre les différents groupes).**

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

Pour les données du tableau, la valeur observée du carré moyen intergroupe est

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

— **Variabilité intragroupe (variabilité à l'intérieur de chaque groupe).**

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

Pour nos données

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

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 la statistique F suit la loi de Fisher à $(k-1, n-k)$ degrés de liberté, que l'on note $F(k-1, n-k)$. Soit f_{obs} la valeur observée de la statistique F.

Dans l'exemple, la valeur observée de F est $f_{\text{obs}} = 500/62,5 = 8$.

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 $\alpha = 5\%$.

La statistique R^2 (*rapport de corrélation*)

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

Dans l'exemple, la valeur observée de R^2 est $r^2 = 1000/(1000 + 750) = 0,57$.

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

0.1 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.