Chapitre 1

Tests d'ajustement et de comparaison

Nous présentons dans cette section les tests de Kolmogorov-Smirnov, Cramér-von Mises

1.1 Le test d'ajustement de de Kolmogorov

Le test de Kolmogorov-Smirnov est un test non paramétriques d'ajustement ou (d'adéquation), qui s'étend à la comparaison de deux fonctions de répartition empiriques, et permet alors de tester l'hypothèse que deux échantillons sont issus de la même loi.

On l'utilise de préférence au test d'adéquation du chi-deux lorsque le caractère observé peut prendre des valeurs continues.

Ce test non paramétrique effectue une comparaison entre la fonction de répartition empirique $F_n(x)$ et la fonction de répartition théorique $F(x) = P(X < x)$ de la loi de probabilités considérée.

La fonction de répartition empirique des valeurs x_i d'un n-échantillon étant définie par :

$$F_n(x) = \begin{cases} 0 \text{ si } x < x_1 \\ \frac{i-1}{n} \text{ si } x_{i-1} < x < x_i \\ 1 \text{ si } x \geq x_n, \end{cases}$$

1.1.1 Variable de décision

Kolmogorov propose d'utiliser la statistique de test suivante :

$$D_n = D(F_n, F) = \sup |F_n(x) - F(x)| .$$

La loi de cette statistique de test est données dans la table de Kolmogorov.

D'après le théorème de Kolmogorov, sous l'hypothèse nulle H_0 d'identité des deux distributions

$$\begin{cases} H_0 : F(x) = F_n(x) \\ H_1 : F(x) \neq F_n(x) \end{cases}$$

la statistique D_n suit asymptotiquement la distribution de probabilités (distribution de Kolmogorov Smirnov) définie par

$$P\left(\sup |F_n(x) - F(x)| < \frac{y}{\sqrt{n}}\right) \rightsquigarrow k(y) = \sum (-1)^k \exp\left\{\left(-2k^2y^2\right)\right\}.$$

La région critique étant définie par $D_n > d_n$.

– Au seuil $\alpha = 0.05$ et si $n > 80$, la région critique est $D_n > \dfrac{1.3581}{\sqrt{n}}$ pour $\alpha = 0.01$ $D_n > \dfrac{1.6276}{\sqrt{n}}$.

– Si $n < 80$ on se reportera alors à la table test de **Kolmogorov-Smirnov**

– Une valeur élevée de D_n est une indication que la distribution de l'échantillon s'éloigne sensiblement de la distribution de référence $F(x)$, et qu'il est donc peu probable que H_0 soit correcte.

Proposition 1 *Soit $X, ..., X_n$ un échantillon de fonction de répartition F continue, si F_0 est une fonction de répartition continue, alors :*

$$D_n = \max\left\{\left|\frac{i}{n} - F_0\left(X_{(i)}\right)\right|, \left|F_0\left(X_{(i)}\right) - \frac{i-1}{n}\right|\right\}.$$

Exemple 2 *Une nouvelle clientèle étrangère est attendue dans une station balnéaire. Afin de mieux connaître leurs goûts, des brasseurs ont commandé une étude de marché. En début de saison, on demande à vingt de ces nouveaux touristes de donner leur préférence parmi cinq types de bières, de la moins amère (bière 1) à la plus amère (bière 5). A l'aide d'un test de K-S, le chargé d'études décide de comparer les résultats avec une loi uniforme, c'est-à-dire une situation où chaque bière aurait eu la préférence de quatre répondants. Les résultats de l'enquête sont les suivants :*

1 3 2 5 1 2 2 4 1 2 2 1 3 3 2 4 5 1 1 2

On se fixe un risque d'erreur de 5%. L'hypothèse H_0 à tester est celle de l'égalité avec une loi uniforme. Résumons les écarts entre observations et répartition uniforme

Classe	*Effectif*	*Uniforme*	f_i	F_n *(Empirique)*	F *(Théorique)*	D
1	6	4	$6/20 = 0.30$	$0,30$	0.20	0.10
2	7	4	$7/20 = 0.35$	$13/20 = 0,65$	0.40	0.25
3	3	4	$3/20 = 0.15$	$16/20 = 0,80$	0.60	0.20
4	2	4	$2/20 = 0.1$	$18/20 = 0,90$	0.80	0.10
5	2	4	$2/20 = 0.1$	1	1	0.00
Total	20	20	1	//////////	///////	/////

La distance la plus élevée s'établit à $D_n = 0,25$.

On calcule pour $n = 20$ et $\alpha = 5\%$ la valeur de $d_n = 0,294$. Bien que ces touristes semblent préférer les bières les moins amères, on ne peut pas rejeter l'hypothèse selon laquelle ils n'ont pas de préférence particulière.

1.2 Test d'identité de deux distributions

On peut généraliser le test de Kolmogorov au cas de deux échantillons a fin de comparer leurs distributions.

1.2.1 Test de Kolmogorov-Smirnov

L'hypothèse nulle est que les deux échantillons proviennent de la même distribution ; l'alternative est qu'ils proviennent de distributions ayant des répartitions différentes.

On ne spécifie aucune forme particulière pour leur différence. Et la statistique de test est basée sur un écart en valeur absolue entre la fonctions de répartition empiriques des deux suites d'observations.

Exemple 3 *Un psychologue fait passer un test de rapidité à des enfants normaux et d'autres considérés comme mentalement retardés. Les temps qu'ils mettent pour accomplir une série de tâches sont les suivants*

Enfants normaux	183	191	197	204	218	227	233	
Enfants retardés	202	220	228	239	242	243	261	270

On pose les hypothèses de test

$$\begin{cases} H_0 : F_{EN} = F_{ER} \\ \quad contre \\ H_1 : F_{ER} \neq F_{ER} \end{cases}$$

On cherche à obtenir une estimation de la fonction de répartition à partir de l'échantillon observé afin de la comparer ensuite à la fonction de répartition de la loi théorique.

Pour cela, on commence par trier par ordre croissant les valeurs x_i de l'échantillon. On les appelle traditionnellement des statistiques d'ordre. La fonction de répartition empirique est définie par :

$$F_n(x) = \begin{cases} 0 \ si \ x < x_1 \\ \frac{i}{n} \ si \ x_i < x < x_{i+1} \\ 1 \ si \ x \geq x_n, \end{cases}$$

On estime donc $F(x) = P(X \leq x)$ au moyen de la proportion $F_n(x)$ d'éléments de l'échantillon qui sont inférieurs ou égaux à x.

Donc on estime les fonctions de répartition des deux groupes

Obs	$\hat{F}_{EN}$	$\hat{F}_{ER}$	*Ecart*
183	1/7	0	$1/7 = 8/56$
191	2/7	0	$2/7 = 16/56$
197	3/7	0	$3/7 = 24/56$
202	3/7	1/8	17/56
204	4/7	1/8	25/56
218	5/7	1/8	33/56
220	5/7	2/8	26/56
227	6/7	2/8	34/56
228	6/7	3/8	27/56
233	1	3/8	40/56
239	1	4/8	
242	1	5/8	
243	1	6/8	
261	1	7/8	
270	1	1	

L'écart en valeur absolue le plus grand

$$D_n = \sup \left| \hat{F}_{EN} - \hat{F}_{ER} \right| = |1 - 3/8| = 0.62.$$

La loi du supremum des écarts en valeur absolue est tabulée dans la table de Smirnov. On rejette H_0 si $D_n > 0,71$. Donc ici, on ne peut pas rejeter l'hypothèse selon laquelle les distributions sont différentes pour les deux groupes d'enfants.

Exemple 4 *Les notes de mathématiques d'un groupe d'étudiants lors d'un examen sont réparties selon le tableau :*

Notes x	*Nombre d'étudiants* n_i	*Notes* x	*Nombre d'étudiants* n_i
$2 \leq x < 4$	2	$10 \leq x < 12$	16
$4 \leq x < 6$	5	$12 \leq x < 14$	12
$6 \leq x < 8$	8	$14 \leq x < 16$	9
$8 \leq x < 10$	14	$16 \leq x < 18$	4

1− *Par quelle loi de probabilité on peut procéder à un ajustement de la distribution empirique proposée. trouver l'estimation des paramètres de cette loi.*

2− *Vérifier l'ajustement par un test du χ^2 à un seuil de signification $\alpha = 0.05$.*

3− *Vérifier l'ajustement par le test de Kolmogorov à un à un seuil de signification $\alpha = 0.05$.*

On a $\bar{x} = 10.69$, $\acute{S}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = (3.45)^2 = 11.90$

Notes	*Centre de classe* x_i	n_i	f_i
$2 \leq x < 4$	3	2	0.0286
$4 \leq x < 6$	5	5	0.071
$6 \leq x < 8$	7	8	0.114
$8 \leq x < 10$	9	14	0.2
$10 \leq x < 12$	11	16	0.229
$12 \leq x < 14$	13	12	0.171
$14 \leq x < 16$	15	9	0.129
$16 \leq x < 18$	17	4	0.057
Total		70	1

Les probabilités $P(a \leq X \leq b)$ *sont calculées au moyen de la table normale.*

Notes	*Effectifs* n_i	$P(a \leq X \leq b)$	$n_i p_i$	$n_i - n_i p_i$	$\frac{(n_i - n_i p_i)^2}{n_i p_i}$
$2 \leq x < 4$	2	0.0203	1.421	0.579	0.2359
$4 \leq x < 6$	5	0.0607	4.249	0.751	0.1327
$6 \leq x < 8$	8	0.1337	9.359	$-1,359$	0.1973
$8 \leq x < 10$	14	0.2001	14.007	-0.007	0.000003
$10 \leq x < 12$	16	0.2273	15.911	0.089	0.00049
$12 \leq x < 14$	12	0.1835	12.845	-0.845	0.0556
$14 \leq x < 16$	9	0.1067	7.49	1.531	0.3129
$16 \leq x < 18$	4	0.0448	3.136	0.864	0.238

Les effectifs calculées $n_i p_i$ *doivent être* > 5. *On regroupe les classes*

Notes	*Effectifs* n_i	$P(a \leq X \leq b) = p_i$	$n_i p_i$	$\frac{(n_i - n_i p_i)^2}{n_i p_i}$
$2 \leq x < 6$	7	0.081	5.67	0.3119
$6 \leq x < 8$	8	0.1337	9.359	0.1973
$8 \leq x < 10$	14	0.2001	14.007	0.000003
$10 \leq x < 12$	16	0.2273	15.911	0.00049
$12 \leq x < 14$	12	0.1835	12.845	0.0556
$14 \leq x < 18$	13	0.1515	10.605	0.5409
Total	70	////////	////////	1.1062

La valeur de $\chi^2_{calc} = 1.1062$ *le nombre de degrés de liberté*

$$k = (6 - 1) - 2 = 3$$

car on deux paramètres estimés m *et* σ^2 *par* $\bar{x}$ *et* $\acute{S}^2$

a un seuil $\alpha = 0.05$ $\chi^2_3(\alpha) = 7.81$ *on a* $\chi^2_{calc} < \chi^2_3(\alpha)$ *on accepte l'hypothèse d'ajustement par une loi normale* $\mathcal{N}(10.69, 3.45)$

3− *On vérifie l'ajustement par le test de Kolmogorov*

Notes	$F_n(x)$	$F(x) = P(X \leq x)$	$\lvert F_n(x) - F(x)\rvert$
$2 \leq x < 4$	0.0286	0.0203	0.0083
$4 \leq x < 6$	0.1	0.081	0.019
$6 \leq x < 8$	0.2143	0.2147	0.0004
$8 \leq x < 10$	0.4143	0.4148	0.0005
$10 \leq x < 12$	0.6429	0.0.6421	0.0008
$12 \leq x < 14$	0.8143	0.8256	0.0113
$14 \leq x < 16$	0.9429	0.9323	0.0106
$16 \leq x < 18$	1	0.9771	0.0229

L'ecart type maximum est : $D_n = \max \lvert F_n(x) - F(x)\rvert = 0.0229$

Exemple 5 *La valeur critique* d_n *correspondante est* (*pour* $n = 70$)
$d_n = 0.15975$ d_n *est donnée par la table du test de* ***Kolmogorov-Smirnov***
pou $\alpha = 0.05$ *et* $n = 70$, *on a* $D_n < 0.15975$.
On conclu donc que l'hypothèse de normalité ne doit pas être rejeté

1.3 Le test d'ajustement de Cramer-von Mises

Soient $X_1, ..., X_n$ des v.a. *i.i.d.* de fonction de répartition F inconnue continue et F_0 une fonction de répartition continue donnée. Pour tester l'hypothèse

$$\begin{cases} H_0 : F = F_0 \\ \text{contre} \\ H_1 : F \neq F_0, \end{cases}$$

L'indicateur d'écart de ce test est :

$$I = \int_{-\infty}^{+\infty} \left(\hat{F}_n(t) - F_0(t)\right) dt$$

où $\hat{F}_n$ est la fonction de répartition empirique des $(X_i)_{1 \leq i \leq n}$.

Sa distribution a été tabulée (voir recueil de tables, fonction de repartiion de la statistique de Cramer-Von Mises).

On démontre que

$$I = \frac{1}{12n} + \sum \left(\frac{2i-1}{2n} - F(x_i)\right)^2,$$

si les x_i sont les valeurs ordonnées de l'échantillon $((x_1 < x_2 < ... < x_n)$.

On rejette Ho si I est supeneur à une valeur que la variable I aléatoire à une probabilité α de dépasser.

Au seuil a 0.05 on rejette Ho si $I > 0.46136$ pour n grand.

Chapitre 2

Analyse de variance

2.1 Analyse de variance à un facteur

2.1.1 Exemple introductif

Examinateur	A	B	C
Notes	$10, 11, 11$ $12, 13, 15$	$8, 11, 11, 13$ $14, 15, 16, 16$	$10, 13, 14, 14$ $15, 16, 16$
Effectif	$n_1 = 6$	$n_2 = 8$	$n_3 = 7$
Moyenne	12	13	14

21 candidats, 3 examinateurs (resp $6, 8$ et 7 étudiants)
"**effet d'examinateu**r" ?
ANOVA : pour étudier l'effet des variables qualitatives sur une variable quantitative

- **Facteur** (variable qualitative) : prend un nombre fini de valeurs,

  une **valeur = une classe**. Exemple : facteur **"examinateur"**
- **Niveau** (ou population) : les différentes valeurs prises par un facteur Ex : niveaux A, B, C
- **Test** de l'effet d'un facteur : tester si les moyennes des populations sont égales.
- **La variable** étudiée : X, à valeurs numériques (**note**)

Donc l'analyse de la variance est un ensemble de techniques permettant de comparer plusieurs échantillons de données. Cette comparaison est le plus souvent limitée à celle des moyennes dans un cas gaussien. On l'utilise également pour étudier l'effet d'un facteur qualitatif externe. Nous nous limiterons ici une présentation résumèe dans le cas où il y a un seul facteur explicatif.

Le but principal de l'analyse de la variante (Anova) est de comparer les moyennes de plusieurs populations vérifiant certaines conditions à partir d'échantillons prélevés dans ces populations.

On dispose de k échantillons de tailles respectives $n_1, n_2, ..., n_k$, correspondant chacun à un niveau diftérent d'un facteur A. On pose $n = \sum_{i=1}^{k} n_i$ et on dresse le tableau (2.1)

On suppose que le facteur A influe uniquement sur les moyennes des distributions et non sur leur variance. II s'agit donc d'un test de confusion des k moyennes $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$

Définition 6 *L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant k modalités (groupes) sur les moyennes d'une variable quantitative X.*

Les problèmes concernés par la technique ANOVA 1 s'écrivent en générale de la manière suivante :

Facteur	A_1	A_2	.....	A_j	.....	A_k	
	$x_{1,1}$	$x_{1,2}$	.....	.. $x_{1,j}$...	.....	$x_{1,k}$	
	$x_{2,1}$	$x_{2,2}$	.....	.....	.....	$x_{2,k}$	
	...	...		...			
	$x_{i,1}$	$x_{i,2}$		$x_{i,j}$		$x_{i,k}$	
	...	...	.....	...	.....	...	
	...	...		...		...	
	$x_{n_1,1}$	$x_{n_2,2}$		$x_{n_i,j}$		$x_{n_k,k}$	
Total	$\sum_{i=1}^{n_1} x_{i,1}$	$\sum_{i=1}^{n_2} x_{i,2}$	.....	$\sum_{i=1}^{n_i} x_{i,j}$		$\sum_{i=1}^{n_k} x_{i,k}$	
Moyennes	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_j$		$\bar{x}_k$	
Nombre d'observations	n_1	n_2	.....	n_i		n_k	$\bar{x}$
Vriances	$\bar{S}_1^2$	$\bar{S}_2^2$	.....	$\bar{S}_j^2$		$\bar{S}_k^2$	S^2

(2.1)

2.1.2 Les données de l'analyse

On considère que chaque échantillon est issu d'une v.a. X_i suivant une loi $\mathcal{N}(m_i, \sigma_i)$. En terme de test, nous avons donc

$$\begin{cases} H_0 : m_1 = m_2 = ... = m_k = m \\ H_1 : \exists i, j \ m_i \neq m_j \end{cases} \tag{2.2}$$

et le modèle mathématique leurs associés est donné par

$$x_{ij} = m_i + \epsilon_i^j \text{ avec } i = \overline{1,n},\ j = \overline{1,k}.$$

où x_{ij} est la $j^{ième}$ réalisation de la variable quantitative X dans le $i^{ième}$ échantillon et ϵ_i^j les erreurs de mesure telque $\epsilon_i^j \backsim \mathcal{N}(0, \sigma_i)$, ou encore

$$x_{ij} = \mu + \alpha_i + \epsilon_i^j$$

où μ est une valeur moyenne constante et α_i l'effet du niveau i du facteur explicatif.

Dans le cas où l'hypothèse H_0 est rejetée le problème se posera donc d'estimer m_i (ou μ et les α_i).

- ▶ Si H_0 est vraie alors la variation due au facteur doit être petite par rapport à la variation résiduelle .
- ▶ Par contre, si H_1 est vraie alors la variation due au facteur doit être grande par rapport à la quantité .

2.1.3 Les étapes de l'ANOVA 1

Afin de réaliser le test définie dans (2.2), trois conditions doit être vérifiées préalablement, à savoir :

- Les k échantillons comparés sont indépendants.
- La variable quantitative étudiée suit une loi normale dans les k populations comparées.
- Les k populations comparées ont même variance : Homogénéité des variances ou homoscédasticité.

Si ces dernièeres conditions sont vérifiées alors, on peut utiliser la technique ANOV A 1 pour réaliser le test (2.2), et pour ce faire nous avons besoin des quantités (statistiques) suivantes :

2.1.4 Le test

- n nombre total d'observations.
- On note la moyenne des observations (chaque échantillon) i,

$$\bar{x}j = \frac{1}{n_i}\sum_{i=1}^{n_i} x_{ij} \quad i = 1, ..., k$$

- Variance de chaque échantillon

$$S_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_i)^2, \quad i = 1, ..., k$$

- La moyenne de toutes les observations ($Totale$) est la moyenne des moyennes de chaque échantillon

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{x}_i$$

- La variance totale S^2 est estimée par

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2,$$

  et en remarquant que : $x_{ij} - \bar{x} = x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x}$.

On montre facilement que cette variance totale peut se décomposer en la somme de la variance des moyennes, S_A^2 (aussi appelée variance **inter-classes**) plus la moyenne des variances, S_R^2 (aussi appelée variance **intra-classes**).

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \frac{1}{n}\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

La variance S_A^2 représente la variation du au facteur explicatif A, la variance S_R^2 est elle considérée comme la variabilité résiduelle.

$$S_A^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \quad S_R^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

on a donc

$$S^2 = S_A^2 + S_R^2 \text{ “formule d'analyse de variance”}$$

- La variance totale S^2 est égale à la somme de la variance des moyennes et de la moyenne des variances

Si on écrit $S_R^2 = \frac{1}{n}\sum_{i=1}^{k} n_i S_i^2$, avec $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ en introduisant les dispersions de chaque échantillon, on trouve que $\frac{nS_R^2}{\sigma^2}$ suit une loi du χ^2 à $n-k$ degrés de liberté, car chaque quantité $\frac{n_i S_i^2}{\sigma^2}$ suit une loi du $\chi^2_{n_i-1}$ et

$$\frac{nS_R^2}{\sigma^2} = \frac{\sum_{i=1}^{k} n_i S_i^2}{\sigma^2}$$

Sous l'hypothèse H_0, les v.a. X_i sont de même loi donc on a également le fait que la quantité $\frac{nS^2}{\sigma^2}$ suit une loi du χ^2 à $n-1$ degrés de liberté, et $\frac{nS_A^2}{\sigma^2}$, une loi du χ^2 à $k-1$ degrés de liberté.

On peut donc construire l'indicateur de notre test par

$$\mathbf{F}_{obs} = \mathbf{F}(\mathbf{k-1}, \mathbf{n-k}) = \frac{\frac{\mathbf{S}_A^2}{\mathbf{k-1}}}{\frac{\mathbf{S}_R^2}{\mathbf{n-k}}} \backsim f_\alpha\,(k-1, n-k)$$

dont la loi f_α est celle de *Fisher-Snédécor*.

2.1.5 Conclusion

1. On choisit un seuil de confiance α.
2. On garde l'hypothèse H_o, le facteur contrôlé n'apas d'influence, donc la population est homogène, si

$$\mathbf{f}_{obs} < \mathbf{f}_\alpha.$$

3. On rejette l'hypothèse H_0, le facteur exerce une influence et donc, la population n'est pas homogène si

$$\mathbf{f}_{obs} > \mathbf{f}_\alpha.$$

Avec $\mathbf{f}_{obs}$ est la réalisation de la variable (statistique) $\mathbf{F}_{obs}$, et $\mathbf{f}_\alpha$ la valeur critique, lue sur les tables de Fisher, dépend du seuil α choisi.

2.1.6 Calcul rapide des différentes statistiques

Les résultats suivants sont faciles à démontrer après développement des différents termes carrés

$$nS^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 \quad \text{degré de liberté } (n-1)$$

La quantité $\Delta = \dfrac{1}{n}\left(\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i} x_{ij}\right)^2$ est un terme correctif

$$\begin{aligned} nS_A^2 &= \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \Delta \quad \text{degré de liberté } (k-1) \\ nS_R^2 &= nS^2 - nS_A^2 \text{ degré de liberté } (n-k) \end{aligned}$$

2.1.7 Application et tableau de variation

Tous ces résultats sont résumés dans le tableau d'analyse de la variance.

Variation	Somme des carrés	Degré de liberté	Quotient
Variation due au facteur	nS_A^2	$k-1$	$\dfrac{n\mathbf{S}_A^2}{\mathbf{k-1}}$
Variation résiduelle	nS_R^2	$n-k$	$\dfrac{n\mathbf{S}_R^2}{\mathbf{n-k}}$
Variation totale	nS^2	$n-1$	

Exemple 7 *On veut comparer l'usure de quatre types de pneumatiques P_1, P_2, P_3 et P_4. Sur chacun d'eux, on fait un certain nombre d'essais, 4 ou 5, les coefficients d'usure sont donnés dans le tableau* (2.2) *(en excès au-delà de la valeur* 80*)*

$N°$ *de l'essai*	P_1	P_2	P_3	P_4
1	3	1	2	3
2	3	1	5	3
3	4	2	6	2
4	5	4	4	1
5			4	4
Total	$\sum_{i=1}^{4} x_{i,1} = 15$	$\sum_{i=1}^{4} x_{i,2} = 8$	$\sum_{i=1}^{4} x_{i,3} = 21$	$\sum_{i=1}^{4} x_{i,4} = 13$

Peut-on considérer que les quatre types de pneumatiques sont équivalents ?

Solution 8 *Pneumatique* $P_1 : \bar{x}_1 = 3,75$ $s_1^2 = 0,6875$
Pneumatique $P_2 : \bar{x}_2 = 2, \quad s_2^2 = 1,5$
Pneumatique $P_3 : \bar{x}_3 = 4,\ 2 \quad s_3^2 = 1,76$
Pneumatique $P_4 : \bar{x}_4 = 2,6 \quad s_4^2 = 1,04$

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + n_4\bar{x}_4}{n}$$

- Test sur l'égalité des variances : on compare les estimations des variances des échantillons I et III (la plus petite et la plus grande) . Si ces variances peuvent être considérées ' comme egales, 1e rapport : $\frac{5 \times l,76}{4} \times \frac{3}{4 \times 0,68756} = 240$ est la réalisation d'une variable de Fisher $F(4\ ,3)$.

$$P_r\left\{F(4,3) > 9,12\right\} = 0,05$$

On ne peut pas rejeter l'hypothèse d'égalité des variances des échantillons I et III. Les quatre variances peuvent être considérées comme égales.

Analyse de la variance :

1- Nombre total d'observations, $n = \sum_{i=1}^{4} n_i = 18$

2- Somme de tous les termes,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^{4} x_{i,1} + \sum_{i=1}^{4} x_{i,2} + \sum_{i=1}^{5} x_{i,3} + \sum_{i=1}^{5} x_{i,4} = 15 + 8 + 21 + 13 = 57$$

3- Variation totale,

$$nS^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 = 3^2 + 3^2 + ... + 1^2 + 4^2 - \frac{57^2}{18} = 36,5$$

4- Variation due au facteur,

$$nS_A^2 = \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 = \frac{15^2}{4}+\frac{8^2}{4}+\frac{21^2}{5}+\frac{13^2}{5}-\frac{57^2}{18} = 13,75$$

5- Variation résiduelle, $nS_R^2 = 36,5 - 13,75 = 22,75$

Variation	Somme des carrés	Degré de liberté	Quotient
due au facteur	$13,75$	3	$\frac{n \times S_A^2}{3} = 4,58$
résiduelle	$22,75$	14	$\frac{n \times S_R^2}{14} = 1,625$
totale	$36,50$	17	

$$\begin{aligned} F_{obs} &= \frac{\frac{n \times S_A^2}{3}}{\frac{n \times S_R^2}{14}} = 2,82,\ \mathbf{f}_{0.05} = f(3,14) = 3.34 \\ &\Rightarrow \quad Pr(F(3,14) > 2,82) = 0,95 \end{aligned}$$

On peut admettre que la population est homogène, il n'y a pas de différence entre les quatre types de pneumatiques.

L'estimation de l'usure moyenne est égale à $57/18 = 3,17$ (moyenne générale) et celle de la variance au quotient $\frac{nS_R^2}{14} = 1,625$.

Exemple 9 *Dans une clinique de réhabilitation on veut vérifier si la condition physique avant une intervention chirurgicale au genou a un effet sur le nombre de jours de physiothérapie pour conduire à une réhabilitation complète. La condition physique est évaluée selon un barème qui donne moyenne, sous la moyenne ou au dessous de la moyenne. Voici les données en jours de traitement pour obtenir une réhabilitation complète Condition 0*

	Inférieure	*Moyenne*	*Supérieure*	
	29	30	26	
	42	35	32	
	38	39	21	
	40	28	20	
	43	31	23	
	40	33	22	
	30	29		
	42	35		
		29		
		31		
Total	304	320	144	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = 768$
Moyennes	$\bar{x}_1 = 38$	$\bar{x}_2 = 32$	$\bar{x}_3 = 24$	$\bar{x} = 32$
Nombres d'observations	$n_1 = 8$	$n_2 = 10$	$n_3 = 6$	

Donner l'évaluation du nombre moyen de jours de réhabilitation par groupe. Peut-on dire au niveau 5% *que la condition physique in‡uence la temps de réhabilitation ? La conclusion resterait-elle la même pour un niveau de* 1%*.*

Solution 10 *Le facteur étudier c'est condition physique avec* $k = 3$*, et* $n_1 = 8 : n_2 = 10 : n_3 = 6 : n = \sum_{i=1} n_i = 24 :$

L'hypothèse nulle H_0 *: "Il n'y'a pas une influence du facteur A sur les nombres des jours"*

$\bar{x}_1 = 38 : \bar{x}_2 = 32$, $\bar{x}_3 = 24$ $\bar{x} = 32$

$\sum\sum x_{ij}^2 = 25664$, $\sum n_j \bar{x}_j^2 = 25248$

$nS_A^2 = SCE_{inter} = \sum n_j {\bar{x}_j}^2 - n\bar{x}^2 = 25248 - 24(32)^2 = 672 > 0.$

$SCE_{intra} = \sum\sum x_{ij}^2 - \sum n_j(\bar{x}_j)^2 = 25664 - 25248 = 416 > 0.$

Tableaux des variations

Sources de variation	*Degrés de liberté*	*Somme des Carrés des Ecarts*	*Carée Moyen*	*Test de Fisher-Snédécor*
Totale	$N - 1 = 23$			
Facteur	$p - 1 = 2672$	672	$CM_{inter} = 336$	$Fobs = 16,961$
Résiduelle	$N - p = 21$	416	$CM_{intra} = 19,810$	

$$f_{0,95}(p-1, n-p) = f_{0,95}(2, 21) = 3,44$$

Il est clair que $F_{obs} > f_{0,95}(p-1, n-p)$. *Donc on rejette* H_0 *: il y'a un effet du facteur A sur les nombres des jours.*

Exemple 11 *Supposons que nous ayons* 3 *forêts contenant un type d'arbre bien déterminé oú nous désirons savoir si ces forêts ont une influence sur la hauteur des arbres ou non. A cet effet, nous avons réalisés un recueil de hauteur de six* (06) *arbres dans chaque forêt, dont les mesures sont rangées dans le tableau suivant*

N^0	*forêt* 1	*forêt* 2	*forêt* 3	
1	23.3	18.9	22.5	
2	24.4	21.1	22.9	
3	24.6	21.1	23.7	
4	24.9	22.1	24.0	
5	25.0	22.5	24.0	
6	26.2	23.5	24.5	
Total	148.4	129.2	141.6	$\sum_{i=1}^{3}\sum_{j=1}^{n_i} x_{ij} = 419.2$
Moyennes	$\bar{x}_1 = 24.73$	$\bar{x}_2 = 21.53$	$\bar{x}_3 = 23.60$	$\bar{x} = 23.2889$
Nombres d'observations	$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	

$$\begin{aligned}
nS^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 \\
&= (23.3)^2 + (24.4)^2 + (24.6)^2 + ... + (24.0)^2 + (24.5)^2 - \frac{1}{18}(419.2)^2 \\
&= 9814 - \frac{1}{18}(419.2)^2 = 51.297 \\
nS_A^2 &= \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 \\
&= \frac{(148.4)^2}{6} + \frac{(129.2)^2}{6} + \frac{(141.6)^2}{6} - \frac{1}{18}(419.2)^2 = 31.58 \\
nS_R^2 &= nS^2 - nS_A^2 = 19.717
\end{aligned}$$

Variation	*Somme des carrés*	*Degré de liberté*	*Quotient*	f_{obs}	*Ficher*
Due au facteur	31.58	2	$\frac{n \times S_A^2}{2} = \frac{31.58}{2} = \mathbf{15.79}$	$\frac{\mathbf{15.79}}{\mathbf{1.314}} = \mathbf{12.02}$	***3.6823***
Résiduelle	19.717	15	$\frac{n \times S_R^2}{14} = \frac{19.717}{15} = \mathbf{1.314}$		
Totale	$51,297$	17			

Décision *: On constate que* $f_{obs} = 12,02 > f_\alpha = 3,6823$ *(pour un risque de* $\alpha = 5\%$*), donc les hauteurs moyennes des arbres sont significativement différentes d'une forêt á une autre. Cela signifié que le facteur forêt influe sur la hauteur des arbres*

Exemple 12 *Nous souhaitons comparer quatre traitements, notés* A, B, C *et* D*. Nous répartissons par tirage au sort les patients, et nous leur affectons l'un des quatre traitements. Nous mesurons sur chaque patient la durée, en jours, séparant de la prochaine crise d'asthme. Les mesures sont reportées dans le tableau ci-dessous :*

Traitement A	*Traitement B*	*Traitement C*	*Traitement D*
36	42	26	42
37	38	26	45
35	39	30	50
38	42	38	56
41	44	34	58

Pouvons-nous conclure, à un seuil de risque 1%*, que les facteur traitement a une influence sur le critère retenue*

Solution

$n = 20$, $n_1 = n_2 = n_3 = n_4 = 5$ $\bar{x}_1 = 24.73$, $\bar{x}_2 = 21.53$, $\bar{x}_3 = 23.60$, $x_4 =$, $\bar{x} = 23.2889$

	Traitement A	*Traitement B*	*Traitement C*	*Traitement D*
	36	42	26	42
	37	38	26	45
	35	39	30	50
	38	42	38	56
	41	44	34	58
$Total$ $\sum_{i=1}^{4}\sum_{j=1}^{n_j} x_{ij} = 797$	187	205	154	251
Moyennes	$\bar{x}_1 = 37.4$	$\bar{x}_2 = 41$	$\bar{x}_3 = 30.8$	$\bar{x}_4 = 50.2$
Nombres d'observations	5	5	5	5

$$
\begin{aligned}
nS^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 = \\
&= (36)^2 + (37)^2 + (35)^2 + ... + (56)^2 + (58)^2 - \frac{1}{20}(797)^2 \\
&= 33085 - \frac{1}{20}(797)^2 = 1324.55 \\
nS_A^2 &= \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 \\
&= \frac{(187)^2}{5} + \frac{(205)^2}{5} + \frac{(154)^2}{5} + \frac{(251)^2}{5} - \frac{1}{20}(797)^2 = 981.75 \\
nS_R^2 &= nS^2 - nS_A^2 = 1324.55 - 981.75 = 342.8
\end{aligned}
$$

	SC	*ddl*	*CM*	*f*	f_α
Inter-groupes	981.75	3	327.25	15. 274	5.29
Intra-groupes	342.8	16	21.425		
Total	1324.55	19			

On constate que $f > f_\alpha$ cela signifie qu'on doit rejeter H_0. C'est-à-dire le facteur traitement a une influence significative sur les durées séparant deux crise d'asthme.