Université de Picardie Jules Verne **2017-2018**
UFR des Sciences

Licence mention Mathématiques - Semestre 3
Statistique

Analyse de la variance

Dans ce qui suit, on considère un entier $k \geq 2$.

L'analyse de la variance permet de comparer les moyennes de plusieurs populations à partir d'échantillons indépendants, afin de tester l'influence d'un ou de plusieurs facteurs.

En toute rigueur, elle n'est valable que pour des échantillons extraits de populations gaussiennes de même variance. En général, le non-respect de ces conditions n'a pas trop d'influence sur la validité du test (on parle de méthode robuste). L'erreur introduite est cependant d'autant plus forte que les effectifs des échantillons sont faibles et inégaux.

1. Comparaison de k variances : test de Bartlett

Dans k populations P_1, P_2, ..., P_k on étudie le même caractère. Soient X_1, X_2, ..., X_k des variables aléatoires représentant le caractère dans chaque population, de moyennes respectives μ_1, μ_2, ..., μ_k, d'écart-types respectifs σ_1, σ_2, ..., σ_k. Pour tout $i = 1,\dots,k$, on extrait de P_i un échantillon $E_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n_i})$ de taille n_i de X_i ; les moyennes d'échantillon sont alors $\overline{X_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}$, et les variances corrigées d'échantillon $S_{c,i}^2 = \frac{n_i}{n_i - 1} S_i^2$ avec $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}^2 - \overline{X_i}^2$.

On suppose que les échantillons E_i sont indépendants et que les X_i suivent les lois normales $\mathcal{N}(\mu_i;\sigma_i)$.

On pose $n = \sum_{i=1}^{k} n_i$ et $S_R^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)S_{c,i}^2$ la *variance résiduelle (ou intragroupe)* des échantillons.

On pose $B = \frac{1}{\lambda}\left((n-k)\ln S_R^2 - \sum_{i=1}^{k}(n_i - 1)\ln S_{c,i}^2\right)$, avec $\lambda = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right)$.

Test de $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$ contre H_1 : les variances ne sont pas égales.

Sous l'hypothèse H_0, B suit approximativement une loi du khi-deux à $k - 1$ degrés de liberté.

On calcule $s_R^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_{c,i}^2$, estimation de σ^2 sous l'hypothèse H_0, et $b = \frac{1}{\lambda}\left((n-k)\ln s_R^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_{c,i}^2\right)$. On détermine le réel b_α tel que $P(B \geq b_\alpha) = \alpha$ (table 4), et on décide que :

- si $b < b_\alpha$, alors on ne peut rejeter H_0 ;
- si $b \geq b_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

2. Comparaison de k moyennes : analyse de la variance à un facteur

On reprend la situation du paragraphe 1.

Les échantillons E_i sont supposés indépendants. On suppose de plus que les X_i suivent les lois normales $\mathcal{N}(\mu_i;\sigma_i)$ et que $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$. Cette dernière hypothèse peut être validée en effectuant le test d'égalité des variances décrit ci-dessus (paragraphe 1.).

En général, les k populations correspondent aux k modalités d'un facteur controlé (par exemple k groupes de malades, chaque groupe recevant un traitement différent).

On désigne par $\overline{X}$ et S_T^2 la moyenne et la variance corrigée de la réunion des k échantillons. On a alors $n = \sum_{i=1}^{k} n_i$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n}\sum_{i=1}^{k} n_i\overline{X_i}$ et $S_T^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \overline{X})^2$.

On définit la variance résiduelle (ou intragroupe) $S_R^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)S_{c,i}^2$, qui caractérise la dispersion des valeurs à l'intérieur des échantillons, et la variance factorielle (ou intergroupe) $S_F^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{X_i}-\overline{X})^2$, qui caractérise la dispersion des valeurs d'un échantillon à l'autre, i.e. la variation due à l'influence du facteur étudié. On a alors $(n-1)S_T^2 = (n-k)S_R^2 + (k-1)S_F^2$. Ainsi, S_T^2 est une moyenne pondérée de S_R^2 et S_F^2.

Test de $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ **contre** H_1 : **les moyennes ne sont pas égales**.

Sous l'hypothèse H_0, $F = \frac{S_F^2}{S_R^2}$ suit la loi de Snédécor à $(k-1, n-k)$ degrés de liberté. On calcule $s_R^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i-1)s_{c,i}^2$, estimation de σ^2 d'après l'hypothèse d'égalité des variances, et s_F^2 , estimation de σ^2 sous l'hypothèse H_0 ; on peut utiliser la relation $(n-1)s_T^2 = (n-k)s_R^2 + (k-1)s_F^2$. On calcule alors $f = \frac{s_F^2}{s_R^2}$. On détermine f_α tel que $P(F \geq f_\alpha) = \alpha$ (tables 5 et 6), et on décide que :

- si $f < f_\alpha$, alors on ne peut rejeter H_0 ;
- si $f \geq f_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper, i.e. que l'on attribue une influence significative au facteur étudié.

3. Analyse de la variance à deux facteurs

On considère un facteur A à r modalités et un facteur B à s modalités ; ces deux facteurs déterminent $k = rs$ populations $P_{i,j}$, $1 \leq i \leq r$ et $1 \leq j \leq s$.

Dans les k populations $P_{i,j}$ on étudie le même caractère. Soit $X_{i,j}$ la variable aléatoire représentant le caractère dans la population $P_{i,j}$, de moyennes respectives $\mu_{i,j}$ et d'écart-type respectifs $\sigma_{i,j}$.

De chaque population $P_{i,j}$ on extrait un échantillon $E_{i,j} = (X_{i,j,1}, X_{i,j,2}, \ldots, X_{i,j,n_i})$ de taille $n_{i,j}$; la moyenne d'échantillon est alors $\overline{X_{i,j}} = \frac{1}{n_{i,j}}\sum_{k=1}^{n_{i,j}} X_{i,j,k}$, et la variance corrigée d'échantillon $S_{c,i,j}^2 = \frac{n_{i,j}}{n_{i,j}-1}S_{i,j}^2$ avec $S_{i,j}^2 = \frac{1}{n_{i,j}}\sum_{k=1}^{n_{i,j}} X_{i,j,k}^2 - \overline{X_{i,j}}^2$.

Les échantillons $E_{i,j}$ sont supposés **indépendants et de même taille** n ($n_{i,j} = n$). On suppose de plus que les $X_{i,j}$ suivent les lois normales $\mathcal{N}(\mu_{i,j}; \sigma_{i,j})$, les variances $\sigma_{i,j}^2$ étant égales à σ^2.

L'analyse de la variance à deux facteurs permet de comparer les moyennes des $k = rs$ échantillons et de tester l'influence du facteur A seul, l'influence du facteur B seul et l'influence de l'interaction des deux facteurs (il y a interaction lorsque l'influence d'un facteur sur la moyenne des populations est différente en l'absence ou en la présence de l'autre facteur). Il y aura donc trois tests d'égalité des moyennes.

On désigne par $\overline{X}$ et S_T^2 la moyenne et la variance corrigée de la réunion des $k = rs$ échantillons (réunion de taille nrs). On a alors $\overline{X} = \frac{1}{nrs}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{i,j}} X_{i,j,k} = \frac{1}{nrs}\sum_{i=1}^{r}\sum_{j=1}^{s} n\overline{X_{i,j}}$ et $S_T^2 = \frac{1}{nrs-1}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{i,j}}(X_{i,j,k}-\overline{X})^2$

De façon analogue à l'analyse de la variance à un facteur, on définit la variance résiduelle $S_R^2 = \frac{1}{(n-1)rs}\sum_{i=1}^{r}\sum_{j=1}^{s}(n-1)S_{c,i,j}^2$, qui caractérise la dispersion des valeurs à l'intérieur des échantillons, et la variance factorielle $S_F^2 = \frac{1}{rs-1}\sum_{i=1}^{r}\sum_{j=1}^{s} n(\overline{X_{i,j}}-\overline{X})^2$, qui caractérise la dispersion des valeurs d'un échantillon à l'autre, i.e. la variation due à l'influence des facteurs étudiés. On a alors $(nrs-1)S_T^2 = (n-1)rsS_R^2 + (rs-1)S_F^2$. Ainsi, S_T^2 est une moyenne pondérée de S_R^2 et S_F^2.

Pour étudier l'influence de chacun des deux facteurs et de leur interaction, on définit :

- $\overline{X_{i,\bullet}} = \frac{1}{s}\sum_{j=1}^{s}\overline{X_{i,j}}$, moyenne conditionnelle à la $i^{ème}$ modalité du facteur A ;
- $\overline{X_{\bullet,j}} = \frac{1}{r}\sum_{i=1}^{r}\overline{X_{i,j}}$, moyenne conditionnelle à la $j^{ème}$ modalité du facteur B ;
- $S_A^2 = \frac{sn}{r-1}\sum_{i=1}^{r}(\overline{X_{i,\bullet}} - \overline{X})^2$, variance factorielle due au facteur A seul ;
- $S_B^2 = \frac{rn}{s-1}\sum_{j=1}^{s}(\overline{X_{\bullet,j}} - \overline{X})^2$, variance factorielle due au facteur B seul ;
- $S_{AB}^2 = \frac{n}{(r-1)(s-1)}\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{X_{i,j}} - \overline{X_{i,\bullet}} - \overline{X_{\bullet,j}} - \overline{X})^2$, variance factorielle due à l'interaction de A et B.

La variance factorielle S_F^2 se decompose alors suivant la formule $(rs-1)S_F^2 = (r-1)S_A^2 + (s-1)S_B^2 + (r-1)(s-1)S_{AB}^2$. Ainsi, S_F^2 est une moyenne pondérée de S_A^2, S_B^2 et S_{AB}^2.

Test de $H_{0,A}$: le facteur A n'a pas d'influence sur la moyenne des populations contre $H_1 = \overline{H_{0,A}}$.

Sous l'hypothèse $H_{0,A}$, $F_A = \frac{S_A^2}{S_R^2}$ suit la loi de Snédécor à $(r-1, (n-1)rs)$ degrés de liberté.

On calcule s_R^2 et s_A^2, d'où $f_A = \frac{s_A^2}{s_R^2}$. On détermine f_α tel que $P(F_A \geq f_\alpha) = \alpha$ (tables 5 et 6),. et on décide que :

- si $f_A < f_\alpha$, alors on ne peut rejeter H_0 ;
- si $f_A \geq f_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper, i.e. que l'on attribue une influence significatice au facteur A.

Test de $H_{0,B}$: le facteur B n'a pas d'influence sur la moyenne des populations contre $H_1 = \overline{H_{0,B}}$.

Sous l'hypothèse $H_{0,B}$, $F_B = \frac{S_B^2}{S_R^2}$ suit la loi de Snédécor à $(s-1, (n-1)rs)$ degrés de liberté. On procède alors de façon analogue au test ci-dessus.

Test de $H_{0,AB}$: il n'y a pas d'interaction entre les facteurs A et B contre $H_1 : \overline{H_{0,AB}}$.

Sous l'hypothèse $H_{0,AB}$, $F_{AB} = \frac{S_{AB}^2}{S_R^2}$ suit la loi de Snédécor à $((r-1)(s-1), (n-1)rs)$ degrés de liberté. On procède alors de façon analogue au test ci-dessus.

Cas particulier : échantillons d'une seule observation

Si chaque échantillon ne comporte qu'une seule observation, i.e. $n = 1$, alors les $S_{i,j}^2$ et S_R^2 sont nulles et les quotients précédents F_A, F_B et F_{AB} ne sont plus définis.

On a alors $(rs-1)S_T^2 = (r-1)S_A^2 + (s-1)S_B^2 + (r-1)(s-1)S_{AB}^2$.

Test de $H_{0,A}$ contre $H_1 = \overline{H_{0,A}}$.

Sous l'hypothèse $H_{0,A}$, $F_A = \frac{S_A^2}{S_{AB}^2}$ suit la loi de Snédécor à $(r-1, (r-1)(s-1))$ degrés de liberté. On procède alors de façon analogue au test ci-dessus.

Test de $H_{0,B}$ contre $H_1 = \overline{H_{0,B}}$.

Sous l'hypothèse $H_{0,B}$, $F_B = \frac{S_B^2}{S_{AB}^2}$ suit la loi de Snédécor à $(s-1, (r-1)(s-1))$ degrés de liberté. On procède alors de façon analogue au test ci-dessus.

On ne peut pas tester $H_{0,AB}$ contre $H_1 : \overline{H_{0,AB}}$.

4. Exercices

Sauf mention explicite, les tests seront réalisés au risque 5%.

Exercice 1.

On a étudié la durée de développement d'un parasite à l'intérieur d'un organisme hôte en fonction de la température d'élevage. Les résultats obtenus sont les suivants :

		durée de développement (en jours)	
température (en °C)	nombre d'animaux	moyennes	écart-type corrigé
16	32	81	6,8
20	33	52	5,2
23	31	46	6,7

La température a-t-elle une influence sur la durée de développement du parasite ? On précisera les hypothèses à faire pour pouvoir appliquer la technique d'analyse de la variance à la résolution du problème posé.

Exercice 2.

Lors d'une expérience pédagogique, on s'intéresse a l'effet comparé de deux pédagogies des mathématiques chez deux groupes de 10 sujets : pédagogie traditionnelle (p1) et pédagogie moderne (p2). On note la performance à une epreuve de statistique :

- p1 : 5.0 , 4.0 , 1.5 , 6.0 , 3.0 , 3.5 , 3.0 , 2.5 , 1.5 , 2.5 ;
- p2 : 4.0 , 5.5 , 4.5 , 6.5 , 4.5 , 5.5 , 1.0 , 2.0 , 4.5 , 4.5.

1) Vérifier que les paramètres des deux échantillons sont donnés par :

	p1	p2
moyenne	3.250	4.250
écart-type corrigé	1.439	1.637
variance corrigée	2.069	2.681

2) Ces données expérimentales permettent-elles d'affirmer que la pédagogie a un effet sur les résultats à l'épreuve de statistique ?

a) Comparer les moyennes à l'aide d'une analyse de variance.

b) Comparer les résultats avec ceux obtenus par un test utilisant la statistique *T*.

Exercice 3.

On veut savoir si l'addition de substances adjuvantes à un vaccin modifie la production d'anticorps. Pour cela, on mesure les quantités d'anticorps produites par des sujets après admininistration de quantités égales de vaccin, additioné ou non d'une substance adjuvante. On a obtenu les taux suivants :

- sans substance adjuvante : 1, 3, 3, 0, 1 ;
- avec de l'alumine : 2, 4, 5, 4, 3, 6 ;
- avec des sels de calcium : 3, 3, 4, 5 ;
- avec des phosphates : 1, 4, 2, 3, 3.

1) Quelle(s) hypothèse(s) faut-il faire pour pouvoir appliquer la technique d'analyse de la variance à la résolution du problème posé ? La validité de ces hypothèses est-elle importante dans le cas présent ?

2) En supposant ces hypothèses satisfaites, l'efficacité du vaccin dépend-elle :

a) de la présence de substances adjuvantes ?

b) de leur nature ?

Exercice 4.

Dans une expérience, on présente à chaque sujet soit oralement soit par écrit des mots qui sont soit familiers, soit non familiers. Après une période d'attente, on interroge le sujet et on calcule le nombre de syllabes non significatives mémorisées. 24 sujets ont participé, répartis en 4 groupes de 6.

Oral	Familier	19	16	18	23	14	16
Oral	Non familier	15	13	7	9	8	11
Ecrit	Familier	10	12	18	16	17	14
Ecrit	Non familier	9	16	14	11	12	8

1) Tester l'hypothèse selon laquelle le mode de présentation (oral ou écrit) n'a pas d'effet sur la mémorisation.

2) Tester l'hypothèse selon laquelle la nature des mots présentés (familier ou non) n'a pas d'effet sur la mémorisation.

3) Tester l'hypothèse selon laquelle il n'y a pas d'effet d'interaction sur la mémorisation.

Exercice 5.

Pour tester la fiabilité de 4 laboratoires d'analyse, on utilise 4 solutions ayant le même titre de glucose dans du sérum physiologique additionné de quantités variables de galactose. Chaque laboratoire reçoit un échantillon de chaque solution et fournit le résultat de ses mesures. L'ensemble des résultats, exprimés en grammes de glucose par litre de solution, est regroupé dans le tableau suivant :

Solution \ Laboratoire	L1	L2	L3	L4
S1	1.05	1.15	1.08	1.13
S2	1.12	1.15	1.11	1.09
S3	1.02	1.10	1.04	1.05
S4	1.09	1.11	1.07	1.10

1) Au risque 5 %, peut-on considérer que le choix du laboratoire a une influence sur la mesure du taux de glucose ?

2) Au risque 5 %, peut-on considérer que la quantité de galactose présente dans la solution a une influence sur la mesure du taux de glucose ?