Ricco Rakotomalala

Comparaison de populations

Tests paramétriques

Version 1.2

Université Lumière Lyon 2

Avant-propos

Ce support est dédié aux tests paramétriques de comparaison de populations.

Comparaison de populations. *Stricto sensu*, les tests de comparaisons de populations cherchent à déterminer si K ($K \geq 2$) échantillons proviennent de la même population relativement à la variable d'intérêt. Nous sommes dans le cadre de la statistique inférentielle : à partir d'échantillons, nous tirons des conclusions sur la population. Au delà de ces aspects purement théoriques, les applications pratiques sont nombreuses. Un test de comparaison répond à des questions très concrètes :

- Vérifier que la teneur en sel du hamburger de la marque A est différente de celle de la marque B. Pour cela, on réalise un prélèvement dans les différents restaurants de chaque marque. On compare les moyennes des teneurs en sel de chaque type de hamburger (cf. *comparaison de moyennes, échantillons indépendants*).
- Évaluer la réduction de la variabilité des pièces produites par une machine en introduisant de nouvelles normes de qualité (cf. *comparaison de variances*).
- Dans un couple marié, composé de 2 personnes actives, l'homme a-t-il un salaire plus élevé que sa compagne (cf. *comparaison sur échantillons appariés*).

On peut aussi considérer la comparaison de populations sous l'angle de l'étude de la liaison entre une variable catégorielle et une variable continue. Par exemple, pour les habitations, on veut analyser l'effet du type de chauffage utilisé et le montant de la facture annuelle. Ou encore, analyser le rôle bénéfique de différents additifs de carburants sur la consommation des véhicules. Dans ce cadre, la variable catégorielle sert à définir les sous populations, la variable continue correspond au caractère que l'on cherche à étudier.

Paramétrique. On parle de tests paramétriques lorsque l'on fait l'hypothèse que les variables qui décrivent les individus suivent une distribution paramétrée. Dans ce support, nous analyserons principalement (mais pas seulement) le cas des **variables continues gaussiennes**. Les paramètres sont estimés à partir des échantillons et, dans ce cas, les tests reviennent simplement à les comparer puisqu'elles définissent de manière non ambiguë la distribution. Ainsi, concernant la distribution gaussienne, les tests porteront essentiellement sur la moyenne et l'écart type. L'hypothèse de normalité n'est pas aussi restrictive qu'on peut le penser, nous en discuterons de manière détaillée plus loin.

Ce support se veut avant tout opérationnel. Il se concentre sur les principales formules et leur mise en oeuvre pratique avec un tableur. Autant que possible nous ferons le parallèle avec les résultats fournis par les logiciels de statistique. Le bien-fondé des tests, la pertinence des hypothèses à opposer sont peu

ou prou discutées. Nous invitons le lecteur désireux d'approfondir les bases de la statistique inférentielle, en particulier la théorie des tests, à consulter les ouvrages énumérés dans la bibliographie.

Un document ne vient jamais du néant. Pour élaborer ce support, je me suis appuyé sur différentes références, des ouvrages disais-je plus tôt, mais aussi des ressources en ligne qui sont de plus en plus présents aujourd'hui dans la diffusion de la connaissance. Les seuls bémols par rapport à ces documents en ligne sont le doute que l'on pourrait émettre sur l'exactitude des informations prodiguées, mais la plupart de leurs auteurs sont des enseignants-chercheurs qui font sérieusement leur travail ; une disponibilité plus ou moins aléatoire, au gré des migrations des serveurs et de la volonté de leurs auteurs, auquel il est très difficile de remédier ; les informations sont disparates, avec une absence d'organisation, à la différence des ouvrages qui suivent une ligne pédagogique très structurante. Néanmoins, ces ressources en ligne renouvellent profondément le panorama des documents disponibles pour les enseignements. La gratuité n'est pas le moindre de leurs atouts.

Concernant ce support de cours, rendons à César ce qui lui appartient, il a été en grande partie inspiré par les références suivantes :

1. Pour les tests univariés : le manuel *Engineering Statistics Handbook* du NIST, disponible en ligne `http://www.itl.nist.gov/div898/handbook/index.htm`, en particulier le chapitre 7 "Product and Process Comparisons". Ce site est d'autant plus remarquable que les techniques présentées sont programmées dans le logiciel DATAPLOT qui fait référence auprès des statisticiens (`http://www.itl.nist.gov/div898/software/dataplot/document.htm`).
2. Pour les tests multivariés : le cours en ligne de S. Rathbun et A. Wiesner du PennState University, "Applied Multivariate Statistical Analysis (STAT 505)", `http://www.stat.psu.edu/online/development/stat505/`
3. De manière plus globale, l'ouvrage de Howell [7]. Voilà un document que je trouvais initialement un peu bavard, voire rebutant pour le fanatique des formules que je suis. En insistant un peu, après plusieurs relectures, je me suis rendu compte de la richesse extraordinaire du texte, du recul serein de l'auteur par rapport aux techniques, et de la profusion exceptionnelle des références bibliographiques. Le site Web associé à l'ouvrage (`http://www.uvm.edu/~dhowell/methods/index.html`) propose différentes ressources : les données utilisées dans le texte, les corrections des exercices, les erratum, des liens vers d'autres sites relatifs aux techniques statistiques.

Enfin, selon l'expression consacrée, ce support n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont le bienvenu.

Table des matières

Partie I Tests pour échantillons indépendants

Partie III Tests multivariés

Concernant l'hypothèse de normalité

L'hypothèse sous-jacente de normalité n'est pas aussi restrictive qu'on peut le penser disions nous plus haut. Il y a plusieurs raisons à cela :

- Tout d'abord, il est possible de **transformer les variables** de manière à se rapprocher de la loi normale. Les fonctions les plus connues sont le log ou la racine carrée. Avec les fonctions de Box-Cox, nous disposons d'outils paramétrables qui nous permettent de réaliser au mieux l'opération [1].
- Certains tests sont **robustes** c.-à-d. même si l'on s'écarte légèrement des hypothèses sous-jacentes initiales, ils restent valables. Il faut vraiment que la violation soit patente (distributions très dissymétriques ou bimodales) pour que la procédure ne soit pas opératoire. D'autres en revanche ne sont pas robustes du tout, les tests d'égalité des variances de Fisher et de Bartlett par exemple s'effondrent totalement dès que l'on s'écarte, même légèrement, de la distribution gaussienne. Nous distinguerons clairement les différences de comportement dans notre texte.
- Lorsque les effectifs augmentent, le **théorème central limite** joue à plein. En effet, il stipule que la somme de variables aléatoires de même moyenne et écart-type tend vers la loi normale [2]. De fait, les statistiques composées à partir d'une somme de variables aléatoires, la moyenne mais aussi la proportion, tendent vers la loi normale dès que les effectifs sont suffisamment élevés (de l'ordre de 30 en pratique), quelle que soit la distribution initiale sous jacente. Ce résultat élargit considérablement le champ d'action des tests que nous présenterons dans ce support de cours. Ce qui explique d'ailleurs leur popularité dans la pratique.

Bien entendu, si aucune des conditions ci-dessus ne sont réunies, il est inutile de s'entêter. On se tournera avantageusement vers les tests non paramétriques. Ils feront l'objet d'un support spécifique prochainement.

Notations

Les données proviennent de K échantillons Ω_k $(k = 1, \ldots, K)$. La variable X est notée en majuscule, la valeur pour l'observation $n^o i$ est notée x_i en minuscule. Parfois, il sera nécessaire de trier les valeurs, dans ce cas la série triée sera notée $x_{(i)}$ c.-à-d. $x_{(1)}$ correspond à la plus petite valeur. L'effectif global est n, les sous-échantillons comportent n_k observations, avec $n = n_1 + \ldots + n_K$.

La moyenne théorique (resp. estimée) est notée μ (resp. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$).

L'écart type théorique (resp. estimée) est notée σ (resp. $s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$).

Toutes les statistiques conditionnelles, associées aux sous-populations, sont indicées par k (ex. la moyenne théorique de la sous-population $n^o 1$ sera μ_1, etc.). La valeur de l'individu $n^o i$ dans l'échantillon Ω_k sera notée x_{ik}

Sauf mention contraire, nous choisirons comme seuil de signification $\alpha = 5\%$ pour tous les tests de ce support.

1. Voir `http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf` concernant les tests de normalité et la transformation de Box-Cox
2. Voir `http://fr.wikipedia.org/wiki/Théorèmes_limites_(probabilités)`

Données

L'analyse d'un fichier de demandeurs de crédits servira de trame dans ce support. Nous l'utiliserons pour illustrer les différents tests que nous présentons. Il comporte $n = 50$ observations. Chaque ligne correspond à un ménage composé d'un homme, d'une femme et éventuellement des personnes à charge (les enfants principalement). Les variables sont les suivantes (Figure 0.1) :

Numéro	Sal.Homme	Sal.Femme	Rev.Tete	Age	Acceptation	Garantie.Supp	Emploi
1	7.92	7.72	7.42	3.69	oui	hypotheque	cdd
2	7.97	7.49	7.76	3.89	oui	caution	cdd
3	6.97	7.10	6.35	3.53	non	non	cdd
4	7.85	7.39	7.24	3.78	oui	caution	cdd
5	6.67	6.76	5.46	3.78	oui	hypotheque	cdd
6	6.89	6.51	6.72	4.16	non	hypotheque	cdd
7	7.29	6.93	6.43	3.37	oui	hypotheque	cdd
8	7.53	7.51	7.52	3.99	oui	hypotheque	cdd
9	7.48	7.25	6.46	3.47	oui	non	cdi
10	7.27	6.60	6.59	3.30	oui	hypotheque	cdi
11	7.28	7.47	6.97	3.56	oui	non	cdi
12	8.40	8.07	7.84	3.76	oui	caution	cdi
13	7.46	6.79	6.26	3.40	oui	hypotheque	cdi
14	8.42	8.01	7.83	3.47	oui	non	cdi
15	7.39	7.44	7.42	3.89	non	non	cdd
16	7.47	7.59	7.53	3.78	oui	non	cdi
17	7.86	7.50	7.29	3.64	oui	hypotheque	cdi
18	6.83	7.06	6.03	3.74	oui	hypotheque	cdi
19	6.98	7.29	6.74	3.83	non	hypotheque	cdd
20	7.80	7.38	7.61	3.97	oui	hypotheque	cdi
21	7.67	7.69	7.27	3.81	oui	hypotheque	cdi
22	7.28	7.05	7.17	3.30	oui	caution	cdi
23	7.17	6.86	6.62	3.40	non	hypotheque	cdd
24	7.42	7.25	6.42	3.40	non	non	cdd
25	7.83	7.77	7.40	3.76	oui	hypotheque	cdi
26	7.33	7.14	7.24	3.18	oui	hypotheque	cdi
27	6.02	6.03	5.11	3.26	oui	hypotheque	cdi
28	7.63	7.77	6.79	3.66	oui	non	cdi
29	6.18	6.40	6.30	4.08	oui	non	cdi
30	7.57	7.53	6.63	3.43	oui	non	cdi
31	7.36	7.78	6.90	3.74	oui	hypotheque	cdi
32	8.03	7.94	7.29	3.78	oui	non	cdi
33	8.46	8.12	8.30	3.69	oui	hypotheque	cdi
34	6.64	7.12	6.22	3.50	oui	hypotheque	cdi
35	7.92	7.92	6.82	3.66	oui	non	cdi
36	7.14	7.20	6.26	3.78	non	hypotheque	cdd
37	7.13	6.85	6.08	3.85	non	caution	cdd
38	7.43	7.20	7.32	4.11	oui	hypotheque	cdi
39	8.78	8.58	8.69	3.78	oui	non	cdi
40	8.28	7.85	7.68	3.74	oui	hypotheque	cdi
41	6.31	6.57	5.75	3.66	non	hypotheque	cdd
42	7.48	6.97	7.26	3.74	non	hypotheque	cdd
43	7.48	6.96	6.85	3.37	non	hypotheque	cdi
44	7.69	7.11	7.44	4.16	non	hypotheque	cdi
45	7.44	7.16	6.91	3.78	non	non	cdi
46	7.47	7.24	6.45	3.66	oui	hypotheque	cdi
47	8.17	8.29	8.23	3.95	oui	non	cdi
48	7.40	7.29	7.35	3.09	non	hypotheque	cdi
49	7.26	6.81	7.06	4.16	non	non	cdi
50	7.50	7.16	7.35	4.08	non	hypotheque	cdi

Fig. 0.1. Fichier des demandeurs de crédits

1. Le logarithme du salaire de l'homme (Sal.Homme) ;
2. Le logarithme du salaire de la femme (Sal.Femme) ;
3. Le logarithme du revenu par tête (Rev.Tete). Le revenu par tête correspond au revenu du ménage (salaire homme + salaire femme) divisé par le nombre de personnes ;
4. Le logarithme de l'âge de l'homme (Age) ;
5. L'accord du crédit par l'organisme prêteur (Acceptation - 2 modalités) ;
6. La garantie supplémentaire demandée à l'emprunteur (Garantie.Supp - 3 modalités).
7. Le type d'emploi occupé par l'emprunteur (la personne inscrite en premier dans le formulaire de demande c.-à-d. la personne de référence) (Emploi - 2 modalités)

Nous avons pris les logarithmes pour les variables continues de manière à corriger une asymétrie à gauche.

Partie I

Tests pour échantillons indépendants

Pour obtenir des échantillons indépendants, il y a 2 manières de procéder :

1. Dans chaque sous population, on décide de prélever n_k observations. Dans ce cas, la valeur n_k résulte de la décision du statisticien, il ne reflète pas *a priori* la taille relative Ω_k. Parfois, il est décidé arbitrairement que $n_1 = n_2 = \cdots = n_K$ afin d'améliorer l'efficacité ou la robustesse de certains tests (voir par exemple l'ANOVA à 1 facteur).
2. On effectue un prélèvement aléatoire dans la population globale, puis on se sert d'une variable catégorielle pour distinguer les observations relatives à chaque sous population. Nous avons également affaire à des échantillons indépendants dans ce cas, à la différence que cette fois-ci la fréquence $f_k = \frac{n_k}{n}$ reflète la taille relative de Ω_k.

Pour nous, qu'importe le mode de tirage, il faut simplement qu'une observation quelconque de Ω_k n'ait aucun lien particulier avec une observation de $\Omega_j \quad (j \neq k)$. Les échantillons sont indépendants de ce point de vue.

De même, mais est-ce nécessaire de le préciser, toutes les observations dans chaque sous échantillon doivent être indépendants et identiquement distribuées (*i.i.d.*).

1

Comparaison de moyennes

1.1 Estimation de la moyenne et théorème central limite

Avant de rentrer dans le vif du sujet, penchons nous un instant sur une propriété remarquable de la moyenne, estimateur sans biais de l'espérance mathématique.

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires (v.a.) *indépendantes de même loi de répartition*, d'espérance μ et de variance σ^2. Alors la v.a.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est de moyenne et variance

$$E(\bar{X}) = \mu$$
$$V(\bar{X}) = \frac{\sigma^2}{n}$$

Concernant les distributions,

- si X suit une loi normale $\mathcal{N}(\mu, \sigma)$ alors la loi exacte de $\bar{X}$ est $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;
- si X suit une loi quelconque, en vertu du **théorème central limite**, lorsque $n \to +\infty$, alors $\bar{X}$ tend vers la loi normale. En pratique, dès que $n \geq 30$, l'approximation devient effective.

Cette propriété, remarquable, élargit considérablement le champ d'application des tests de comparaison de moyenne que nous présentons dans ce chapitre. La restriction imposée par l'hypothèse de normalité sous-jacente de X est levée. Dès que n est suffisamment élevé, la quantité Z avec

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Suit une loi normale centrée réduite, quelle que soit la loi sous-jacente de X.

Cas où l'écart type σ de X n'est pas connu

Si nous ne disposons de la *vraie* valeur de l'écart type, nous utilisons l'estimateur sans biais s de σ avec

$$s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

Dans ce cas, le rapport

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Suit une loi de Student $\mathcal{T}(\nu)$ à $\nu = n - 1$ degrés de liberté. On remarquera que lorsque $n \to +\infty$, la distribution de Student se confond avec la loi normale.

Ces résultats sont importants, nous en ferons largement usage dans ce qui suit.

1.2 Comparaison de 2 moyennes

Notre variable d'intérêt est X. Nous souhaitons comparer la moyenne de X dans 2 sous populations. Le test d'hypothèses s'écrit :

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Le test peut être unilatéral à gauche ou à droite. Le schéma global reste le même, la statistique du test et les degrés de liberté sont identiques. Seule la région critique sera modifiée.

Soient 2 échantillons Ω_1 et Ω_2 prélevés à partir de 2 sous populations. Nous formons les moyennes conditionnelles empiriques :

$$\bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik} \quad , \quad k = 1, 2$$

Le test de comparaison de 2 moyennes consiste à confronter les quantités estimées $\bar{x}_1$ et $\bar{x}_2$ en tenant compte de la dispersion (variance) des valeurs dans chaque sous-groupe. Les calculs diffèrent selon les hypothèses relatives aux variances conditionnelles.

1.2.1 Cas des variances connues

Le cas où les variances sont connues dans les sous-groupes est purement théorique. Pourtant la très grande majorité des présentations du test de comparaison de moyennes commencent par cette configuration. En effet, elle comporte tous les éléments de compréhension du test de comparaison de 2 moyennes.

Nous formons l'écart $D = \bar{X}_1 - \bar{X}_2$. L'espérance de D est

$$E(D) = \mu_1 - \mu_2$$

Les échantillons étant indépendants, sa variance est obtenue directement avec

$$\begin{aligned} V(D) &= V(\bar{X}_1 - \bar{X}_2) \\ &= V(\bar{X}_1) + V(\bar{X}_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

Sous $H_0 : \mu_1 = \mu_2$, la statistique du test de comparaison de moyenne devient

$$Z = \frac{D}{\sigma_D} = \frac{D}{\sqrt{V(D)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1.1}$$

Puisque X est distribuée normalement, Z l'est également. Pour un test bilatéral, la région critique du test (rejet de H_0) s'écrit

$$R.C. : |Z| \geq u_{1-\frac{\alpha}{2}}$$

$u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi normale centrée réduite.

1.2.2 Cas des variances égales

Statistique du test

Dans la pratique, nous ne connaissons pas les valeurs σ_k, il nous faut les obtenir à partir des données, nous utilisons les estimateurs non biaisés

$$s_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2$$

Si l'on fait l'hypothèse que les variances sont identiques dans les sous-groupes, nous pouvons produire un estimateur synthétique de la variance s^2 avec

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

La statistique du test devient

$$T = \frac{D}{\hat{\sigma}_D} = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1.2}$$

Sous H_0, elle suit une loi de Student $\mathcal{T}(\nu)$ à $\nu = (n_1 + n_2 - 2)$ degrés de liberté.

Pour un test bilatéral, la région critique est analogue à la précédente :

$$R.C. : |T| \geq t_{1-\frac{\alpha}{2}}(\nu)$$

où $t_{1-\frac{\alpha}{2}}(\nu)$ est le quantile d'ordre $1 - \frac{\alpha}{2}(\nu)$ de la loi de Student.

Il faut bien comprendre le mécanisme que recèle cette formulation. A écart égal entre les moyennes, plus la variabilité des valeurs sera faible, plus nous serons emmenés à rejeter l'hypothèse nulle : les distributions conditionnelles se démarquent plus fortement même si le paramètre de localisation (la moyenne) n'est pas modifié.

Application numérique : comparaison des salaires

Concernant les demandes de crédit (Figure 0.1), un expert financier affirme que l'accord des banques est subordonné au salaire de la femme dans le ménage. Nous souhaitons vérifier cette hypothèse en comparant le logarithme du salaire moyen des femmes dans les deux groupes : ceux qui se sont vus refuser leur crédit (acceptation = non, groupe 2), et ceux qui ont obtenu un accord (acceptation = oui, groupe 1).

Statistiques descriptives

Avant toutes choses, notre premier réflexe est de réaliser des statistiques descriptives. On ne mesure jamais assez la foule d'enseignements que l'on peut en tirer. Quelques graphiques bien sentis font largement autant que des calculs compliqués. Ils permettent de délimiter la portée des résultats numériques. Bien souvent on pourrait même les substituer aux tests, notamment lorsque les effectifs sont très élevés et que les tests statistiques ont tendance à conclure systématiquement au rejet de l'hypothèse nulle.

L'outil le plus simple est certainement la *boîte à moustaches* conditionnelle. Nous avons idée à la fois des paramètres de localisation et de dispersion des variables dans chaque sous groupe. De plus, nous pouvons détecter visuellement les éventuels points atypiques (Figure 1.1).

Plusieurs résultats sautent aux yeux :

- La distribution des salaires féminins chez les crédits acceptés semble décalée (vers les valeurs élevées).
- La dispersion paraît un peu plus forte aussi. L'égalité des variances est sujette à caution. Notre idée justement est de comparer les résultats obtenus en intégrant ou non cette hypothèse.

Fig. 1.1. Boîtes à moustaches des salaires féminins selon l'acceptation

- Il y a un point suspect chez les crédit acceptés. Après auscultation du fichier, il s'agit de l'observation $n^o 27$, les deux personnes ont des revenus sensiblement plus faibles que les autres, y compris en ce qui concerne le salaire de la femme. Pourtant, ils se sont vu accorder leur crédit. Nous ne céderons pas à la tentation de supprimer cette observation. Toutefois, nous n'oublierons pas cette information par la suite. Elle peut avoir des conséquences sur les résultats des tests [1].

Test de comparaison de moyennes

A l'aide de l'outil "tableau croisé dynamique" d'EXCEL [2], nous calculons les effectifs, les moyennes et les écarts type dans chaque groupe (Figure 1.2) : nous obtenons $n_1 = 34$, $\bar{x}_1 = 7.4394$ et $s_1 = 0.5483$ (resp. $n_2 = 16$, $\bar{x}_2 = 7.0331$ et $s_2 = 0.2615$). A partir de ces résultats, nous pouvons compléter le test de comparaison :

- L'écart entre les moyennes est $D = 7.4394 - 7.0331 = 0.4063$
- L'estimation de l'écart type commun (la racine carrée de la variance intra-classes)

$$s = \sqrt{\frac{(34-1) \times 0.5483^2 + (16-1) \times 0.2615^2}{34+16-2}} = 0.4775$$

- Nous pouvons en extraire l'estimation de l'écart type de D avec $\hat{\sigma}_D = 0.4775 \times \sqrt{\frac{1}{34} + \frac{1}{16}} = 0.1448$
- Finalement, nous produisons la statistique du test $T = \frac{0.4063}{0.1448} = 2.8063$
- Le nombre de degrés de liberté est $\nu = 34 + 16 - 2 = 48$

1. Lorsque ces "anomalies" prennent de l'ampleur, on peut s'interroger sur les méthodes à utiliser. Le passage aux techniques basées sur les rangs, les tests non paramétriques, permettrait par exemple de *gommer* ce type de scories et présenter des résultats plus crédibles. C'est ce type d'interrogations que doivent nous emmener les statistiques descriptives.

2. Pour la mise en oeuvre de ce formidable outil d'EXCEL, plusieurs références en ligne sont recensés sur le site `http://eric.univ-lyon2.fr/~ricco/cours/cours_excel.html`

	A	B	C	D
1				
2				
3		Données		
4	Acceptation	Nombre de Sal.Femme	Moyenne de Sal.Femme2	Écartype de Sal.Femme3
5	oui	34	7.4394	0.5483
6	non	16	7.0331	0.2615
7	Total	50	7.3094	0.5099
8				
9	Test de comparaison de moyennes - Variances égales			
10				
11		Ecart (D)	0.4063	
12				
13		s^2	0.2280	
14		s	0.4775	
15				
16		Sigma(D)	0.1448	
17				
18		T	2.8063	
19		ddl	48	
20				
21		t_0.975(48)	2.0106	
22		p-value	0.0072	

Fig. 1.2. Comparaison des salaires féminins selon l'acceptation du crédit - Variances égales

– Au risque $\alpha = 5\%$, nous devons comparer T avec le seuil $t_{0.975}(48) = 2.0106$, les données ne sont pas compatibles avec l'hypothèse d'égalité des moyennes. Apparemment, l'accord de crédit des banques est bien subordonné au salaire de la femme. Nous aboutissons bien entendu à la même conclusion si nous nous référons à la probabilité critique du test (p-value = 0.0072).

Robustesse du test de comparaison de moyennes

Un écart modéré par rapport à la normalité des distributions ne perturbe pas (trop) le test de comparaison de moyennes, surtout si les distributions conditionnelles sont symétriques. Si les distributions sont dissymétriques, mais qu'elles le sont de la même manière dans les sous-groupes, le test de Student s'applique quand même. Lorsque les effectifs sont élevés, le théorème central limite balaye toutes les hésitations.

En toute rigueur, le test de comparaison de moyenne devrait être précédé par un test de comparaison de variances. En effet, nous émettons une hypothèse d'homoscédasticité, les variances sont identiques dans les sous-groupes, elle doit être vérifiée. En pratique, il semble que ce ne soit pas nécessaire dans la grande majorité des cas, sauf violation flagrante visible dans les statistiques descriptives. Un écart modéré par rapport à cette hypothèse n'est pas problématique, ceci d'autant plus que les effectifs sont équilibrés c.-à-d. $n_1 \approx n_2$.

En revanche, lorsque les effectifs sont déséquilibrés, n_1 très différent de n_2, on privilégiera plutôt le test de la section suivante.

1.2.3 Cas des variances inégales

Statistique du test, estimation de la variance et degrés de liberté

Lorsque nous nous affranchissons de l'hypothèse d'homoscédasticité, le schéma global reste d'actualité, notamment l'utilisation de l'écart D, la distribution de Student et la définition de la région critique. En revanche, il nous faut produire 2 nouveaux éléments importants :

1. L'estimation $\hat{\sigma}_D$ de l'écart type de D, elle devient maintenant

$$\hat{\sigma}_D = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

2. Un calcul approprié des degrés de liberté ν avec

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \tag{1.3}$$

Cette formule est appelée "équation de Welch-Satterthwaite" [3]. Il en existe des variantes plus précises si on se réfère à la littérature (voir [7], page 226), mais les logiciels implémentent principalement celle-ci.

Application numérique : comparaison des salaires selon l'acceptation

Reprenons notre exemple de la section précédente, nous introduisons ces nouvelles informations dans la feuille EXCEL (Figure 1.3) :

- La partie haute de la feuille de calcul n'est pas modifiée, nous avons toujours $D = 7.4394 - 7.0331 = 0.40629$
- L'estimation de la variance est modifiée maintenant, avec $\hat{\sigma}_D^2 = \frac{0.5483^2}{34} + \frac{0.2615^2}{16} = 0.01312$, et $\hat{\sigma}_D = \sqrt{0.01312} = 0.11452$
- La statistique du test de devient $T = \frac{0.40629}{0.11452} = 3.54760$. Largement plus élevée que celle obtenue sous l'hypothèse d'égalité des variances (nous avions $T = 2.08063$).
- Voyons ce qu'il en est des degrés de liberté pour pouvoir statuer sur la conclusion du test. Nous décomposons la formule 1.3, nous obtenons le numérateur

$$A = (\frac{0.5483^2}{34} + \frac{0.2615^2}{16})^2 = 0.000172$$

et le dénominateur

$$B = \frac{0.5483^4}{34^2 \times (34-1)} + \frac{0.2615^4}{16^2 \times (16-1)} = 0.000004$$

. Par conséquent, $\nu = \frac{A}{B} = 47.96$

3. http://en.wikipedia.org/wiki/Welch-Satterthwaite_equation

	A	B	C	D
1				
2				
3		Données		
4	Acceptation	Nombre de Sal.Femme	Moyenne de Sal.Femme2	Écartype de Sal.Femme3
5	oui	34	7.4394	0.5483
6	non	16	7.0331	0.2615
7	Total	50	7.3094	0.5099
8				
9	**Test de comparaison de moyennes - Variances inégales**			
10				
11		Ecart (D)	0.40629	
12				
13		Sigma^2(D)	0.01312	
14		Sigma(D)	0.11452	
15				
16		T	3.54760	
17				
18		A	0.000172	
19		B	0.000004	
20		ddl	47.96	
21				
22		t_0.975(47)	2.01174	
23		t_0.975(48)	2.01063	
24				
25		t_0.975(47.96)	2.01068	

Fig. 1.3. Comparaison des salaires féminins selon l'acceptation du crédit - Variances inégales

- Le seuil critique de la loi de Student au risque $\alpha = 5\%$ n'est pas défini pour ces degrés de liberté. Il nous faut donc passer par une interpolation linéaire pour l'obtenir. Sachant que $t_{0.975}(47) = 2.01174$ et $t_{0.975}(48) = 2.01063$, le *véritable* seuil critique sera

$$t_{0.975}(47.96) = \frac{47.96 - 47}{48 - 47} \times (2.01063 - 2.01174) + 2.01174 = 2.01068$$

.

- Puisque $|T| = 3.54760 > t_{0.975}(47.96) = 2.01068$, nous rejetons l'hypothèse d'égalité des salaires selon l'acceptation du crédit au risque $\alpha = 5\%$

C'est un peu le canon pour tuer la mouche tout ça. Nos effectifs étant relativement importants [4], les résultats sont très similaires que l'on intègre ou non une estimation commune de la variance.

A l'usage, on constate que tenir compte de l'inégalité des variances n'est vraiment déterminant que pour les effectifs déséquilibrés [5] c.-à-d. avec n_1 très différent de n_2. Certains auteurs [6] précisent même que l'on **devrait toujours utiliser la variante pour variances inégales dès que n_1 et n_2 sont très différents**, quand bien même le ratio entre la plus grande et la plus petite variance n'excéderait pas 1.5. Procéder préalablement à un test de comparaison de variances pour choisir la procédure adéquate de comparaison de moyennes est illusoire dès lors que les effectifs sont déséquilibrés.

Remarque 1 (Une règle de décision conservatrice... mais plus simple). Notons qu'il est possible d'adopter une démarche conservatrice [7], elle consiste à choisir $\nu = \min(n_1 - 1, n_2 - 1) = \min(34 - 1, 16 - 1) = 15$.

4. Nous nous rapprochons de $n_k \geq 20, \forall k$; voir [11], page 342
5. `http://en.wikipedia.org/wiki/T-test`
6. Zimmerman, D. W. (1996). *Some properties of preliminary tests of equality of variances in the two-sample location problem*, Journal of General Psychology, 1996, 123, 217-231
7. `http://www.stat.psu.edu/online/development/stat800/08_twogroups/04_twogroups_2means.htm`

Dans notre cas, $t_{0.975}(15) = 2.1315$, nous aboutirons quand même au rejet de l'hypothèse d'égalité des moyennes. Nous gagnons en simplicité ce que nous perdons en précision, il faut tout simplement être conscient (et en tenir compte) que ce faisant nous favorisons l'hypothèse nulle.

1.3 Comparaison de K ($K \geq 2$) moyennes - ANOVA à 1 facteur

1.3.1 Comparaison de K ($K \geq 2$) moyennes

L'ANOVA (*analyse de variance*) est une généralisation de la comparaison de moyennes à K sous populations. Les échantillons sont indépendants. Les hypothèses nulles et alternatives s'écrivent maintenant

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$
$$H_1 : \text{Au moins deux moyennes sont différentes}$$

Les hypothèses sous-jacentes sont : (1) X est distribuée normalement et (2) les variances dans les sous groupes sont identiques (homoscédasticité).

L'ANOVA est une *vraie* généralisation au sens où, pour $K = 2$, nous retrouvons exactement le test de Student de comparaison de 2 moyennes avec égalité des variances dans les sous populations (section 1.2.2).

Les applications sont nombreuses. Il peut s'agir d'une véritable comparaison : par exemple, on cherche à comparer la teneur en sel réelle de différents marques de sandwich disponibles dans les distributeurs automatiques ; comparer le salaire moyen des étudiants selon les filières à la sortie de l'université ; etc. [8].

Il peut aussi s'agir d'analyser l'effet d'un facteur représenté par une variable catégorielle sur une variable continue : par exemple, on cherche à évaluer l'impact des différentes méthodes d'enseignements sur les notes des étudiants ; comparer les émissions polluantes des véhicules selon le type de filtre incorporé dans les pots d'échappement ; etc.

Dans notre exemple des demandeurs de crédit (Figure 0.1), on cherche à confronter le salaire des femmes selon le type de garantie demandé (Garantie.Supp). Plus que la comparaison, il s'agit plutôt d'analyser l'association entre les variables "Salaire.Femme" et "Garantie.Supp". On peut se demander par exemple "est-ce que les banques sont enclins à demander des garanties particulières selon le niveau de salaire de la femme dans le couple " ? L'outil privilégié est le rapport de corrélation [9], pour tester sa significativité, nous retombons sur la même statistique que l'ANOVA.

Enfin, on distingue généralement 2 types d'analyses [10] : le modèle à effets fixes, tous les sous-groupes sont représentés dans l'échantillon à analyser ; le modèle à effets aléatoires, les groupes représentés constituent un échantillon des sous populations. Si les conséquences sur l'interprétation des résultats sont

8. `http://lib.stat.cmu.edu/cgi-bin/dasl.cgi?query=ANOVA&submit=Search!&metaname=methods&sort=swishrank` pour des exemples accompagnées de données.
9. Voir `http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf`, section 3.6
10. `http://en.wikipedia.org/wiki/Analysis_of_variance`

importantes, les calculs, notamment la statistique du test, les distributions et les degrés de liberté associés sont les mêmes. Dans ce support, nous nous focalisons avant tout sur l'aspect *comparaison de moyennes*, nous considérons que nous sommes dans un schéma à effets fixes sans que cela ne réduise la portée de l'exposé. C'est le choix approprié dans la grande majorité des cas (voir [6], page 37).

1.3.2 Décomposition de la variance

Pourquoi nommer *analyse de variance* un processus qui consiste à comparer des moyennes ? Pour le comprendre, nous devons nous pencher sur une égalité très importante appelée **formule de décomposition de la variance** ou **équation d'analyse de variance** :

$$SCT = SCE + SCR$$
$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2$$

- **SCT** traduit la variabilité totale dans l'échantillon, c'est la somme des carrés totaux, elle ne dépend pas des groupes.
- **SCE** traduit la variabilité des moyennes conditionnelles, c'est la somme des carrés inter-classes, expliquée par l'appartenance aux groupes. Une valeur élevée de B indique que les moyennes sont très différentes les unes des autres, cela nous amène à rejeter H_0 dans le test d'hypothèses de l'ANOVA. Lorsque SCE augmente, SCR diminue puisque SCT est fixe.
- **SCR** est variabilité à l'intérieur des groupes, c'est la somme des carrés intra-classes, la variabilité résiduelle. Lorsque $SCR \to 0$, cela veut dire que les valeurs sont agglutinés autour des moyennes conditionnelles à l'intérieur des sous-échantillons, la différenciation entre les groupes est forte, toute la variabilité est expliquée par le décalage entre les moyennes conditionnelles $SCE \to SCT$. Dans ce cas, nous sommes emmenés à rejeter l'hypothèse nulle dans le test de comparaison des moyennes.

On comprend intuitivement que le test ANOVA va reposer sur la confrontation entre SCE et SCR. C'est ce que nous montrons dans la section suivante.

1.3.3 Statistique du test et région critique

On résume la décomposition de la variance dans un tableau dit *tableau d'analyse de variance* (voir [11], page 355 ; [7], page 348) [11], fourni en standard par la très grande majorité des logiciels (Tableau 1.1).

La statistique du test F est donc définie par

$$F = \frac{\frac{SCE}{K-1}}{\frac{SCR}{n-K}} = \frac{CME}{CMR} \tag{1.4}$$

11. Voir `http://www.stat.psu.edu/online/development/stat800/09_anova/02_anova_oneway.htm` qui explicite clairement le contenu et l'interprétation du tableau d'analyse de variance

Source	Somme.Carrés	ddl	Carrés.Moyens	F
Expliquée	$SCE = \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})^2$	$K-1$	$CME = \frac{SCE}{K-1}$	$F = \frac{CME}{CMR}$
Résiduelle	$SCR = \sum_{k=1}^{K} \sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2$	$n-K$	$CMR = \frac{SCR}{n-K}$	-
Total	$SCT = \sum_{i=1}^{n}(x_i - \bar{x})^2$	$n-1$	-	-

Tableau 1.1. Tableau d'analyse de la variance à 1 facteur

Il s'agit du rapport entre la variabilité expliquée et la variabilité résiduelle corrigée par les degrés de liberté. Sous H_0, et lorsque la variable sous-jacente X est gaussienne, F suit une loi de Fisher $\mathcal{F}(K-1, n-K)$ [12]. F est définie dans $\mathbb{R}^+$. La région critique du test au risque α s'écrit :

$$R.C. : F \geq F_{1-\alpha}(K-1, n-K)$$

1.3.4 Application numérique : comparaison des salaires selon la garantie demandée

Analyse graphique

Commençons toujours par une analyse graphique. Nous créons les boîtes à moustaches de "Salaire.Femme" pour chaque valeur de "Garantie.Supp" (Figure 1.4). Une analyse succincte montre qu'il semble y avoir décalage entre les distributions conditionnelles. Les dispersions également semblent un peu disparates. Tout cela demande à être confirmé numériquement. Par ailleurs, on constate qu'il a été demandé une *hypothèque* au ménage $n^o 27$ ayant un revenu particulièrement faible.

Mise en oeuvre du test

De nouveau, nous devons calculer les effectifs, moyennes et écarts type conditionnels (Figure 1.5 - A). Nous observons (pour "Garantie.Supp" : "caution - 1", "hypothèque - 2", "non - 3") pour la première modalité : $n_1 = 5$, $\bar{x}_1 = 7.3700$ et $s_1 = 0.4684$; pour la seconde, etc. Les effectifs ne sont pas équilibrés. L'écart type de "Salaire.Féminin" pour la modalité "non" de "Garantie.Supplémentaire" semble plus élevée que les autres. Nous analyserons cela dans le chapitre suivant.

Passons au tableau d'analyse de variance (Figure 1.5 - B). Nous pouvons déduire toutes les valeurs nécessaires au calcul de F à partir du tableau précédent (A) :

- $SCT = (n-1) \times s^2 = (50-1) \times 0.5099^2 = 12.7423$
- $SCE = \sum_k n_k(\bar{x}_k - \bar{x})^2 = 5 \times (7.3700 - 7.3094)^2 + 29 \times (7.1762 - 7.3094)^2 + 16 \times (7.5319 - 7.3094)^2 = 1.3248$
- $SCR = SCT - SCE = 12.7423 - 1.3248 = 11.4175$
- On en déduit ainsi $CME = \frac{1.3248}{3-1} = 0.6624$, $CMR = \frac{11.4175}{50-3} = 0.2429$ et $F = \frac{0.6624}{0.2429} = 2.7267$.
- Au risque 5%, nous devons comparer cette valeur avec le quantile $F_{0.95}(2, 47) = 3.1951$.
- *A priori*, les données ne contredisent pas l'hypothèse nulle. Les banques se gardent de demander une garantie spécifique selon le salaire de la chef de famille.
- La probabilité critique (p-value = 0.0758) renvoie à la même conclusion bien entendu.

12. Voir [11], page 354, concernant le détail de la formation de la distribution de F. Il s'agit du rapport de 2 distributions du χ^2 normalisées par les degrés de liberté

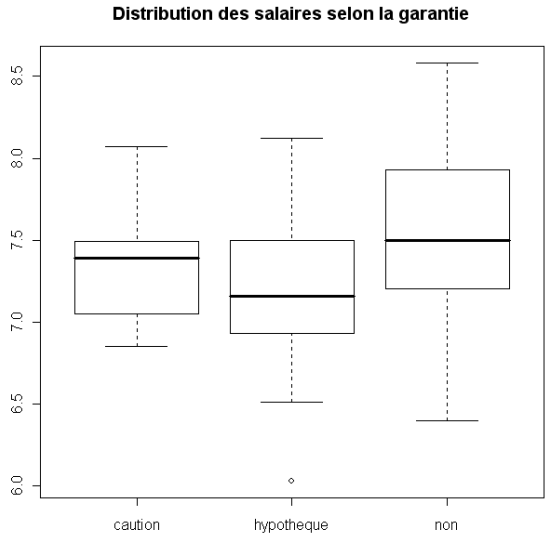


Fig. 1.4. Boîtes à moustaches des salaires féminins selon la garantie demandée

	A	B	C	D	E	F	G
1							
2	Effectifs, moyennes et ecart-types conditionnels						
3		Données					
4	Garantie.Supp	Nombre de Sal.Femme	Moyenne de Sal.Femme	Écartype de Sal.Femme			
5	caution	5	7.3700	0.4684			
6	hypotheque	29	7.1762	0.4615			
7	non	16	7.5319	0.5524			
8	Total	50	7.3094	0.5099			
9							
10							
11	Tableau d'analyse de variance						
12	Source	SC	ddl	CM	F	p-value	
13	Expliquée	1.3248	2	0.6624	2.7267	0.0758	
14	Résiduelle	11.4175	47	0.2429			
15	Totale	12.7423	49				

(A)

(B)

Fig. 1.5. ANOVA - Salaires féminins selon la garantie demandée

1.3.5 Robustesse de l'ANOVA

L'ANOVA est très robuste par rapport à l'hypothèse de normalité. Il suffit que les distributions conditionnelles aient des formes similaires, même asymétriques. Un graphique (boîte à moustaches ou histogramme de fréquences) permet de poser un diagnostic immédiat.

Il en est de même en ce qui concerne l'hypothèse d'homoscédasticité. Le test est d'autant plus robuste que les sous échantillons sont équilibrés. Dans ce cas, la variance conditionnelle la plus élevée peut être jusqu'à 4 fois supérieure à la plus petite variance ([7], page 363).

De manière générale, on gagne toujours à équilibrer les sous échantillons c.-à-d. faire de manière à ce que $n_1 = n_2 = \ldots = n_K$. Comme nous le disions plus haut, cela permet de réduire l'impact des variances inégales. Mais cela permet aussi de réduire le risque de 2e espèce du test. Par la suite, lorsque nous concluons au rejet de H_0 et qu'il va falloir décomposer les moyennes pour déceler les principales différences, les procédures subséquentes tirent parti de cette caractéristique (voir [6], page 37).

Remarque 2 (Une démarche conservatrice... mais simple). Enfin, si véritablement nous pensons que l'hétéroscédasticité peut entacher les résultats, nous pouvons conserver les calculs de l'ANOVA et modifier la région critique du test, à l'instar de ce qui a été présenté dans le test de comparaison de 2 moyennes. La démarche consiste à utiliser le F standard de l'ANOVA (équation 1.4), et de la comparer au quantile d'ordre $(1-\alpha)$ de la loi de Fisher à $(1, n-1)$ degrés de liberté. La région critique est ainsi réduite, nous favorisons fortement l'hypothèse nulle, nous ne risquons pas de la rejeter à tort intempestivement.

1.3.6 ANOVA avec variances conditionnelles hétérogènes

Lorsque nous cumulons deux obstacles, les variances sont manifestement hétérogènes et les effectifs sont déséquilibrés, nous pouvons adopter la procédure de Welch ([7], page 364 ; [12], pages 62 à 64) [13]. Il faut en revanche que la distribution sous-jacente de X soit gaussienne.

En anglais *Variance-weighted one-way ANOVA*, cette technique consiste à introduire une pondération particulière des données, dépendante des effectifs et de la variance des sous-groupes.

Statistique du test et région critique

Soit ω_k la pondération définie de la manière suivante :

$$\omega_k = \frac{n_k}{s_k^2}$$

La moyenne marginale s'écrit comme une moyenne pondérée des moyennes conditionnelles :

$$\bar{x}' = \frac{\sum_k \omega_k \bar{x}_k}{\sum_k \omega_k}$$

La statistique du test devient

$$F = \frac{\frac{\sum_k \omega_k (\bar{x}_k - \bar{x}')^2}{K-1}}{1 + \frac{2(K-2)}{K^2-1} \sum_k (\frac{1}{n_k - 1})(1 - \frac{\omega_k}{\sum_k \omega_k})^2} \tag{1.5}$$

Sous H_0, F suit une loi de Fisher à $(K-1, \nu)$ degrés de liberté avec

$$\nu = \frac{K^2 - 1}{3 \sum_k (\frac{1}{n_k - 1})(1 - \frac{\omega_k}{\sum_k \omega_k})^2}$$

La forme de la région critique du test n'est pas modifiée. Au risque α, on rejette l'hypothèse nulle si

$$R.C. : F > F_{1-\alpha}(K-1, \nu)$$

	A	B	C	D	E
1					
2	Effectifs, moyennes et ecart-types conditionnels				
3		Données			
4	Garantie.Supp	Nombre de Sal.Femme	Moyenne de Sal.Femme	Écartype de Sal.Femme	omega_k
5	caution	5	7.3700	0.4684	22.789
6	hypotheque	29	7.1762	0.4615	136.185
7	non	16	7.5319	0.5524	52.431
8	Total	50	7.3094	0.5099	211.405
9					
10					
11		x_barre'	7.2853		
12					
13		A	4.9720		
14		B	0.2412		
15					
16		F	2.3446		
17					
18		dd1	2		
19		ddl2	11.05		
20					
21		F_0.95(2, 11)	3.9823		
22					
23		F_0.9(2, 11)	2.8595		

Fig. 1.6. ANOVA de Welch - Salaires féminins selon la garantie demandée

Application : garantie exigée selon le salaire féminin

Reprenons notre exemple de la garantie demandée en fonction du salaire de la chef de famille. Nous y retrouvons les conditions d'application de la technique : les effectifs conditionnels sont relativement déséquilibrés, et il y a suspicion d'hétéroscédasticité (*dixit* l'analyse graphique avec les boxplot).

Complétons la feuille de calcul de manière à intégrer les nouveaux calculs (Figure 1.6) :

- Le calcul des effectifs, moyennes et écarts type conditionnels nous est familier maintenant.
- Nous rajoutons une nouvelle colonne dédiée aux ω_k. Pour la première valeur par exemple, nous avons $\omega_1 = \frac{5}{0.4684^2} = 22.789$. En faisant la somme, nous obtenons $\sum_k \omega_k = 211.405$.
- La moyenne pondérée est $\bar{x}' = \frac{22.789 \times 7.3700 + 136.185 \times 7.1762 + 52.431 \times 7.5319}{211.405} = 7.2853$
- Nous formons la quantité A, elle est égale à $A = \sum_k \omega_k (\bar{x}_k - \bar{x}')^2 = 4.9720$
- Puis B, avec $B = \sum_k (\frac{1}{n_k - 1})(1 - \frac{\omega_k}{\sum_k \omega_k})^2 = 0.2412$
- Nous pouvons en déduire $F = \dfrac{\frac{A}{3-1}}{1 + \frac{2 \times (3-2)}{3^2-1} \times B} = 2.3446$
- Pour obtenir le seuil critique au risque α, nous devons disposer des degrés de liberté, le premier $ddl_1 = K - 1 = 3 - 1 = 2$ ne pose pas de difficultés. Le second $ddl_2 = \nu$ demande quelques calculs supplémentaires $\nu = \frac{3^2-1}{3 \times B} = 11.05$
- Le second degré de liberté étant fractionnaire, en toute rigueur nous devrions calculer le véritable seuil critique par interpolation, à l'instar de ce qui a été fait pour le test de comparaison de 2 moyennes avec variances inégales (qui est aussi de Welch d'ailleurs) (section 1.2.3). Dans notre exemple, nous simplifions quelque peu la chose en prenant directement la valeur entière la plus

13. Note : Assez curieusement, il est très difficile de trouver de la documentation concernant cette méthode sur le web. Invariablement, les moteurs de recherche nous ramènent vers la documentation du logiciel SAS.

proche du degré de liberté, soit $\nu = 11$. Cela nous facilitera la vie quand nous aurons à effectuer plusieurs comparaisons associées à des niveaux de signification différents.

- Au risque $\alpha = 5\%$, le seuil critique est $F_{0.95}(2, 11) = 3.9823$. L'hypothèse nulle d'égalité des moyennes conditionnelles est compatible avec nos données.
- Si nous passons à $\alpha = 10\%$, le seuil critique est $F_{0.90}(2, 11) = 2.8595$. Nous aboutissons également à l'acceptation de H_0, à la différence de l'ANOVA standard qui, avec une p-value de 0.0758 rejetait l'hypothèse nulle (Figure 1.5). Il y a différence de comportement des tests dans ce cas.

Encore une fois, c'est un peu se compliquer la vie tout ça. Si nous avons prise sur les conditions de recueil des données, nous avons tout intérêt à produire des sous échantillons équilibrés. Nous pouvons ainsi utiliser en toute confiance l'ANOVA à 1 facteur standard, connue de tout le monde. Nous nous dispensons ainsi d'une difficulté supplémentaire lors de la présentation des résultats : avoir à expliquer les subtilités des méthodes, toujours un peu périlleuse face à des non-statisticiens.

1.3.7 Que faire suite au rejet de l'hypothèse nulle ? Les comparaisons multiples

Une fois que nous avons rejeté l'hypothèse nulle d'égalité des moyennes, nous savons que 2 des moyennes au moins sont différentes. Mais lesquelles ? Quelles sont les moyennes qui sont différentes ? Quels sont les écarts les plus importants ? Quelles sont celles qui s'écartent significativement d'une moyenne de référence ?

Pour répondre à ces questions, on procède aux comparaisons multiples des moyennes. Ce thème dépasse largement le propos de ce support. Il pourrait même faire l'objet d'un support à part tant le domaine est fertile. Pour notre part, nous nous contenterons de conseiller la lecture de l'excellent document en ligne du NIST - `http://www.itl.nist.gov/div898/handbook/prc/section4/prc47.htm`.

2

Comparaison de variances

Les tests présentés dans le chapitre précédent (chapitre 1) sont subordonnés à l'égalité des variances conditionnelles. La vérification de cette propriété semble donc un préalable nécessaire, même si par ailleurs nous avons montré que, sous certaines conditions, les tests de comparaison de moyennes pouvaient se révéler très robustes. Lorsque ces conditions ne sont pas réunies, déterminer une différence significative entre les variances nous permet de choisir en connaissance de cause les procédures appropriées (sections 1.2.3 et 1.3.6).

Ce n'est pas le seul usage des tests d'égalité des variances. Comparer la variabilité dans les sous-groupes peut être la finalité intrinsèque d'une étude : comparer la variance des notes des étudiants en fonction de leur disposition dans la salle de classe (en cercle, en rangées, etc.) ; comparer la variance de la taille des pièces produites par différentes machines ; etc.

De manière générale, nous considérons que X suit une loi normale dans ce chapitre. Certains tests sont très sensibles à cette propriété, d'autres en revanche sont plus robustes. Nous en discuterons lors de la description des techniques.

2.1 Estimation de la variance et distribution

Soit la v.a. X distribuée normalement de paramètres $\mathcal{N}(\mu, \sigma)$. On veut estimer σ à partir d'un échantillon de taille n. Dans un premier temps, on considère que μ est connu, nous utilisons la quantité s'^2 pour estimer σ^2 :

$$s'^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

On montre facilement que $\frac{ns'^2}{\sigma^2}$ suit une loi du $\chi^2(n)$ à n degrés de liberté. En effet,

$$\frac{ns'^2}{\sigma^2} = \sum_i (\frac{x_i - \mu)}{\sigma})^2 \equiv \sum_{i=1}^{n}[\mathcal{N}(0,1)]^2 \equiv \chi^2(n)$$

Nous pouvons également écrire

$$\frac{s'^2}{\sigma^2} \equiv \frac{\chi^2(n)}{n}$$

Cette configuration est purement théorique. Dans la pratique, nous devons estimer l'espérance mathématique μ à l'aide de la moyenne empirique $\bar{x}$, **l'estimateur sans biais** de la variance s'écrit

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

On montre alors que

$$\frac{s^2}{\sigma^2} \equiv \frac{\chi^2(n-1)}{n-1} \tag{2.1}$$

Ce résultat est très important, nous l'utiliserons constamment dans ce chapitre. En réalité, sans y faire référence explicitement, il a été mis à contribution pour calculer les lois de statistiques des tests de comparaison des moyennes du chapitre précédent.

2.2 Comparaison de 2 variances - Test de Fisher

2.2.1 Test d'hypothèses, statistique et région critique

Le test de comparaison de Fisher compare les variances de 2 sous populations, il confronte les hypothèses suivantes :

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

A partir des résultats de la section précédente (section 2.1), la statistique du test calcule le rapport entre les variances estimées dans chaque sous-échantillon [1].

Si F s'éloigne significativement de la valeur 1, on peut considérer que les variances conditionnelles sont différentes. Formellement, sous H_0, il a été établi que F suit une loi de Fisher à (ν_1, ν_2) degrés de liberté [2] avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$. La région critique du test au risque α s'écrit alors

$$R.C. : F \leq F_{\frac{\alpha}{2}}(n_1 - 1, n_2, -1) \quad \text{ou} \quad F \geq F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2, -1)$$

$F_a(n_1 - 1, n_2 - 1)$ est le quantile de d'ordre a de la loi de Fisher à $(n_1 - 1, n_2 - 1)$ degrés de liberté.

	A	B	C
1			
2			
3		Données	
4	Acceptation	Nombre de Sal.Femme	Var de Sal.Femme
5	oui	34	0.3006
6	non	16	0.0684
7	Total	50	0.2600
8			
9		F	4.3962
10			
11		ddl1	15
12		ddl2	33
13			
14		F_0.025(15,33)	0.3813
15		F_0.975(15,33)	2.2610

Fig. 2.1. Test d'homogénéité de 2 variances - Salaire vs. Acceptation

2.2.2 Variabilité des salaires féminins selon l'acceptation du crédit

Nous avons là l'occasion de vérifier l'hypothèse d'homoscédasticité mise en avant dans nos illustrations des tests de comparaisons de moyennes. Nous voulons comparer les variances des salaires féminins selon l'acceptation du crédit (à rapprocher avec le test de comparaison de moyennes, cf. section 1.2.2).

Les calculs sont très simples, l'outil "Tableau croisé dynamique" d'EXCEL nous fournit directement les variances conditionnelles (Figure 2.1) :

- Les effectifs ne sont pas modifiés par rapport à l'analyse précédente (Figure 1.2), à savoir $n_1 = 34$ et $n_2 = 16$. Nous en déduisons directement $\nu_1 = 34 - 1 = 33$ et $\nu_2 = 16 - 1 = 15$
- Nous obtenons $s_1^2 = 0.3006$ et $s_2^2 = 0.0684$
- Nous construisons le rapport $F = \frac{0.3006}{0.0684} = 4.3962$
- Au risque 5%, le seuil critique inférieur est $F_{0.025}(15, 33) = 0.3813$, le seuil supérieur est $F_{0.975}(15, 33) = 2.2610$. La valeur calculée n'est pas dans l'intervalle $]0.3813; 2.2610[$, on conclut au rejet de l'hypothèse nulle, les variances sont significativement différentes.
- Il semble donc que lors du traitement de cet exemple dans les sections consacrées aux comparaisons des moyennes (sections 1.2.2 et 1.2.3), il fallait opter pour le test adapté aux cas des variances inégales.

Remarque 3 (Obtenir les valeurs critiques de la loi de Fisher). La loi de Fisher est souvent tabulée pour les quantiles d'ordre élevé ($a = 0.95, 0.975, 0.99, \ldots$). Pour obtenir le seuil inférieur, nous pouvons utiliser la propriété suivante

$$F_a(\nu_1, \nu_2) = \frac{1}{F_{1-a}(\nu_2, \nu_1)}$$

1. voir `http://www.itl.nist.gov/div898/handbook/prc/section3/prc32.htm` et `http://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm`
2. Bien évidemment, puisqu'il s'agit du rapport de lois de χ^2 normalisées par leurs degrés de liberté (Equation 2.1

2.2.3 Robustesse et pratique du test F pour la comparaison de variance

Le test de comparaison de 2 variances de Fisher n'est pas robuste du tout. Un écart, même minime, de la distribution normale fausse les résultats. Il faut absolument s'assurer du caractère gaussien de X avant de l'utiliser. Ce qui en limite considérablement la portée. Dans la pratique, on se tournera avantageusement vers les autres techniques présentées dans ce chapitre.

Pourtant le test de Fisher est systématiquement présenté dans les ouvrages, il est également disponible dans la très grande majorité des logiciels. Peut être parce que les concepts à manipuler sont accessibles, calculer le rapport entre 2 variances est une opération très simple.

2.3 Comparaison de K variances - Test de Bartlett

2.3.1 Test, statistique du test et région critique

Le test de Bartlett sert à éprouver l'homogénéité de K variances. C'est une généralisation du test de Fisher. Lui également repose pesamment sur la normalité des données. Un faible écart par rapport à cette propriété remet en cause fortement ses résultats, le test de Bartlett n'est absolument pas robuste [3]. Il n'est vraiment performant que si l'hypothèse de normalité est établie. Dans la pratique, on lui préférera les techniques présentées plus loin dans ce chapitre.

Pourtant il est très largement répandu, utilisé dans les études et disponible dans le logiciel. Peut être faut-il y voir encore une fois des raisons purement historiques. Le test de Bartlett est largement antérieur aux autres.

Les hypothèses à confronter sont :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$$
$$H_1 : \text{au moins 2 variances sont différentes}$$

La statistique du test s'écrit [4]

$$T = \frac{(n-K)\ln s_p^2 - \sum_{k=1}^{K}(n_k - 1)\ln s_k^2}{1 + \frac{1}{3(K-1)}\left(\sum_{k=1}^{K}\frac{1}{n_k-1} - \frac{1}{n-K}\right)} \tag{2.2}$$

Où s_p^2 est l'estimateur non biaisé de σ^2, il s'agit de la variance intra-classes :

$$s_p^2 = \frac{\sum_k (n_k - 1)s_k^2}{n-K}$$

3. http://en.wikipedia.org/wiki/Bartlett's_test
4. http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm

Sous H_0, T suit une loi du $\chi^2(K-1)$ à $(K-1)$ degrés de liberté. L'approximation par la loi du χ^2 est satisfaisante dès lors que $n_k \geq 5$ $(\forall k)$ et K petit par rapport à n.

La région critique du test au risque α s'écrit

$$R.C. : T \geq \chi^2_{1-\alpha}(K-1)$$

$\chi^2_{1-\alpha}(K-1)$ est le quantile d'ordre $(1-\alpha)$ de la loi du $\chi^2(K-1)$.

Remarque 4 (Une écriture simplifiée du test de Bartlett). Dans certaines références, on ne retient que le numérateur de l'équation 2.2. On sait dans ce cas que les résultats sont légèrement biaisés (voir `http://www.itl.nist.gov/div898/handbook/prc/section4/prc42.htm`). Le dénominateur de l'équation 2.2 doit être compris comme un facteur de correction.

2.3.2 Exemple : variance des salaires selon les garanties demandées

Reprenons notre exemple de la section 1.3.4 : nous souhaitons comparer les variances des salaires féminins selon la garantie demandée par les banques. On ne peut pas dire que l'interprétation des résultats d'une telle étude soit réellement transcendante. On dira ici qu'il s'agit d'un test préparatoire pour l'ANOVA.

	A	B	C	D	E	F	G
1		Données					
2	Garantie.Supp	Nombre de Sal.Femme	Var de Sal.Femme		n_k-1	1/n_k	ln(s^2_k)
3	caution	5	0.2194		4	0.2000	-1.5169
4	hypotheque	29	0.2129		28	0.0345	-1.5467
5	non	16	0.3052		15	0.0625	-1.1869
6	Total	50	0.2600				
7							
8		s^2_p	0.2429				
9							
10		A	-66.5049				
11		B	-67.1792				
12		A - B	0.6743				
13							
14		C	1.0552				
15							
16		T	0.6390				
17							
18		KHI-2(0.95 ; 2)	5.9915				
19							
20		p-value	0.7265				

Fig. 2.2. Test de Bartlett - Salaire vs. Garantie

Si la formule 2.2 semble rébarbative, sa mise en oeuvre dans un tableur est finalement assez simple (Figure 2.2) :

- Il nous faut une petite phase préparatoire avant de procéder aux calculs. En colonnes E, F et G, nous calculons respectivement $n_k - 1$, $\frac{1}{n_k}$ et $\ln(s_k^2)$
- Nous pouvons calculer la variance intra-classes $s_p^2 = \frac{4\times0.2194+28\times0.2129+15\times0.3052}{50-3} = 0.2429$
- Ensuite nous calculons la quantité A qui est la première partie du numérateur de l'équation 2.2, à savoir $A = (50-3) \times \ln(0.2429) = -66.5049$

- Puis la seconde partie du numérateur $B = 4\times(-1.5169)+28\times(-1.5467)+15\times(-1.1869) = -67.1792$
- L'écart $A - B = 0.6743$
- Passons maintenant au facteur de correction au dénominateur. Nous formons $C = 1 + \frac{1}{3(3-1)} \times ((0.2 + 0.0345 + 0.0625) - \frac{1}{50-3}) = 1.0552$. La correction est légère.
- Nous en déduisons ainsi $T = \frac{A-B}{C} = \frac{0.6743}{1.0522} = 0.6390$
- A partir de la loi du χ^2, nous obtenons le seuil critique du test pour un risque α : $\chi^2_{0.95}(3-1) = 5.9915$. Nous sommes dans la région d'acceptation de H_0, les variances ne sont pas significativement différentes d'un groupe à l'autre.
- La probabilité critique (p-value = 0.7265) aboutit bien évidemment à la même conclusion.

2.4 Comparaison de K variances - Test de Cochran et test de Hartley

Les tests qui viennent cumulent les désavantages. Néanmoins, nous les présentons car on les retrouve parfois dans les études et ils sont disponibles dans les logiciels. Il faut donc les connaître pour ne pas être pris au dépourvu si nous les rencontrons. Leur utilisation n'est pas vraiment conseillée.

Deux conditions doivent être réunies pour utiliser ces tests : la distribution sous-jacente de X doit être normale, les procédures sont peu robustes par rapport à cette hypothèse ; les effectifs doivent être parfaitement équilibrés c.-à-d. $n_1 = n_2 = \ldots = n_K = m$. Une contrainte supplémentaire vient se rajouter : les statistiques de test ne suivent pas une loi de probabilité d'usage courant, elles sont tabulées spécifiquement.

Test de Hartley (ou Test Fmax)

La statistique du test repose sur le rapport entre la plus grande variance et la plus petite variance conditionnelle

$$H = \frac{\hat{\sigma}^2_{max}}{\hat{\sigma}^2_{min}}$$

La région critique du test au niveau de signification α est définie comme suit

$$H \geq H_{1-\alpha}(K, m-1)$$

Où le seuil critique $H_{1-\alpha}(K, m-1)$ est lue dans une table spécifique,où K et $m-1$ sont les degrés de liberté [5].

5. Voir par exemple `http://www.gseis.ucla.edu/courses/help/fmax.html` pour un risque $\alpha = 5\%$

Test de Cochran

Le test de Cochran utilise le rapport entre la variance conditionnelle maximale et la somme non pondérée des variances :

$$C = \frac{\hat{\sigma}^2_{max}}{\sum_k \hat{\sigma}^2_k}$$

La région critique s'écrit

$$C \geq C_{1-\alpha}(K, m-1)$$

Où le seuil critique $C_{1-\alpha}(K, m-1)$, avec K et $m-1$ les degrés de liberté, est lue dans une table spécifique (Voir [6], Table 5, pour les risques $\alpha = 0.05$ et $\alpha = 0.01$).

2.5 Comparaison de K variances - Test de Levene

2.5.1 Principe, statistique de test, région critique

Le test de Levene est une alternative crédible du test de Bartlett (et de Fisher). Il **est robuste** c.-à-d. il est moins sensible à un écart par rapport à l'hypothèse de normalité. De fait, si la distribution sous jacente de X n'est pas gaussienne : il aura moins tendance à détecter des faux positifs (conclure à l'inégalité des variances alors que l'hypothèse nulle est vraie) ; et il sera plus apte à détecter les vrais positifs (conclure à juste titre à l'inégalité des variances) [6].

Pour tester l'homogénéité des variances dans K groupes, le test de Levene procède en 2 temps. Une transformation des variables est tout d'abord opérée, nous calculons

$$z_{ik} = |x_{ik} - \bar{x}_k| \tag{2.3}$$

où $\bar{x}_k$ est la moyenne des valeurs dans le sous-échantillon Ω_k.

Puis, dans un second temps, la statistique W est calculée

$$W = \frac{(n-K)\sum_{k=1}^{K} n_k(\bar{z}_k - \bar{z})^2}{(K-1)\sum_{k=1}^{K}\sum_{i=1}^{n_k}(z_{ik} - \bar{z}_k)^2} \tag{2.4}$$

A y regarder de plus près, nous nous rendons compte qu'il s'agit là du rapport entre les carrés moyens expliqués et les carrés moyens résiduels calculés sur les valeurs z_{ik} : le test de Levene est donc une analyse de variance sur la variable transformée.

Sous H_0, W suit une loi de Fisher à $(K-1, n-K)$ degrés de liberté. La région critique au risque α est définie tout naturellement de la manière suivante :

$$R.C. : W \geq F_{1-\alpha}(K-1, n-K)$$

Avec $F_{1-\alpha}(K-1, n-K)$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher.

6. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm`

2.5.2 Application sur la variance des salaires selon la garantie

La mise en oeuvre du test de Levene dans un tableur demande une petite préparation. Nous devons notamment calculer les données transformées z_{ik}. Par la suite, nous retrouvons le schéma de l'ANOVA avec les calculs des effectifs et moyennes conditionnelles, la somme des carrés des écarts, etc. (Figure 2.3).

	A	B	C	D	E	F	G	H
1								
2		Sal.Femme	Garantie.Supp			caution	hypotheque	non
3		7.49	caution		moyenne.cond	7.37	7.18	7.53
4		7.39	caution				z	
5		8.07	caution			0.12	0.5437931	0.431875
6		7.05	caution			0.02	0.4162069	0.281875
7		6.85	caution			0.7	0.6662069	0.061875
8		7.72	hypotheque			0.32	0.2462069	0.478125
9		6.76	hypotheque			0.52	0.3337931	0.091875
10		6.51	hypotheque				0.5762069	0.058125
11		6.93	hypotheque				0.3862069	0.281875
12		7.51	hypotheque				0.3237931	0.238125
13		6.60	hypotheque				0.1162069	1.131875
14		6.79	hypotheque				0.1137931	0.001875
15		7.50	hypotheque				0.2037931	0.408125
16		7.06	hypotheque				0.5137931	0.388125
17		7.29	hypotheque				0.3162069	1.048125
18		7.38	hypotheque				0.5937931	0.371875
19		7.69	hypotheque				0.0362069	0.758125
20		6.86	hypotheque				1.1462069	0.721875
21		7.77	hypotheque				0.6037931	
22		7.14	hypotheque				0.9437931	
23		6.03	hypotheque				0.0562069	
24		7.78	hypotheque				0.0237931	
25		8.12	hypotheque				0.0237931	
26		7.12	hypotheque				0.6737931	
27		7.20	hypotheque				0.6062069	
28		7.20	hypotheque				0.2062069	
29		7.85	hypotheque				0.2162069	
30		6.57	hypotheque				0.0662069	
31		6.97	hypotheque				0.0637931	
32		6.96	hypotheque				0.1137931	
33		7.11	hypotheque				0.0162069	
34		7.24	hypotheque					
35		7.29	hypotheque		n_k	5	29	16
36		7.16	hypotheque		z_barre_k	0.3360	0.3499	0.4221
37		7.10	non		somme.carrés	0.3131	2.4126	1.7266
38		7.25	non					
39		7.47	non					
40		8.01	non			z_barre	0.3716	
41		7.44	non					
42		7.59	non			A	0.0609	
43		7.25	non			B	4.4524	
44		7.77	non					
45		6.40	non			W	0.3212	
46		7.53	non					
47		7.94	non			ddl1	2	
48		7.92	non			ddl2	47	
49		8.58	non					
50		7.16	non			F_0.95(2,47)	3.1951	
51		8.29	non					
52		6.81	non			p-value	0.7269	

Fig. 2.3. Test de Levene - Salaire vs. Garantie

Détaillons cela :

- Tout d'abord, les données ont été triées selon la variable "Garantie.Sup" afin d'en faciliter la lecture (colonnes B et C).
- Nous calculons les moyennes conditionnelles des $\bar{x}_k$ que nous utilisons pour centrer les données *à l'intérieur des sous-échantillons*. Nous avons $\bar{x}_1 = 7.37$, $\bar{x}_2 = 7.18$ et $\bar{x}_3 = 7.53$. En colonne F, G et H (ligne 5 à ligne 33), nous formons les séries z_{ik} selon la formule 2.3.

- Pour chaque sous échantillon, nous calculons n_k (F35 à H35), $\bar{z}_k$ (F36 à H36) et $\sum_{i=1}^{n_k}(z_{ik}-\bar{z}_k)^2$ (F37 à H37). Ainsi, pour la première modalité "garantie = caution", nous avons $n_1 = 5$, $\bar{z}_1 = 0.3360$ et $\sum_{i=1}^{5}(z_{i1}-\bar{z}_1)^2 = 0.3131$
- La moyenne marginale des z_{ik} est $\bar{z} = 0.3716$
- Nous formons la quantité $A = \sum_{k=1}^{K} n_k(\bar{z}_k - \bar{z})^2 = 5\times(0.3360-0.3716)^2 + 29\times(0.3499-0.3716)^2 + 16\times(0.4221-0.3716)^2 = 0.0609$
- Puis la quantité $B = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(z_{ik}-\bar{z}_k)^2 = 0.3131 + 2.4126 + 1.7266 = 4.4524$
- Reste à former le rapport pour obtenir la statistique de Levene

$$W = \frac{(50-3)\times A}{(3-1)\times B} = \frac{(50-3)\times 0.0609}{(3-1)\times 4.4524} = 0.3212$$

- Au risque $\alpha = 5\%$, nous le comparons au seuil critique $F_{0.95}(2,47) = 3.1951$
- Manifestement, les variances sont identiques dans les sous-groupes.
- La probabilité critique (p-value = 0.7269) confirme cela.

Nous remarquerons la similitude des résultats avec ceux du test de Bartlett (concernant la p-value du test principalement). La normalité des données étant crédible selon les tests réalisés en annexe **??**, les résultats sont proches, c'est assez logique. Dans le cas contraire, distributions non gaussiennes, les résultats peuvent être très différents voire contradictoires. On privilégiera alors systématiquement le test de Levene.

2.6 Comparaison de K variances - Test de Brown-Forsythe

2.6.1 Principe, statistique de test, région critique

Le test de Brown-Forsythe est une généralisation du test de Levene. Il en précise les conditions de robustesse. En effet, la formule de transformation de données z_{ik} est mise en relation avec les caractéristiques des distributions [7] :

- La transformation originelle n'est vraiment performante que si X est symétrique, avec une queue de distribution modérée. On prendra alors

$$z_{ik} = |x_{ik} - \bar{x}_k|$$

- Lorsque la distribution est à queue lourde (loi de Cauchy par exemple), nous aurons intérêt à utiliser un autre type de transformation

$$z_{ik} = |x_{ik} - \bar{x}'_k|$$

où $\bar{x}'_k$ est la moyenne des données comprises entre le quantile d'ordre 0.05 et le quantile d'ordre 0.95. En d'autres termes, la moyenne des données pour lesquelles nous aurons retiré 5% des valeurs les plus basses, et 5% des valeurs les plus élevées. L'idée bien entendu est de *lisser* les données en retirant les valeurs extrêmes, la moyenne n'en sera que plus robuste.

7. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm`

– Lorsque la distribution est asymétrique à gauche ($\chi^2(4)$ par exemple). On aura intérêt à passer par une autre transformation

$$z_{ik} = |x_{ik} - \tilde{x}_k|$$

où $\tilde{x}_k$ est la médiane conditionnelle.
Lorsque nous n'avons pas de connaissances précises sur les distributions, cette transformation est conseillée. **Elle réalise le meilleur compromis quelle que soit la distribution sous-jacente de X. C'est la procédure à utiliser en priorité pour tester l'homogénéité des variances dans un contexte générique**.

2.6.2 Application sur la variance des salaires selon la garantie

Sur les mêmes données que précédemment, comparer la variance des salaires féminins selon la garantie, nous mettons en oeuvre la variante de Brown-Forsythe basée sur la médiane. L'organisation de la feuille de calcul (Figure 2.4) est exactement la même que celle du test de Levene (section 2.5.2), à la différence que nous utilisons la médiane pour transformer les données (cellules F3 à H3) : $\tilde{x}_1 = 7.39$, $\tilde{x}_2 = 7.16$ et $\tilde{x}_3 = 7.50$.

Au final, nous obtenons $W = 0.3268$ avec une probabilité critique (p-value = 0.7229). Les résultats de cette variante sont très proches de ceux du test de Levene et du test de Bartlett sur nos données.

Dans notre contexte, la répartition de la variable d'intérêt étant compatible avec la loi normale, le test de Bartlett est certainement le plus approprié. Pourtant, on remarquera que les autres méthodes, censées plus performantes pour les autres types de distributions (asymétriques ou à queue lourde) ne s'effondrent pas et donnent des résultats cohérents. Ce qui milite encore une fois pour leur utilisation, notamment la variante de Brown-Forsythe basée sur la médiane conditionnelle, quelle que soit la loi sous-jacente des données.

2.6.3 Une autre variante - Le test de O'Brien

Le test de O'Brien est une autre variante du test de Levene. Encore une fois, il s'agit de convertir les données de manière à ce qu'elles reflètent la variabilité des valeurs originelles [8]. Puis, comme dans le test de Levene, pour détecter l'hétérogénéité des variances, nous réalisons une comparaison de moyennes (ANOVA) sur les valeurs transformées.

Son principal intérêt est que la transformation est paramétrée. Nous disposons d'un outil supplémentaire pour améliorer l'adéquation du test avec la véritable nature des données.

Les valeurs z_{ik} sont maintenant définies comme suit :

$$z_{ik} = \frac{(\omega + n_k - 2)n_k(x_{ik} - \bar{x}_k)^2 - \omega s_k^2(n_k - 1)}{(n_k - 1)(n_k - 2)} \tag{2.5}$$

8. `http://www.utd.edu/~herve/Abdi-OBrien2007-pretty.pdf`

	A	B	C	D	E	F	G	H
1								
2		Sal.Femme	Garantie.Supp			caution	hypotheque	non
3		7.49	caution		mediane.cond	7.39	7.16	7.50
4		7.39	caution				z	
5		8.07	caution			0.1	0.56	0.4
6		7.05	caution			0	0.4	0.25
7		6.85	caution			0.68	0.65	0.03
8		7.72	hypotheque			0.34	0.23	0.51
9		6.76	hypotheque			0.54	0.35	0.06
10		6.51	hypotheque				0.56	0.09
11		6.93	hypotheque				0.37	0.25
12		7.51	hypotheque				0.34	0.27
13		6.60	hypotheque				0.1	1.1
14		6.79	hypotheque				0.13	0.03
15		7.50	hypotheque				0.22	0.44
16		7.06	hypotheque				0.53	0.42
17		7.29	hypotheque				0.3	1.08
18		7.38	hypotheque				0.61	0.34
19		7.69	hypotheque				0.02	0.79
20		6.86	hypotheque				1.13	0.69
21		7.77	hypotheque				0.62	
22		7.14	hypotheque				0.96	
23		6.03	hypotheque				0.04	
24		7.78	hypotheque				0.04	
25		8.12	hypotheque				0.04	
26		7.12	hypotheque				0.69	
27		7.20	hypotheque				0.59	
28		7.20	hypotheque				0.19	
29		7.85	hypotheque				0.2	
30		6.57	hypotheque				0.05	
31		6.97	hypotheque				0.08	
32		6.96	hypotheque				0.13	
33		7.11	hypotheque				0	
34		7.24	hypotheque					
35		7.29	hypotheque		n_k	5	29	16
36		7.16	hypotheque		z_barre_k	0.3320	0.3493	0.4219
37		7.10	non		somme.carrés	0.3285	2.4316	1.7460
38		7.25	non					
39		7.47	non					
40		8.01	non			z_barre	0.3708	
41		7.44	non					
42		7.59	non			A	0.0627	
43		7.25	non			B	4.5061	
44		7.77	non					
45		6.40	non			W	0.3268	
46		7.53	non					
47		7.94	non			ddl1	2	
48		7.92	non			ddl2	47	
49		8.58	non					
50		7.16	non			F_0.95(2,47)	3.1951	
51		8.29	non					
52		6.81	non			p-value	0.7229	

Fig. 2.4. Test de Brown-Forsythe - Salaire vs. Garantie

Nous pouvons fixer la valeur de ω. Les auteurs proposent la valeur $\omega = 0.5$ par défaut (voir par exemple [7], page 227). Mais nous pouvons faire mieux en la modulant de manière à ce que les caractéristiques de la distribution des z_{ik} concorde avec celles de la variable originelle X.

Au final, il semble que le choix de ω ne soit pas très décisif. Le test de O'Brien d'ailleurs n'est pas plus performant que les autres variantes du test de Levene[9].

9. Voir `http://v8doc.sas.com/sashtml/stat/chap30/sect37.htm`

3

Comparaison de proportions

3.1 Comparaison de 2 proportions

Dans ce chapitre, nous considérons que X est binaire, définie dans $\{1;0\}$. Elle prend la valeur 1 lorsque l'individu possède la caractéristique qui nous intéresse. Bien évidemment, la variable X ne peut pas être gaussienne, néanmoins nous **restons dans un cadre paramétrique** car X est distribuée selon une loi de Bernoulli $\mathcal{B}(1,p)$ où p est la paramètre que nous manipulons. La connaissance de p définit totalement la distribution. Lors des tests de comparaisons, ce sont les estimations de p dans chaque sous échantillon que nous confronterons.

3.1.1 Estimation et distribution

Nous disposons d'un échantillon de taille n, les individus sont tirés de manière indépendante. La statistique S

$$S = \sum_{i=1}^{n} x_i$$

suit une loi binomiale de paramètres $\mathcal{B}(n,p)$, d'espérance $E(S) = np$ et de variance $\sigma_S^2 = np(1-p)$. Lorsque n est suffisamment élevé, le théorème central limite s'applique, S tend vers la loi normale. En pratique, on juge que l'approximation est bonne dès que $np(1-p) > 9$ (voir [5], page 264).

Un estimateur sans biais de p est la fréquence observée

$$f = \frac{S}{n}$$

avec cette fois-ci $E(f) = p$ et $\sigma_f^2 = \frac{p(1-p)}{n}$.

En y regardant de plus près, on constate que f est une moyenne calculée sur une variable codée 0/1. Toujours sous les conditions ci-dessus, la distribution de f est asymptotiquement gaussienne

$$f \approxeq \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Nous utiliserons abondamment ces résultats dans les sections consacrées aux tests.

3.1.2 Test de comparaison, statistique du test et région critique

Nous disposons de 2 échantillons Ω_1 et Ω_2, nous souhaitons savoir si la proportion des individus portant le caractère étudié est la même dans les sous populations. Auquel cas, les populations sont homogènes du point de vue de la variable d'intérêt. Les hypothèses à confronter sont

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

Nous formons la statistique $D = f_1 - f_2$. A la lumière de la section précédente, elle suit asymptotiquement une loi normale d'espérance

$$E(D) = p_1 - p_2$$

Et de variance

$$V(D) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Sous $H_0 : p_1 = p_2 = p$, la proportion est la même dans les sous-populations, les caractéristiques deviennent $E(D) = 0$ et

$$V(D) = p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})$$

Il ne reste plus qu'à passer aux estimations des proportions p à partir des fréquences f. Pour les fréquences conditionnelles, nous produisons

$$\hat{p}_k = f_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}$$

Et pour la fréquence globale,

$$\hat{p} = f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Statistique du test

Sous H_0, et lorsque n_1 et n_2 sont suffisamment élevés, la statistique du test U suit une loi normale centrée réduite :

$$U = \frac{D}{\hat{\sigma}_D} = \frac{f_1 - f_2}{\sqrt{f(1-f)(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{3.1}$$

Région critique

Pour un test bilatéral au risque α, l'hypothèse nulle est rejetée lorsque

$$R.C. : |U| \geq u_{1-\frac{\alpha}{2}}$$

3.1.3 Un exemple numérique : proportion de l'acceptation selon le type d'emploi

	A	B	C	D	E
1					
2					
3		Données	Emploi		
4		Nombre de Acceptation2		Nombre de Acceptation	
5	Acceptation	cdd	cdi	cdd	cdi
6	oui	6	28	37.50%	82.35%
7	non	10	6	62.50%	17.65%
8	Total	16	34	100.00%	100.00%
9					
10					
11		f1	0.3750		
12		f2	0.8235		
13					
14		f	0.6800		
15					
16		D	-0.4485		
17		sigma	0.1414		
18					
19		U	-3.1716		
20					
21		u_0.975	1.9600		
22					
23		p-value	0.0015		

Fig. 3.1. Comparaison des proportions - Acceptation selon emploi

Toujours à partir de notre fichier de travail (Figure 0.1), nous souhaitons savoir si la proportion des dossiers acceptés est différent selon le type d'emploi de la personne de référence du dossier. Avec l'outil "Tableaux croisés dynamiques", nous obtenons à la fois les effectifs et les proportions (Figure 3.1). Nous observons ainsi :

- La première sous-population est définie par les individus en "cdd", la seconde en "cdi".
- Les effectifs respectifs des personnes en "cdd" et "cdi" sont $n_1 = 16$ et $n_2 = 34$
- Le nombre de personnes ayant été accepté parmi les "cdd" (resp. "cdi") est de 6 parmi 16 (resp. 28 parmi 34), la fréquence observée est $f_1 = \frac{6}{16} = 0.3750$ (resp. $f_2 = 0.8235$)
- La fréquence de "Acceptation" globalement, dans les 2 sous-échantillons, est $f = \frac{16\times0.3750+34\times0.8235}{16+34} = 0.6800$
- Nous pouvons en déduire $D = 0.3750 - 0.8235 = -0.4485$, l'estimation de son écart type $\hat{\sigma}_D = \sqrt{0.68(1-0.68)(\frac{1}{16}+\frac{1}{34})} = 0.1414$ et la statistique du test $U = \frac{-0.4485}{0.1414} = -3.1716$
- Le seuil critique au risque 5% pour un test bilatéral est $u_{1-\frac{\alpha}{2}} = 1.96$. Puisque nous sommes dans la zone de rejet $|U| = 3.1716 > u_{1-\frac{\alpha}{2}} = 1.96$, nous concluons que la probabilité d'acceptation du crédit est différent selon le type d'emploi de la personne de référence.
- La p-value du test (0.0015) est cohérente avec à cette conclusion.

Au final, nous dirons qu'il faut quand même être très circonspect par rapport à cet exemple. Le faible effectif, surtout ($n_1 = 16$), rend l'approximation normale un peu (beaucoup) hasardeuse.

3.2 Test d'homogénéité du χ^2 pour 2 populations

3.2.1 Position du problème, statistique du test et région critique

Ce test est d'une certaine manière une généralisation du test de proportion. Nous considérons maintenant que X est une variable discrète prenant R valeurs : X peut être une variable catégorielle, une variable ordinale ou une variable continue découpée en intervalles. Le test est non paramétrique, peu importe la distribution sous-jacente de X.

Nous souhaitons vérifier si la distribution de X est la même dans les deux sous populations. Soit $p_{r/k}$ la probabilité d'obtenir $X = r$ dans la sous-population k, l'hypothèse nulle du test s'écrit :

$$H_0 : p_{r/1} = p_{r/2}, \quad \forall r$$

On sert pour cela des échantillons Ω_1 et Ω_2. Nous noterons o_{rk} le nombre d'observations $X = r$ dans l'échantillon Ω_k. La statistique du test(voir [2], pages 314 et 315) s'écrit

$$\chi^2_{obs} = n_1 n_2 \sum_{r=1}^{R} \frac{\left(\frac{o_{r1}}{n_1} - \frac{o_{r2}}{n_2}\right)^2}{o_{r1} + o_{r2}} \tag{3.2}$$

On remarquera que le rapport $\frac{o_{rk}}{n_k}$ est la proportion du caractère $X = r$ dans le sous échantillon k c.-à-d. $\hat{p}_{r/k} = \frac{o_{rk}}{n_k}$. Sous $H0$, χ^2_{obs} suit asymptotiquement une loi du χ^2 à $(R-1)$ degrés de liberté.

La région critique du test pour un niveau de signification α s'écrit

$$R.C. : \chi^2_{obs} \geq \chi^2_{1-\alpha}(R-1)$$

$\chi^2_{1-\alpha}(R-1)$ est le quantile d'ordre $1-\alpha$ de la loi du $\chi^2(R-1)$.

3.2.2 Un exemple numérique : distribution de l'acceptation selon le type d'emploi

Réitérons notre exemple de test de comparaison de proportions (section 3.1.3) en l'appréhendant sous l'angle du test de comparaison de distributions. Nous souhaitons savoir si la proportion des acceptation et refus de crédit sont les mêmes selon le type d'emploi. Le tableau croisé dynamique initial est le même (Figure 3.1), les ratios calculés sont différents pour obtenir la nouvelle statistique du test (Figure 3.2) :

- A partir des distributions conditionnelles, nous formons $A = \frac{(0.3750-0.8235)^2}{6+28} + \frac{(0.625-0.1765)^2}{10+6} = \frac{0.2012}{34} + \frac{0.2012}{16} = 0.0185$
- Nous en déduisons $\chi^2_{obs} = 16 \times 34 \times 0.0185 = 10.0589$
- Le nombre de degrés de liberté est $R - 1 = 2 - 1 = 1$. La loi du χ^2 nous fournit le seuil critique $\chi^2_{0.95}(1) = 3.8415$
- Au risque $\alpha = 5\%$, nous concluons que les disparités entre les proportions ne sont pas dues au hasard, aux fluctuations d'échantillonnage. Il y a véritablement un différence de comportement dans les 2 sous populations.

	A	B	C	D	E	F	G	H	I
1									
2									
3		Données	Emploi						
4		Nombre de Acceptation2		Nombre de Acceptation					
5	Acceptation	cdd	cdi	cdd	cdi		Ecart	Ecart²	Eff.Marginal
6	oui	6	28	37.50%	82.35%		-0.4485	0.2012	34
7	non	10	6	62.50%	17.65%		0.4485	0.2012	16
8	Total	16	34	100.00%	100.00%				
9									
10									
11		A	0.0185						
12		chi^2_obs	10.0589						
13									
14		ddl	1						
15		chi^2_0.95(1)	3.8415						
16									
17		p-value	0.0015						

Fig. 3.2. Comparaison des distributions - Acceptation selon emploi

L'intérêt de cet exemple est que nous pouvons faire le parallèle avec les résultats de la comparaison des proportions (section 3.1.3). Les résultats doivent être équivalents.

Rappelons la relation entre la loi normale et la loi du χ^2, dans notre cas nous savons que $[\mathcal{N}(0,1)]^2 \equiv \chi^2(1)$. La statistique du test précédent $U = -3.1716$ suit une loi normale, lorsque nous la passons au carré $U^2 = (3.1716)^2 = 10.0589$, exactement la valeur de la statistique obtenue dans la seconde approche $\chi^2_{obs} = 10.0589$.

Pour le cas particulier ($R = 2$), les deux approches sont totalement équivalentes.

3.2.3 Un exemple numérique : distribution de la garantie selon le type d'emploi

	A	B	C	D	E	F	G	H	I
1									
2									
3		Données	Emploi						
4		Nombre de Emploi		Nombre de Emploi2					
5	Garantie.Supp	cdd	cdi	cdd	cdi		Ecart	Ecart²	Eff.Marginal
6	caution	3	2	18.75%	5.88%		0.1287	0.0166	5
7	hypotheque	10	19	62.50%	55.88%		0.0662	0.0044	29
8	non	3	13	18.75%	38.24%		-0.1949	0.0380	16
9	Total	16	34	100.00%	100.00%				
10									
11									
12		A	0.0058						
13		chi^2_obs	3.1745						
14									
15		ddl	2						
16		chi^2_0.95(1)	5.9915						
17									
18		p-value	0.2045						

Fig. 3.3. Comparaison des distributions - Garantie supplémentaire selon emploi

L'avantage du test d'homogénéité des distributions est que nous pouvons appréhender les cas ($R > 2$). Nous voulons cette fois-ci comparer la structure de la garantie supplémentaire apportée par les demandeurs de crédit selon le type d'emploi de la personne de référence : ici, $R = 3$.

Toujours avec l'outil "Tableaux croisés dynamiques" d'Excel, nous élaborons le tableau des effectifs et des distributions conditionnelles (Figure 3.3) :

- Mis à part le fait que nous avons $R = 3$ lignes dans le tableau croisé dynamique, la structure de la feuille de calcul est exactement la même. Ainsi $A = \frac{(0.1875-0.0588)^2}{3+2} + \frac{(0.6250-0.5588)^2}{10+19} + \frac{(0.1875-0.3824)^2}{3+13} = \frac{0.0166}{5} + \frac{0.0044}{19} + \frac{0.0380}{16} = 0.0058$
- Nous en déduisons $\chi^2_{obs} = 16 \times 34 \times 0.0058 = 3.1745$.
- Avec $R - 1 = 3 - 1 = 2$ degrés de liberté, le seuil critique du test au risque $\alpha = 5\%$ est $\chi^2_{0.95}(2) = 5.9915$.
- L'hypothèse d'homogénéité des distributions n'est pas contredite par les données. La structure de la "garantie supplémentaire" est la même quel que soit le type d'emploi de la personne de référence.
- La p-value du test est 0.2045.

3.2.4 Généralisation du test d'homogénéité à $K > 2$ populations

Le test d'homogénéité du χ^2 peut être généralisé à $K > 2$ populations. Introduisons les nouvelles notations suivantes :

- $o_{.k} = \sum_{r=1}^{R} o_{rk}$
- $o_{r.} = \sum_{k=1}^{K} o_{rk}$

La statistique du test s'écrit (voir [2], page 314 ; [11], page 345 et 346) :

$$\chi^2_{obs} = \sum_{r=1}^{R}\sum_{k=1}^{K} \frac{\left(o_{rk} - \frac{o_{r.}o_{.k}}{n}\right)^2}{\frac{o_{r.}o_{.k}}{n}} = n\left[\sum_r \sum_k \frac{o_{rk}^2}{o_{r.}o_{.k}} - 1\right]$$

Sous H_0, la distribution de X est la même quelle que soit la sous population, la statistique du test suit une loi du χ^2 à $(R-1)(K-1)$ degrés de liberté.

Remarque 5 (Test d'indépendance du χ^2 ?). Le test de χ^2 est très répandu. On le retrouve sous différentes formes. La plus populaire est certainement le test d'indépendance entre 2 variables catégorielles, calculée à partir d'un tableau de contingence. **La problématique n'est pas (exactement)** [1] **la même que la notre**, mais on se rend compte que les formules sont identiques.

1. Entre parenthèses le terme "exactement" car, comme nous l'annoncions plus haut, une comparaison de populations peut être vue comme l'analyse d'une association de la variable d'intérêt X avec une variable catégorielle qui permet de définir les sous populations. La distinction n'est peut être pas aussi tranchée finalement.

Partie II

Autres tests

4

Tests pour échantillons appariés

4.1 Principe de l'appariement

L'objectif de l'appariement est de réduire la variabilité due aux observations. Prenons un exemple simple pour expliciter l'idée. Un industriel affirme que son additif pour essence permet de réduire la consommation des automobiles. Pour vérifier cette assertion, nous choisissons au hasard n_1 véhicules, nous leur faisons emprunter un parcours routier, nous notons la consommation de chaque véhicule. Puis nous extrayons un second échantillon de n_2 observations, nous rajoutons l'additif dans le réservoir, sur le même parcours routier, nous mesurons les consommations. Pour tester la réduction la consommation, nous confrontons les deux moyennes observées $\bar{x}_1$ et $\bar{x}_2$. Nous sommes dans un schéma de test sur échantillons indépendants dans ce cas.

En y regardant de plus près, on se rend compte qu'il y a des éléments non maîtrisés dans notre expérimentation. Avec un peu de (mal)chance, il se peut que les petites berlines soient majoritaires dans le premier échantillon, les grosses berlines dans le second. Cela faussera totalement les résultats, laissant à penser que l'additif a un effet néfaste sur les consommations. Le principe de l'appariement est d'écarter ce risque en créant des paires d'observations. Dans notre exemple, nous choisissons en effet n véhicules au hasard [1] dans la population : nous leur faisons faire le trajet normalement une première fois, puis nous rajoutons l'additif dans réservoir, nous leur refaisons parcourir le même chemin. L'écart entre les consommations sera un bon indicateur des prétendues bénéfices introduits par l'additif. Ce schéma "avant-après" est la forme la plus populaire de l'appariement. Elle permet de réduire le risque de second espèce du test c.-à-d. nous augmentons la puissance du test.

L'appariement est en réalité plus large que le seul schéma "avant-après". Il est efficace à partir du moment où nous réunissons les deux conditions suivantes : les individus dans chaque paire se ressemblent le plus possible, ou appartiennent à une même entité statistique (un ménage, des jumeaux, etc.) ; les paires d'observations sont très différentes les unes des autres.

Reprenons notre exemple des additifs pour carburants : nous souhaitons comparer les mérites respectifs de 2 additifs concurrents. On ne peut pas mettre le premier additif, faire faire le trajet, puis ajouter le second additif. Quand bien même nous aurions vidangé le réservoir entre temps, nous ne savons pas si les effets du premier additif sur le moteur se sont estompés. Pour dépasser cet écueil, il serait plus judicieux

1. pas tellement, nous verrons cela plus loin

d'échantillonner des paires de modèles identiques (marque, modèle, kilométrage), et de comparer leurs consommations deux à deux. Nous y gagnerons si les paires sont différentes les unes des autres c.-à-d. couvrant aussi largement que possible le spectre des véhicules existants (petites citadines, familiales, grosses berlines, etc.).

L'appariement, que l'on retrouve sous différentes appellation (mesures répétées, échantillons dépendants, *paired samples* ou *matched pairs samples* en anglais,) est une procédure très populaire en statistique. Elle permet une analyse fine des différences entre les populations. Un excellent document en ligne explique en détail les motivations, les contraintes et les bénéfices associés à cette stratégie - `http://www.tufts.edu/~gdallal/paired.htm`.

4.2 Comparaison des moyennes

4.2.1 Test d'hypothèses, statistique du test et région critique

Nous considérons maintenant que nous disposons d'un échantillon de n observations. Chaque observation étant constituée d'une paire de valeurs. Nous formons une nouvelle variable aléatoire D dont les valeurs d_i sont obtenues par différences des paires de valeurs c.-à-d.

$$d_i = x_{i1} - x_{i2}$$

X étant gaussienne, D l'est également. Nous savons de plus que $E(D) = \mu_D = \mu_1 - \mu_2$. Le test de comparaison de moyennes pour échantillons appariés s'écrit dès lors

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

Il s'agit ni plus ni moins que d'un test de conformité de la moyenne à un standard à partir d'un échantillon de n observations [2].

Notons $\bar{d}$ la moyenne empirique, avec

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

. L'estimation de la variance [3] de D à partir d'un échantillon s'écrit

2. Voir `http://www.itl.nist.gov/div898/handbook/prc/section2/prc22.htm`

3. Les variables ne sont pas indépendantes, la variance ne peut pas se résumer à la somme des variances des variables individuelles. Il faudrait prendre en compte la covariance entre X_1 et X_2, soit

$$V(D) = \sigma_D^2 = \sigma_{X_1 - X_2}^2 = \sigma_1^2 + \sigma_2^2 - 2 \times COV(X_1, X_2)$$

Mais de toute manière, comme nous devons estimer les variances, nous passons directement par l'estimation s_D^2 à partir des observations d_i

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

Sous H_0, la statistique du test s'écrit

$$t = \frac{\bar{d}}{s_D/\sqrt{n}} \tag{4.1}$$

Elle suit une loi de Student à $(n-1)$ degrés de liberté. Nous rejetons l'hypothèse nulle si

$$R.C. : |t| \geq t_{1-\frac{\alpha}{2}}(n-1)$$

Ce schéma implique qu'il ne doit pas y avoir interaction entre les objets et les facteurs. Reprenons l'exemple de comparaison des additifs, si le premier s'avère meilleur chez les petits véhicules, et le second meilleur pour les berlines. La statistique basée sur la moyenne des écarts sera faussée, proche de 0, poussant à l'acceptation de l'hypothèse nulle alors que la situation est en réalité plus complexe.

4.2.2 Un exemple : la comparaison des salaires

Poussés par une curiosité irrépressible, nous souhaitons savoir si dans les ménages, les hommes ont un salaire plus élevé que leur épouse. Bien évidemment, il ne faut surtout pas mettre en oeuvre un test pour échantillons indépendants. La comparaison doit se faire **à l'intérieur** des ménages.

A partir des colonnes de "salaire.homme" et "salaire.femme", nous construisons la variable D. Ensuite, nous vérifions que la moyenne de D s'écarte significativement de la valeur 0 (Figure 4.1) :

- Nous disposons bien de $n = 50$ valeurs de D
- Nous calculons la moyenne empirique $\bar{d} = 0.1546$, son écart type $s_D = 0.2825$
- La statistique du test est obtenue directement à partir des ces valeurs $t = \frac{0.1546}{0.2825/\sqrt{50}} = \frac{0.1546}{0.0400} = 3.8697$
- Au risque $\alpha = 5\%$, le seuil critique est $t_{0.975}(49) = 2.3124$
- Nous nous situons dans la région critique, nous concluons que les salaires dans les ménages ne sont pas identiques, plutôt à l'avantage de l'homme au regard de $\bar{d}$

Remarque 6 (Test sur échantillons indépendants). Si nous traitons le même exemple, en considérant maintenant que les échantillons sont indépendants, nous disposons de 100 observations ($n_1 + n_2 = 50 + 50 = 100$). La procédure de comparaison des moyennes (cf. section 1.2.2) nous fournit $t = \frac{0.1546}{0.1073} = 1.4407$, et nous sommes dans la région d'acceptation de H_0, les salaires moyens seraient identiques... et c'est comme ça qu'on sauve la paix des ménages. Dans la construction du test t, par rapport au test spécifique pour échantillons appariés, nous perdons en précision (écart type plus élevé) ce que nous avions gagné en degrés de liberté.

	A	B	C	D	E	F	G
1	Numéro	Sal.Homme	Sal.Femme	D			
2	1	7.92	7.72	0.20		n	50
3	2	7.97	7.49	0.48			
4	3	6.97	7.10	-0.13		Moyenne (D)	0.1546
5	4	7.85	7.39	0.46		Ecart-type (D)	0.2825
6	5	6.67	6.76	-0.09			
7	6	6.89	6.51	0.38		t	3.8697
8	7	7.29	6.93	0.36			
9	8	7.53	7.51	0.02		t_0.975(49)	2.3124
10	9	7.48	7.25	0.23			
11	10	7.27	6.60	0.67		p-value	0.0003
12	11	7.28	7.47	-0.19			
13	12	8.40	8.07	0.33			
14	13	7.46	6.79	0.67			
15	14	8.42	8.01	0.41			
16	15	7.39	7.44	-0.05			
17	16	7.47	7.59	-0.12			
18	17	7.86	7.50	0.36			
19	18	6.83	7.06	-0.23			
20	19	6.98	7.29	-0.31			
21	20	7.80	7.38	0.42			
22	21	7.67	7.69	-0.02			
23	22	7.28	7.05	0.23			
24	23	7.17	6.86	0.31			
25	24	7.42	7.25	0.17			
26	25	7.83	7.77	0.06			
27	26	7.33	7.14	0.19			
28	27	6.02	6.03	-0.01			
29	28	7.63	7.77	-0.14			
30	29	6.18	6.40	-0.22			
31	30	7.57	7.53	0.04			
32	31	7.36	7.78	-0.42			
33	32	8.03	7.94	0.09			
34	33	8.46	8.12	0.34			
35	34	6.64	7.12	-0.48			
36	35	7.92	7.92	0.00			
37	36	7.14	7.20	-0.06			
38	37	7.13	6.85	0.28			
39	38	7.43	7.20	0.23			
40	39	8.78	8.58	0.20			
41	40	8.28	7.85	0.43			
42	41	6.31	6.57	-0.26			
43	42	7.48	6.97	0.51			
44	43	7.48	6.96	0.52			
45	44	7.69	7.11	0.58			
46	45	7.44	7.16	0.28			
47	46	7.47	7.24	0.23			
48	47	8.17	8.29	-0.12			
49	48	7.40	7.29	0.11			
50	49	7.26	6.81	0.45			
51	50	7.50	7.16	0.34			

Fig. 4.1. Comparaison des salaires - Échantillons appariés

4.3 Comparaison de K moyennes - Les blocs aléatoires complets

4.3.1 Principe

Le test basé sur les plans d'expériences en blocs aléatoires complets est à l'ANOVA ce que le test pour échantillons appariés est pour le test de Student pour échantillons indépendants. L'idée fondatrice est toujours l'appariement, mais nous gérons cette fois-ci K populations (K traitements, voir [3], chapitre 6, pages 141 à 167).

En anglais, le terme consacré est *randomized blocks*[4]. Reprenons notre exemple des additifs pour carburants. Nous souhaitons maintenant comparer $K = 5$ marques différentes. De la même manière

4. Voir `http://www.socialresearchmethods.net/kb/expblock.php` pour une description détaillée de la stratégie et des bénéfices qu'on peut en attendre

que précédemment, nous constituons n unités statistiques (n blocs), chaque unité étant composé de 5 véhicules. Nous attribuons totalement au hasard le traitement à l'intérieur de chaque unité. Plus les individus à l'intérieur d'un bloc se ressemblent, plus nous réduisons la variabilité intra-blocs, en revanche nous avons tout intérêt à élaborer des blocs aussi différents que possible les uns des autres.

L'appariement peut également faire référence aux mesures répétées (*repeated measures* en anglais). Il s'agit en quelque sorte d'une généralisation du canevas "avant-après" présenté dans le cas de 2 traitements. Par exemple, nous souhaitons analyser la résistance à la déchirure de K combinaisons de motards. Nous demandons à des cascadeurs de simuler des chutes. Le plus judicieux serait de demander à chaque cascadeur de répéter K fois la chute avec chaque combinaison, ce faisant nous réduisons autant que possible la variabilité due à l'échantillon. Bien sûr, il ne faut pas qu'il y ait un phénomène d'apprentissage ou d'habitude de la part des sujets. Si les cascadeurs s'enhardissent au point de provoquer des glissades de plus en plus spectaculaires au fil de l'expérimentation, et si nous passons les différents types de combinaisons dans le même ordre pour chaque individu, les résultats seront complètement faussés.

Les techniques présentées dans cette section s'appliquent exactement de la même manière que l'on soit dans un schéma de "mesures répétées" ou de "blocs aléatoires complets".

4.3.2 Statistique du test - Décomposition de la variance

Puisqu'il s'agit d'un cas particulier de comparaison de plusieurs moyennes, nous devons décomposer la variance de manière à mettre en évidence celle due aux K traitements.

Sans trop rentrer dans les détails, nous sommes dans le cadre de l'ANOVA à deux facteurs. Les K traitements (K populations) représentent le premier facteur (souvent fixe) ; les n blocs représentent le second facteur (forcément aléatoire) (voir [1], chapitre VI, pages 179 à 197).

Hypothèses de calcul

Nous appliquons sur n blocs K traitements, x_{ik} est la valeur observée du traitement k sur le bloc $n^o i$. Pour un modèle à facteur fixe, nous supposons vérifiées les hypothèses suivantes :

- $x_{ik} \equiv \mathcal{N}(\mu_{ik}, \sigma)$;
- Les effets blocs et traitements sont additifs c.-à-d. $\mu_{ik} - \mu_{ik'} = \mu_{i'k} - \mu_{i'k'}$. Quel que soit le bloc, les facteurs agissent de la même manière, avec le même écart.
- Cette égalité peut être ré-écrite avec $\mu_{ik} - \mu_{i'k} = \mu_{ik'} - \mu_{i'k'}$. Quel que soit le traitement, les écarts entre les moyennes d'un bloc à l'autre sont les mêmes.

Notons $\mu_{.k}$ l'espérance de l'effet du traitement $n^o k$. Les hypothèses à confronter de l'ANOVA sont "aucun traitement n'a un effet meilleur que les autres" (ou "tous les traitements produisent le même résultat") vs. un des traitements au moins se démarque d'un autre :

$$H_0 : \mu_{.1} = \cdots = \mu_{.K}$$
$$H_1 : \text{deux des moyennes au moins sont différentes}$$

Tableau d'analyse de variance

A l'instar de l'ANOVA à un facteur (section 1.3), il nous faut décomposer la variance totale pour obtenir une réponse au test. Introduisons tout d'abord de nouvelles notations :

- $T_{i.} = \sum_k x_{ik}$ est la somme des valeurs à l'intérieur du bloc i ;
- $\bar{x}_{i.} = \frac{T_{i.}}{K}$ est la moyenne des valeurs à l'intérieur d'un bloc ;
- $T_{.k} = \sum_i x_{ik}$ est la somme des valeurs associées au traitement k ;
- $\bar{x}_{.k} = \frac{T_{.k}}{n}$ est la moyenne associée au traitement k ;
- $\bar{x}_{..} = \frac{1}{n \times K} \sum_i \sum_k x_{ik}$ est la moyenne globale.

L'équation d'analyse de variance s'écrit (voir [6], pages 67 à 71 pour les démonstrations) :

$$SCT = SCE + SCB + SCR' \tag{4.2}$$

$$\sum_i \sum_k (x_{ik} - \bar{x}_{..})^2 = \sum_i \sum_k (\bar{x}_{.k} - \bar{x}_{..})^2 + \sum_i \sum_k (\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_i \sum_k (x_{ik} - \bar{x}_{i.} - \bar{x}_{.k} + \bar{x}_{..})^2 \tag{4.3}$$

où SCT traduit la variabilité totale ; SCE la variabilité expliquée par les traitements ; SCB la variabilité expliquées par les blocs ; SCR' la variabilité résiduelle, non expliquée par la modélisation.

En faisant le parallèle avec la décomposition de la variance de l'ANOVA à 1 facteur pour échantillons indépendants (section 1.3.2), nous nous rendons compte que la partie non expliquée est maintenant décomposée en 2 parties

$$SCR = SCB + SCR'$$

Entre d'autres termes, nous avons réussi à isoler de l'erreur les différences entre les blocs [5]. Sachant que $SCB \geq 0$, en passant par les blocs aléatoires complets nous ne pouvons que réduire la partie non modélisée ($SCR' \leq SCR$) par rapport à l'ANOVA à échantillons indépendants (que l'on appelle aussi *expériences complètement aléatoires* dans la terminologie des plans d'expériences - voir [3], chapitre 5, pages 109 à 139).

Nous pouvons résumer les différentes sources de variabilité dans le tableau d'analyse de variance (Tableau 4.1).

Source	Somme.Carrés	ddl	Carrés.Moyens	F
Expliquée (Traitement)	$SCE = \sum_i \sum_k (\bar{x}_{.k} - \bar{x}_{..})^2$	$K-1$	$CME = \frac{SCE}{K-1}$	$F = \frac{CME}{CMR'}$
Blocs	$SCB = \sum_i \sum_k (\bar{x}_{i.} - \bar{x}_{..})^2$	$n-1$	$CMB = \frac{SCR}{n-1}$	-
Résiduelle	$SCR' = \sum_i \sum_k (x_{ik} - \bar{x}_{i.} - \bar{x}_{.k} + \bar{x}_{..})^2$	$(n-1)(K-1)$	$CMR' = \frac{SCR'}{(n-1)(K-1)}$	-
Total	$SCT = \sum_i \sum_k (x_{ik} - \bar{x}_{..})^2$	$nK-1$	-	-

Tableau 4.1. Tableau de l'analyse de la variance - Blocs aléatoires complets

La statistique du test est donc $F = \frac{CME}{CMR'}$.

5. voir `http://davidmlane.com/hyperstat/within-subjects.html`

Sous H_0, elle suit une loi de Fisher à $[K-1,(n-1)(K-1)]$ degrés de libertés. Nous rejetons l'hypothèse nulle si F est plus grand que le quantile d'ordre $1-\alpha$ de la loi de Fisher à $[K-1,(n-1)(K-1)]$ degrés de liberté.

Remarque 7 (Lorsque $K = 2$). Lorsque nous avons $K = 2$, la solution mise en avant dans cette section concorde exactement avec celle de la comparaison de moyennes pour échantillons appariés (section 4.2)(Voir [6], page 74-75 pour la démonstration).

4.3.3 Un exemple : analyse de l'endurance de pneumatiques

Notre fichier des demandeurs de crédit ne se prête pas à ce type d'analyse. Nous utilisons des données spécifiques pour illustrer cette section. Nous reproduisons un exercice proposé dans l'ouvrage de Guenther ([6], exercice 3.5, page 74).

Nous souhaitons comparer la distance que l'on peut parcourir à l'aide de 4 marques de pneumatiques (A, F, G, R), pour un type de véhicule spécifique. Nous les montons neufs sur le véhicule. Nous les faisons rouler jusqu'à ce que le témoin d'usure soit apparent. Nous mesurons alors la distance totale parcourue (en millier de miles).

	A	B	C	D	E	F	G
1			Marque de pneumatiques (Traitement)				
2			A	F	G	R	**Moyenne**
3	Blocs	1	38	29	41.5	39	36.88
4		2	24.5	36	35.5	25	30.25
5		3	37.5	38.5	31.5	29.5	34.25
6		4	20.5	33.5	29.5	21.5	26.25
7		5	29.5	35	34	11	27.38
8		6	22	33	35	17	26.75
9		7	29	37.5	38	18.5	30.75
10		8	25	21.5	28	18	23.13
11		9	26	35.5	28	17	26.63
12		10	17	23.5	34	16.5	22.75
13		**Moyenne**	26.90	32.30	33.50	21.30	**28.50**
14							
15			**Tableau d'analyse de variance**				
16		Source	Somme.Carrés	ddl	Carrés.Moyens	F	p-value
17		Traitement	938.40	3	312.80	11.5796	0.00005
18		Bloc	744.75	9	82.75		
19		Résiduel	729.35	27	27.01		
20		Total	2412.50	39			

Fig. 4.2. Endurance de différentes marques de pneumatiques - Blocs aléatoires complets

On sait que l'endurance des pneumatiques est largement influencée par le type de parcours et les conditions climatiques. L'idée est de réduire la variabilité en créant 10 catégories (blocs) aussi homogènes que possibles au regard de ces éléments. A l'intérieur de chaque bloc, composé de 4 véhicules aux caractéristiques identiques, les pneumatiques sont attribués aléatoirement.

Les données sont saisies dans une feuille EXCEL, nous mettons en place les calculs (Figure 4.2) :

– Dans le tableau de calcul, nous avons en ligne les $n = 10$ blocs. La marge correspond à la moyenne des blocs. Par exemple, $\bar{x}_{1.} = 36.88$

- En colonne, nous avons les $K = 4$ types de pneumatiques. En marge, les moyennes des traitements. Par exemple, $\bar{x}_{.1} = 26.90$
- La moyenne de l'ensemble des valeurs est $\bar{x}_{..} = 28.50$
- Nous pouvons construire le tableau d'analyse de variance, conformément à la structure décrite dans le Tableau 4.1. Nous obtenons successivement $SCE = 938.40$, $SCB = 744.75$, $SCT = 2412.5$, et par différence $SCR' = 729.35$
- Après avoir calculé les degrés de liberté et les carrés moyens, nous pouvons former le statistique du test $F = \frac{312.8}{27.01} = 11.5796$
- A 5%, le seuil critique du test est $F_{0.95}(3, 27) = 2.9603$. Nous sommes dans la région critique, nous concluons que les pneumatiques ont des durées de vie différentes.
- La probabilité critique (p-value = 0.00005) aboutit bien évidemment à la même conclusion.

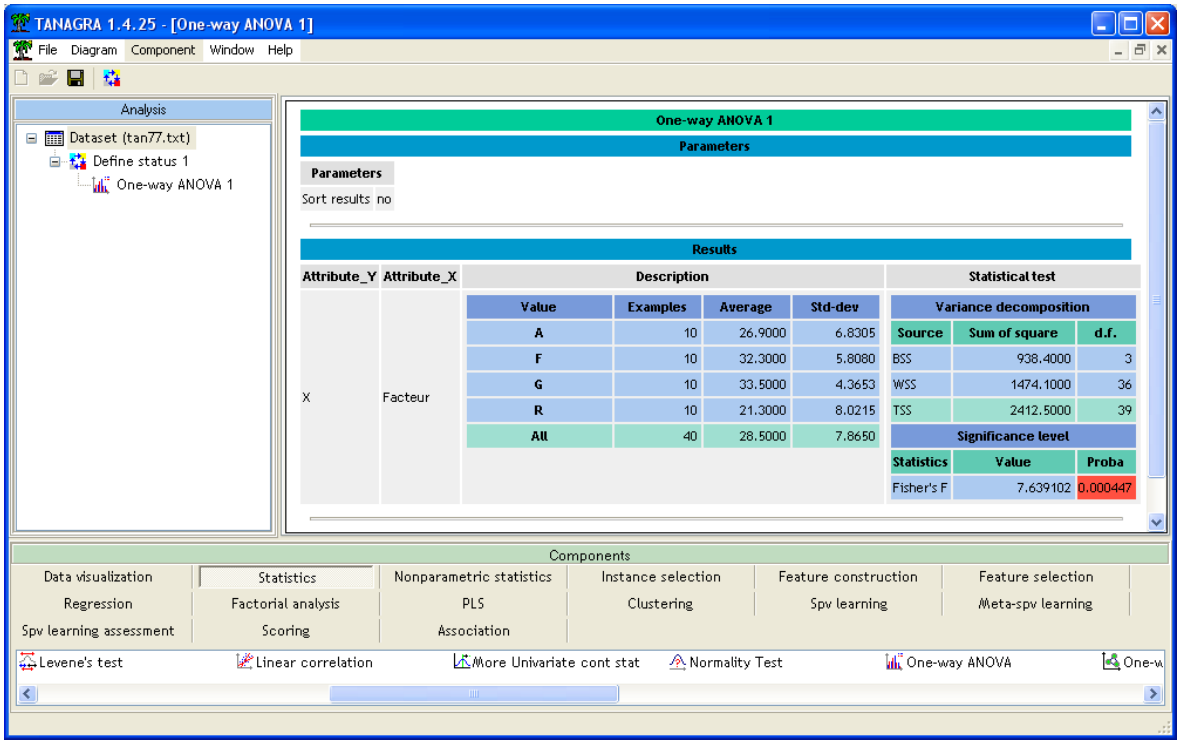


Fig. 4.3. Endurance de différentes marques de pneumatiques - ANOVA échantillons indépendants

Remarque 8 (Et si on ne tenait pas compte des blocs ?). Nous avons procédé au même test en considérant maintenant que les échantillons sont indépendants. Nous omettons des informations très importantes ce faisant. Voyons ce qu'il en est des résultats (Figure 4.3, nous avons utilisé le logiciel Tanagra) :

- Nous retrouvons les valeurs des moyennes des traitements et la moyenne globale.
- La variabilité totale est la même $TSS = SCT = 2412.5$, celle expliquée par les traitements également $BSS = SCE = 938.4$. La différence maintenant est que la variabilité due aux blocs (SCB) et résiduelle (SCR') sont confondues dans la variabilité intra-classes $WSS = SCR = 1474.1 = 744.75 + 729.35$.

- De fait, la quantité au numérateur de la statistique n'est pas modifiée, le dénominateur en revanche est augmentée, F diminue mécaniquement, elle passe à $F = 7.6391$.
- Cette baisse n'est pas compensée par une augmentation des degrés de liberté. Nous constatons que l'ANOVA pour échantillons indépendants signale comme moins significative les écarts entre les performances des pneumatiques (p-value = 0.00045). Il reste toutefois qu'au niveau de signification 5%, les conclusions des 2 approches concordent.

4.4 Comparaison des variances pour 2 échantillons appariés

4.4.1 Test, statistique du test, région critique

De la même manière que pour les échantillons indépendants, nous pouvons être emmenés à procéder à une test de comparaison de variances pour 2 échantillons appariés. Dans notre exemple d'additifs pour carburants, on veut tester si la moyenne les consommations baissent, mais on peut aussi vouloir tester si les valeurs constatées se resserrent, le produit homogénéise-t-il le comportement des véhicules ?

Nous disposons d'un échantillon de taille n, nous mesurons les variables X_1 et X_2 sur les mêmes individus (ou tout du moins sur un échantillon apparié). Nous souhaitons tester l'égalité des variances c.-à-d.

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Le test le plus répandu (ou le plus souvent cité) est celui de Pitman (1939) [6]. Il repose sur un résultat fondamental. Soient les variables synthétiques U et V avec $U = X_1 + X_2$ et $V = X_1 - X_2$, on montre la relation suivante

$$r_{uv} = 0 \Leftrightarrow \frac{\sigma_1}{\sigma_2} = 1$$

r_{uv} est le coefficient de corrélation de Pearson. De fait, le test d'égalité des variances est totalement équivalent au test d'hypothèses suivant

$$H_0 : r_{uv} = 0$$
$$H_1 : r_{uv} \neq 0$$

Il s'agit d'un test de significativité de la corrélation, bien connue [7]. On notera néanmoins que ce test est relativement peu robuste par rapport à l'hypothèse sous-jacente de normalité bivariée du couple (U, V). Ce qui limite quelque peu sa portée.

6. E.G. Pitman, *A note on normal correlation*, Biometrika, 31, 9-12, 1939.
7. R. Rakotomalala, *Analyse de Corrélation - Etude des dépendances, variables quantitatives*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf`

La statistique du test est donc celui du test de significativité de la corrélation, soit

$$t = \frac{\hat{r}_{uv}}{\sqrt{\frac{1-\hat{r}_{uv}^2}{n-2}}} \tag{4.4}$$

$\hat{r}_{uv}$ est l'estimation de r_{uv} sur l'échantillon. Sous H_0, t suit une loi de Student à $(n-2)$ degrés de liberté. La région critique du test au risque α s'écrit

$$R.C. : |t| \geq t_{1-\frac{\alpha}{2}}(n-2)$$

4.4.2 Un écriture équivalente de la statistique du test

On rencontre parfois dans la littérature une autre formulation de la même statistique de test. Elle est exprimée directement à partir des variables X_1 et X_2, elle ne fait donc pas intervenir les variables transformées U et V.

Soient les estimateurs de la variance s_1^2 et s_2^2 de X_1 et X_2, nous en déduisons $s_{max}^2 = \max(s_1^2, s_2^2)$ et $s_{min}^2 = \min(s_1^2, s_2^2)$. Nous définissons F comme le rapport

$$F = \frac{s_{max}^2}{s_{min}^2}$$

.

La statistique du test de comparaison de variables s'écrit

$$t = \frac{(F-1)\sqrt{n-2}}{2\sqrt{F(1-\hat{r}_{x_1x_2}^2)}} \tag{4.5}$$

où $\hat{r}_{x_1x_2}^2$ est le coefficient de corrélation empirique entre les variables X_1 et X_2.

Les équations 4.4 et 4.5 sont totalement équivalentes. La distribution sous H_0 et la région critique sont bien évidemment identiques.

4.4.3 Exemple : dispersion des salaires hommes-femmes

Nous souhaitons comparer la variance des salaires des hommes (X_1) et des femmes (X_2) de notre fichier de travail (Figure 0.1). Nous ne pouvons pas mettre en oeuvre un test pour échantillons indépendants (test de Bartlett, Levene, etc.), en effet les individus vont en couple. Nous allons utiliser les 2 approches ci-dessus (équations 4.4 et 4.5) et comparer les résultats (Figure 4.4).

Pour la première approche (**Solution 1**) :

- Nous créons les deux colonnes U ($u_i = x_{i1} + x_{i2}$) et V ($v_i = x_{i1} - x_{i2}$)
- Nous obtenons la corrélation $\hat{r}_{uv} = 0.1903$ et la statistique du test $t = \frac{0.1903\sqrt{50-2}}{\sqrt{1-0.1903^2}} = 1.3427$.
- t suit une loi de Student à $(n-2 = 50-2 = 48)$ degrés de liberté. La probabilité critique du test est p-value $= 0.1857$.

	A	B	C	D
1	Sal.Homme (X1)	Sal.Femme (X2)	U=X1+X2	V=X1-X2
2	7.92	7.72	15.64	0.20
3	7.97	7.49	15.46	0.48
4	6.97	7.10	14.07	-0.13
5	7.85	7.39	15.24	0.46
6	6.67	6.76	13.43	-0.09
7	6.89	6.51	13.40	0.38
8	7.29	6.93	14.22	0.36
9	7.53	7.51	15.04	0.02
10	7.48	7.25	14.73	0.23
11	7.27	6.60	13.87	0.67
12	7.28	7.47	14.75	-0.19
13	8.40	8.07	16.47	0.33
14	7.46	6.79	14.25	0.67
15	8.42	8.01	16.43	0.41
16	7.39	7.44	14.83	-0.05
17	7.47	7.59	15.06	-0.12
18	7.86	7.50	15.36	0.36
19	6.83	7.06	13.89	-0.23
20	6.98	7.29	14.27	-0.31
21	7.80	7.38	15.18	0.42
22	7.67	7.69	15.36	-0.02
23	7.28	7.05	14.33	0.23
24	7.17	6.86	14.03	0.31
25	7.42	7.25	14.67	0.17
26	7.83	7.77	15.60	0.06
27	7.33	7.14	14.47	0.19
28	6.02	6.03	12.05	-0.01
29	7.63	7.77	15.40	-0.14
30	6.18	6.40	12.58	-0.22
31	7.57	7.53	15.10	0.04
32	7.36	7.78	15.14	-0.42
33	8.03	7.94	15.97	0.09
34	8.46	8.12	16.58	0.34
35	6.64	7.12	13.76	-0.48
36	7.92	7.92	15.84	0.00
37	7.14	7.20	14.34	-0.06
38	7.13	6.85	13.98	0.28
39	7.43	7.20	14.63	0.23
40	8.78	8.58	17.36	0.20
41	8.28	7.85	16.13	0.43
42	6.31	6.57	12.88	-0.26
43	7.48	6.97	14.45	0.51
44	7.48	6.96	14.44	0.52
45	7.69	7.11	14.80	0.58
46	7.44	7.16	14.60	0.28
47	7.47	7.24	14.71	0.23
48	8.17	8.29	16.46	-0.12
49	7.40	7.29	14.69	0.11
50	7.26	6.81	14.07	0.45
51	7.50	7.16	14.66	0.34

Solution 1

r(U,V)	0.1903
n	50
t	1.3427
p-value	0.1857

Solution 2

s^2(X1)	0.3157
s^2(X2)	0.2600
F	1.2140
r(X1,X2)	0.8654
t	1.3427
p-value	0.1857

Fig. 4.4. Comparaison de 2 variances - Échantillons appariés

- Au seuil de signification 5%, nous dirons que l'hypothèse d'égalité des variances ne peut être rejetée.

Pour la seconde approche (**Solution 2**) :

- Nous calculons les variances empiriques $s_1^2 = 0.3157$ et $s_2^2 = 0.2600$. Nous en déduisons $s_{max}^2 = 0.3157$ et $s_{min}^2 = 0.2600$
- Dès lors $F = \frac{0.3157}{0.2600} = 1.2140$
- Par ailleurs, nous calculons la corrélation entre les variables originelles $\hat{r}_{x_1x_2} = 0.8654$ (on observe au passage que les salaires dans les ménages sont fortement corrélés)
- Nous appliquons la formule 4.5 pour obtenir la statistique du test $t = \frac{(1.2140-1)\sqrt{50-2}}{2\sqrt{1.2140(1-0.8654^2)}} = 1.3427$
- t suit toujours une loi de Student à $(50-2=48)$ degrés de liberté, la p-value du test est exactement la même (0.1857).

Partie III

Tests multivariés

Cette partie de notre support se démarque des précédentes. **Nous analysons simultanément plusieurs variables d'intérêt**. La variable aléatoire X est à p dimensions.

Nous nous situons toujours dans un cadre paramétrique, nous faisons l'hypothèse que X **suit une loi normale multidimensionnelle ou loi multinormale, de paramètre μ, son barycentre, et Σ, sa matrice de variance covariance**. Nous utiliserons la notation

$$X \equiv \mathcal{N}_p(\mu, \Sigma)$$

L'objectif de la comparaison de populations n'est pas fondamentalement modifié. Il s'agit toujours de s'assurer que les paramètres de la distribution des données est la même dans $(K \geq 2)$ sous populations. La démarche est identique, seules les statistiques de test et leurs distributions seront modifiées.

Trouver de la documentation détaillée sur le sujet que nous traitons dans cette partie est assez ardue. Par le plus grand des hasards, j'ai pu dénicher un extraordinaire travail accessible gratuitement en ligne (`http://www.stat.psu.edu/online/development/stat505/`). Quasiment tout y est concernant l'inférence statistique multidimensionnelle. Soyons honnête, mon principal mérite dans cette histoire est d'inscrire le sujet dans le canevas de ce support, en préservant autant que possible la cohérence de la présentation ; de rédiger le tout en français ; et de détailler les calculs dans un tableur, afin de décrire finement les principales étapes du calcul, la source ci-dessus se contentant de commentaires de sorties de logiciels commerciaux. A ce titre, nous utiliserons les mêmes exemples que le site pour que tout un chacun puisse comparer les résultats en détail. Les données sont accessibles sur l'URL suivant : `http://www.stat.psu.edu/online/development/stat505/data.htm`

Pour ceux qui souhaitent approfondir le sujet, le mieux est de consulter ce site (que je trouve réellement extraordinaire). Vous y trouverez, entre autres, les parties que j'ai éludé car elles ne rentrent pas directement dans la trame de ce support : la construction pratique des ellipsoïdes de confiance, l'analyse discriminante, etc.

5

Notations et bases inférentielles

5.1 Notations

La nature du problème étant quelque peu modifiée, nous allons redéfinir nos notations. Attention, dans certains cas, elles ne seront pas cohérentes avec les parties précédentes. Nous avons préféré prendre ce risque pour être en accord avec notre principale source, plutôt que d'inventer des nouveaux sigles venus de nulle part, trop complexes à force de vouloir être trop précis.

X correspond maintenant à un ensemble de p variables c.-à-d. $X = (X_1|\dots|X_p)$. Elle suit une loi multinormale $\mathcal{N}_p(\mu, \Sigma)$. μ correspond au vecteur des espérances mathématiques, elle est de dimension $(p, 1)$

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

avec $\mu_j = E(X_j), \quad j = 1, \dots, p$.

Σ est la matrice de variance covariance, de dimension (p, p). Elle est symétrique, la quantité de coordonnée (l, c) est

$$\sigma_{lc} = COV(X_l, X_c)$$

Sur la diagonale principale, nous avons la variance

$$\sigma_{ll} = \sigma_l^2 = V(X_l)$$

A partir d'un échantillon Ω, nous avons une matrice de données avec n observations et p variables. Nous construisons les estimateurs usuels. La moyenne empirique s'écrit

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

avec $\bar{x}_j = \frac{1}{n}\sum_i x_{ij}$; x_{ij} étant la valeur observée de la variable X_j pour l'individu $n^o i$

Concernant la matrice de variance covariance estimée, nous avons S de dimensions (p, p), avec pour contenu de la cellule (l, c)

$$s_{lc} = \frac{1}{n-1}\sum_i (x_{il} - \bar{x}_l)(x_{ic} - \bar{x}_c) \tag{5.1}$$

La matrice est symétrique, sur la diagonale principale nous avons l'estimation non biaisée de la variance

$$s_{ll} = s_l^2 = \frac{1}{n-1}\sum_i (x_{il} - \bar{x}_l)^2$$

Enfin, nous rajoutons l'indice k lorsque nous travaillons sur le sous échantillon Ω_k, nous obtenons ainsi :

- μ_k pour le barycentre théorique ;
- $\bar{X}_k$ pour le barycentre empirique, estimé sur un échantillon de taille n_k ;
- Σ_k pour la matrice de variance covariance théorique ;
- S_k pour la matrice de variance covariance empirique.

5.2 Loi normale multidimensionnelle et autres lois importantes

5.2.1 Loi multinormale

Fonction de densité

La fonction de densité de la loi multinormale s'écrit

$$f(X) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} exp[-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)]$$

où (det Σ est de déterminant de la matrice de variance covariance, Σ^{-1} son inverse.

La fonction de densité prend son maximum autour de la moyenne théorique, la forme du nuage de points dépend de l'intensité du lien entre les variables. Visualisons cela dans le cas à 2 variables, on se rend compte que le nuage devient de plus en plus effilé à mesure que la corrélation r entre les variables augmente (Figure 5.1, voir `http://www.stat.psu.edu/online/development/stat505/05_multnorm/03_multnorm_example.html`).

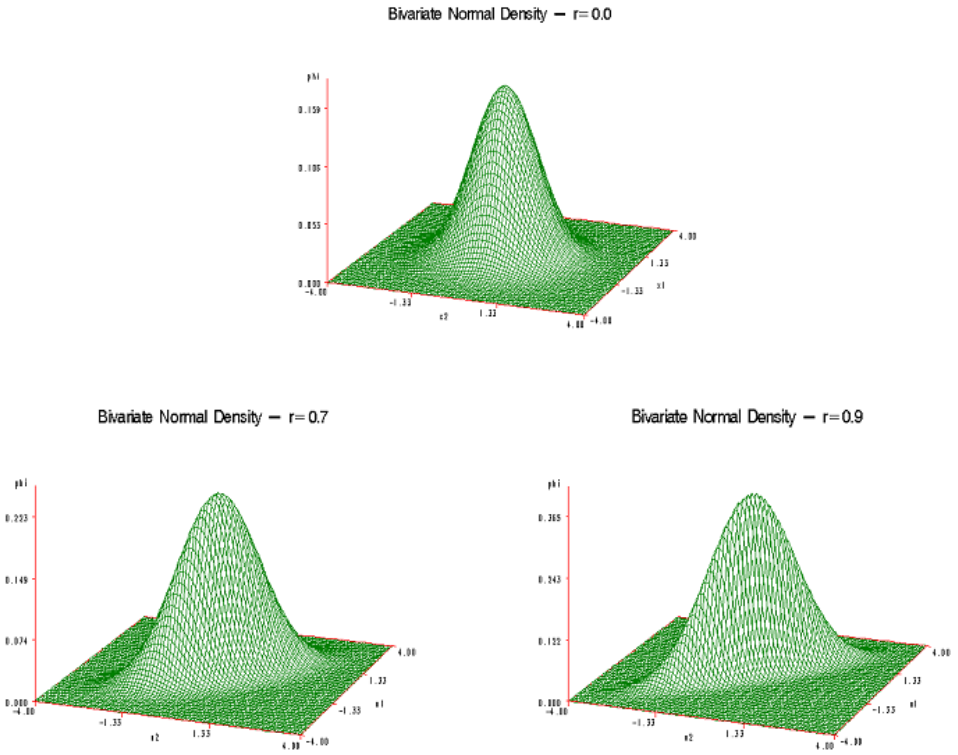


Fig. 5.1. Fonction de densité de la loi multinormale ($p = 2$) en fonction de la corrélation r

Distance de Mahalanobis et variance généralisée

La quantité Δ_p^2 est le carré de la distance de Mahalanobis :

$$\Delta_p^2 = (X - \mu)' \Sigma^{-1} (X - \mu)$$

Elle définit la distance de tout point par rapport au centre de gravité, *en tenant compte de la forme du nuage de points.* Si X est gaussien, on sait que Δ_p^2 suit une loi du χ^2 à p degrés de liberté

$$\Delta_p^2 \equiv \chi^2(p)$$

Ainsi, pour un niveau de confiance $(1-\alpha)$, nous pouvons obtenir les contours de l'ellipsoïde théorique qui a une probabilité $(1 - \alpha)$ de contenir les observations dans notre espace de représentation. Elle est définie par

$$(X - \mu)' \Sigma^{-1} (X - \mu) = \chi^2_{1-\alpha}(p)$$

$\chi^2_{1-\alpha}(p)$ est le quantile d'ordre $1 - \alpha$ de la loi du χ^2

Dans le cas à $p = 2$ dimensions, nous pouvons visualiser cette ellipse (Figure 5.2, voir `http://www.stat.psu.edu/online/development/stat505/05_multnorm/04_multnorm_geom.html`)

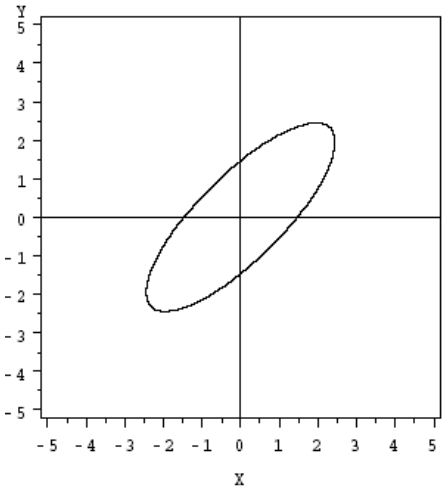


Fig. 5.2. Ellispoïde de confiance théorique dans un espace à $p = 2$ dimensions

Le déterminant de la matrice de variance covariance $\det \Sigma$ est la **variance généralisée**. C'est la généralisation à p dimensions de la variance usuelle. Sa racine carrée est proportionnelle au volume de l'ellipsoïde théorique de confiance (les quantités p et α interviennent également dans le calcul ; voir [2], page 122).

5.2.2 Loi de Wishart, loi de Hotelling et loi de Wilks

Nous présentons succinctement dans cette section les lois de probabilités utilisées pour le calcul statistique dans le cadre multidimensionnel (pour plus de détails, voir [11], pages 103 à 106). Nous les mettrons constamment en parallèle avec l'équivalent univarié.

Loi de Wishart

Une matrice M suit une loi de Wishart $W_p(n, \Sigma)$ si elle peut s'écrire sous la forme $M = X'X$, où X est une matrice de n observations sur p variables. Les observations sont *i.i.d*, distribuées selon une loi multinormale centrée $\mathcal{N}_p(0, \Sigma)$.

La loi de Wishart est la généralisation à p dimensions de la loi du χ^2.

Loi du T^2 de Hotelling

Si x suit une loi normale $\mathcal{N}_p(0, I)$, et M est une matrice de Wishart $W_p(n, I)$, alors la quantité

$$T^2 = nx'M^{-1}x$$

Suit une loi de Hotelling, notée $T_p^2(n)$.

On ne manquera par de faire le parallèle avec le carré de la loi de Student qui, rappelons le, est formée à partir du rapport entre une loi normale et la racine carrée d'un χ^2 normalisé par les degrés de liberté.

Loi du Λ de Wilks

Soient A (resp. B), une matrice de Wishart $W_p(m, \Sigma)$ (resp. $W_p(n, \Sigma)$, alors le rapport

$$\Lambda = \frac{|A|}{|A+B|}$$

Suit une loi de Wilks notée $\Lambda(p, m, n)$.

On peut la voir comme une généralisation multidimensionnelle de la loi de Fisher qui est, rappelons le, formée à partir du rapport de 2 loi de χ^2 normalisées par leurs degrés de liberté respectifs.

5.3 Test de comparaison de la moyenne à un standard)

5.3.1 Définition du test

Test multivarié. Commençons préalablement par le test de comparaison à un standard. Il ne s'agit pas vraiment d'un test de comparaison de populations. Nous le présentons néanmoins car nous y trouvons tout le ferment des tests de comparaison de moyennes dans un espace multidimensionnel.

Pour un ensemble de p variables d'intérêt, nous souhaitons mettre en oeuvre le test d'hypothèses suivant :

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

N'oublions pas que nous sommes dans un espace à p dimensions. μ_0 est le vecteur de référence $(\mu_{10}, \ldots, \mu_{p0})'$.Le test équivaut en réalité à :

$$H_0 : \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{1,0} \\ \vdots \\ \mu_{p,0} \end{pmatrix}$$
$$H_1 : \exists j, \quad \mu_j \neq \mu_{j,0}$$

L'impossibilité de décomposer en tests individuels. Il suffit d'une différence sur une des dimensions pour invalider l'hypothèse nulle. La tentation est grande de décomposer le test en une série de p tests individuels

$$H_0 : \mu_j = \mu_{j,p}$$
$$H_1 : \mu_j \neq \mu_{j,p}$$

Cette démarche n'est pas appropriée pour 2 raisons [1] :

1. voir `http://www.stat.psu.edu/online/development/stat505/08_Hotel/03_Hotel_naive.html`

1. Le test est réalisé en une seule fois lorsqu'il est multivarié. Nous avons une répétition des tests dans la démarche univariée, accroissant le risque de rejeter à tort l'hypothèse nulle. Il serait possible de corriger cela en s'inspirant des techniques populaires dans les comparaisons multiples, telles que la correction de Bonferroni. Mais nous nous heurtons alors à un second écueil.
2. Les variables ne sont pas indépendantes, elle sont plus ou moins corrélées. En les testant individuellement, nous omettons totalement les éventuelles interactions. Il semble alors que la stratégie univariée soit trop conservatrice (favorisant l'hypothèse nulle).

Il nous faut donc produire une statistique de test spécifiquement multivariée.

5.3.2 Statistique, distribution et région critique - Σ est connue

Rappelons que les n observations sont *i.i.d.* c.-à-d. indépendantes et suivent la même loi normale multidimensionnelle. Nous estimons le barycentre théorique avec le barycentre empirique. Le calcul est simple, il suffit de calculer la moyenne sur chaque variable X_j. Nous obtenons le vecteur $\bar{X}$ de dimension $(p, 1)$.

La statistique du test, si Σ est connu s'écrit :

$$\chi^2 = n(\bar{X} - \mu_0)'\Sigma^{-1}(\bar{X} - \mu_0)$$

Sous H_0, elle suit une loi du χ^2 à (p) degrés de liberté.

L'intérêt pratique de ce test est pour le moins restreint. Nous signalerons avant tout la similitude de la formulation avec le test de la comparaison de la moyenne unidimensionnelle avec un standard, lorsque l'écart type est connu.

5.3.3 Statistique, distribution et région critique - Σ est inconnue

Dans la pratique, nous devons estimer la matrice de variance covariance à partir des données, nous utilisons S (section 5.1). Dès lors, la statistique du test devient

$$T^2 = n(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0) \tag{5.2}$$

T^2 suit une loi de Hotelling $T_p^2(n)$.

L'utilisation directe de cette statistique impose que nous ayons sous la main les tables de la loi de Wishart. Ce n'est pas toujours évident. On utilise usuellement une transformation qui permet de nous ramener vers les lois usuelles.

La statistique F définie de la manière suivante

$$F = \frac{n-p}{p(n-1)}T^2 \tag{5.3}$$

Suit une loi de Fisher à $(p, n-p)$ degrés de liberté. La région critique du test pour un risque α s'écrit

$$R.C. : F \geq F_{1-\alpha}(p, n-p)$$

5.3.4 Un exemple : la nutrition des femmes

Nous utilisons les données *Women's Nutrition Data*. Il s'agit d'une enquête portant sur l'alimentation de ($n = 737$) femmes. L'objectif est de comparer les apports en éléments divers (calcium, fer, etc.) avec les quantités recommandées par l'administration (μ_0). Le détail des résultats sont décrits en ligne `http://www.stat.psu.edu/online/development/stat505/08_Hotel/05_Hotel_example.html`. Nous allons essayer de les retrouver en réalisant les calculs avec un tableur. C'est à mon sens un exercice très pédagogique qui permet de démystifier les procédures statistiques.

Variable	Référence	Moyenne
Calcium (mg)	1000	624.0
Fer (mg)	15	11.1
Protéines (g)	60	65.8
Vitamine A (μg)	800	839.6
Vitamine C (mg)	75	78.9

Tableau 5.1. Alimentation des femmes (Données USDA) - Quantités recommandées et moyennes mesurées

Les valeurs de référence (recommandées par l'administration) et les moyennes mesurées sur 737 individus sont recensées dans le tableau 5.1. Reste à savoir si les différences sont significatives. Nous décrivons les calculs dans une feuille EXCEL (Figure 5.3) :

- La première étape importante est le calcul de la matrice de variance covariance. Nous pouvons calculer individuellement les cellules de la matrice (équation 5.1). Il n'y a que $\frac{p(p+1)}{2}$ calculs à faire car elle est symétrique. Mais on se rend compte rapidement que ce n'est pas tenable dès que le nombre de variables augmente, il faut adopter une autre stratégie.
- X est la matrice des données, de dimension (n, p). Nous formons la matrice $\overset{o}{X}$ où chaque colonne est centrée c.-à-d. dans chaque colonne, nous retranchons aux valeurs la moyenne (de la colonne). La matrice de variance covariance estimée est directement obtenue avec

$$S = \frac{1}{n-1} \overset{o}{X}{}' \overset{o}{X}$$

  C'est la stratégie que nous avons adopté. En **M6..Q10**, nous avons tout d'abord $\overset{o}{X}{}' \overset{o}{X}$. Puis en **M13..Q17**, la matrice S.
- Nous l'inversons pour obtenir S^{-1} en **M20..Q24**.
- Parallèlement à cela, nous calculons la différence entre la moyenne empirique et la référence

$$\bar{X} - \mu_0 = \begin{pmatrix} 624.0 - 1000 \\ 11.1 - 15 \\ 65.8 - 60 \\ 839.6 - 800 \\ 78.9 - 75 \end{pmatrix} = \begin{pmatrix} -376.0 \\ -3.9 \\ 5.8 \\ 39.6 \\ 3.9 \end{pmatrix}$$

- Nous pouvons alors former $T^2 = 737 \times (\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) = 1758.54$.

	L	M	N	O	P	Q
1		n	737			
2		p	5			
3						
4						
5		XX				
6		116162470.7	691905.8259	4471800.76	75374589.2	4932389.36
7		691905.8259	26356.55419	83946.7115	1754000.912	101326.582
8		4471800.762	83946.71151	688069.385	5394917.924	351219.04
9		75374589.2	1754000.912	5394917.92	1963980945	16238551
10		4932389.356	101326.5822	351219.04	16238550.98	3986370.36
11						
12		S				
13		157829.4	940.1	6075.8	102411.1	6701.6
14		940.1	35.8	114.1	2383.2	137.7
15		6075.8	114.1	934.9	7330.1	477.2
16		102411.1	2383.2	7330.1	2668452.4	22063.2
17		6701.6	137.7	477.2	22063.2	5416.3
18						
19		S^(-1)				
20		8.743E-06	-5.375E-05	-4.686E-05	-1.188E-07	-4.839E-06
21		-5.375E-05	5.008E-02	-5.249E-03	-2.286E-05	-6.508E-04
22		-4.686E-05	-5.249E-03	2.002E-03	8.931E-07	1.139E-05
23		-1.188E-07	-2.286E-05	8.931E-07	4.056E-07	-1.003E-06
24		-4.839E-06	-6.508E-04	1.139E-05	-1.003E-06	2.102E-04
25						
26		Moyenne empirique				
27		calcium	iron	protein	vitamin.a	vitamin.c
28		624.0	11.1	65.8	839.6	78.9
29						
30		Valeurs de référence				
31		calcium	iron	protein	vitamin.a	vitamin.c
32		1000	15	60	800	75
33						
34		Ecart				
35		-376.0	-3.9	5.8	39.6	3.9
36						
37						
38		T^2	**1758.54**			
39						
40		F	**349.80**			
41						
42		ddl1	5			
43		ddl2	732			
44						
45		F_0.95(5,732)	2.2263			
46						
47		p-value	0.00000			

Fig. 5.3. Analyse multivariée - Comparaison de la moyenne à un standard

- Cette valeur est difficilement utilisable directement, nous la transformons pour obtenir une statistique distribuée selon la loi de Fisher, $F = \frac{737-5}{5(737-1)} \times 1758.54 = 349.80$
- Les degrés de liberté sont $ddl_1 = p = 5$ et $ddl_2 = n - p = 737 - 5 = 732$, au risque $\alpha = 5\%$, le seuil critique est $F_{0.95}(5, 732) = 2.2263$.
- Manifestement, $F > F_{0.95}(5, 732)$, nous concluons au rejet de l'hypothèse nulle. En moyenne les femmes ne suivent pas les recommandations de l'administration en matière de nutrition.

6

T^2 de Hotelling - Comparaison de $K = 2$ populations

Nous revenons dans ce chapitre au coeur de notre sujet. Nous voulons savoir si K échantillons proviennent de la même population ou non. Nous situant dans un cadre paramétrique, X étant censée suivre une loi multinormale, les comparaisons seront essentiellement basées sur la confrontation des moyennes empiriques $\bar{X}_k$ et des matrices de variances covariances empiriques S_k.

Notre trame sera la même que dans la première partie de ce support. La différence est que nous tenons compte dorénavant de la situation simultanée de p variables d'intérêt. Nous utiliserons en priorité les outils statistiques développés dans le chapitre précédent (chapitre 5).

6.1 Comparaison de moyennes - 2 échantillons indépendants, homoscédasticité

6.1.1 Test, statistique du test et région critique

Dans cette section, nous souhaitons comparer les distributions de 2 sous échantillons Ω_1 et Ω_2 en nous basant sur la moyenne. Les échantillons sont indépendants.

Nous considérons que les matrices de variances conditionnelles sont inconnues, mais elles sont identiques. Nous aurons donc à produire une estimation commune de la matrice de variance covariance.

Les hypothèses à confronter sont

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

μ_k est un vecteur de moyennes de dimension $(p, 1)$. Nous pouvons préciser le test d'hypothèses en explicitant chaque dimension :

$$H_0 : \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{p1} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \vdots \\ \mu_{p2} \end{pmatrix}$$
$$H_1 : \exists j, \quad \mu_{j1} \neq \mu_{j2}$$

A partir des sous échantillons Ω_1 et Ω_2, nous produisons les estimations $\bar{X}_1$, $\bar{X}_2$, S_1 et S_2. Puisque nous faisons le pari de l'homoscédasticité, les matrices de variances covariances sont identiques dans les sous groupes, nous calculons une estimation commune (la matrice de variance covariance intra-classes)

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

La statistique du test s'écrit alors

$$T^2 = (\bar{X}_1 - \bar{X}_2)' \left[S_p(\frac{1}{n_1} + \frac{1}{n_2}) \right]^{-1} (\bar{X}_1 - \bar{X}_2) \tag{6.1}$$

On ne manquera par de faire le parallèle avec la statistique du test dans le cas univarié (équation 1.2), les similitudes sont très parlantes.

On retrouve la même statistique sous une autre écriture, totalement équivalente (voir [2], page 330 ; voir d'autres écritures dans [11], page 348) :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)' S_p^{-1} (\bar{X}_1 - \bar{X}_2) \tag{6.2}$$

Sous H_0, T^2 suit une loi de Hotelling. Il est plus pratique d'utiliser la transformation qui permet d'utiliser les tables de la loi de Fisher :

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \times T^2 \tag{6.3}$$

F suit une loi de Fisher à $(p, n_1 + n_2 - p - 1)$ degrés de liberté. Au risque α, la région critique du test est définie comme suit

$$R.C. : F \geq F_{1-\alpha}(p, n_1 + n_2 - p - 1)$$

Tout comme le test de Student dans le cas unidimensionnel, ce test multivarié est assez robuste vis à vis de l'hypothèse de distribution multinormale des X. En revanche, il est moins robuste par rapport à l'hypothèse d'égalité des matrices de variance covariance, plus particulièrement lorsque les effectifs sont déséquilibrés.

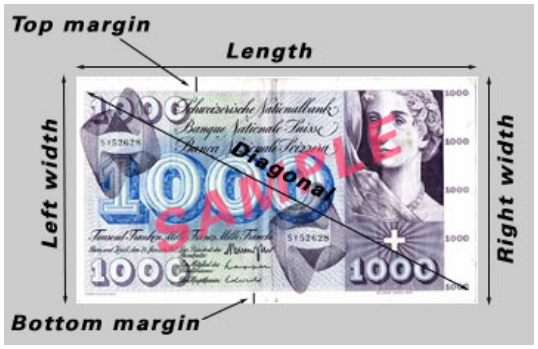


Fig. 6.1. Caractéristiques mesurées des billets de 1000 francs suisses

6.1.2 Un exemple : les billets de banque suisses

L'objectif de l'étude est de distinguer automatiquement les bons des mauvais billets de banque de 1000 francs suisses à partir de leurs caractéristiques physiques (longueur, largeur, ... $p = 6$ mesures en tout) (Figure 6.1, source : `http://www.stat.psu.edu/online/development/stat505/10_2sampHotel/01_2sampHotel_intro.html`). Nous disposons de $n_1 = 100$ billets authentiques et $n_2 = 100$ billets contrefaits, extraits aléatoirement dans leurs populations respectives. Nous avons donc 2 échantillons indépendants.

L'hypothèse nulle du test correspond à : toutes les mesures sont identiques, que le billet soit contrefait ou authentique; l'hypothèse alternative : les billets diffèrent au moins sur une des mesures, permettant de détecter les contrefaçons. Pour produire la statistique du test et statuer sur la réalité des écarts, nous élaborons une nouvelle feuille EXCEL (Figure 6.2) :

- En **R1..Y4**, nous calculons les moyennes conditionnelles. Nous observons des écarts, reste à savoir si elles sont statistiquement significatives. Pour les billets légaux "real", nous lisons : $n_1 = 100$, $\bar{x}_{11} = 214.969$, $\bar{x}_{21} = 129.943$, ..., $\bar{x}_{61} = 141.517$. Pour les imitations "fake", nous observons $n_2 = 100$, $\bar{x}_{12} = 214.823$, ..., $\bar{x}_{62} = 139.450$
- Nous produisons ensuite les matrices de variance covariance conditionnelles estimées S_k. La stratégie est la même : nous centrons la matrice X_k, comportant n_k lignes correspond au sous échantillon Ω_k, à l'aide de la moyenne conditionnelle $\bar{X}_k$. A partir de la matrice centrée $\overset{o}{X}_k$, nous obtenons la matrice S_k en appliquant la formule

$$S_k = \frac{1}{n_k - 1} \overset{o}{X}_k{}' \overset{o}{X}_k$$

  Ainsi, nous obtenons S_1 en **T9..Y14**, et S_2 en **T17..Y22**.
- La matrice de variance covariance intra-classes est obtenue avec $S_p = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$ en **T25..Y30**, son inverse S_p^{-1} en **T33..Y38**.
- Nous formons l'écart entre les moyennes $(\bar{X}_1 - \bar{X}_2)$ en **T42..Y42**.
- Le T^2 de Hotelling est obtenue à l'aide de la formule 6.2

$$T^2 = \frac{100 \times 100}{100 + 100}(\bar{X}_1 - \bar{X}_2)' S_p^{-1} (\bar{X}_1 - \bar{X}_2) = 2412.4507$$

	Q	R	S	T	U	V	W	X	Y
1			Données						
2		statut	Nombre de length	Moyenne de length	Moyenne de left.width	Moyenne de right.width	Moyenne de bot.margin	Moyenne de top.margin	Moyenne de diag.length
3		real	100	214.969	129.943	129.720	8.305	10.168	141.517
4		fake	100	214.823	130.300	130.193	10.530	11.133	139.450
5				**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
6									
7									
8						**S_1**			
9				0.150	0.058	0.057	0.057	0.014	0.005
10				0.058	0.133	0.086	0.057	0.049	-0.043
11				0.057	0.086	0.126	0.058	0.031	-0.024
12				0.057	0.057	0.058	0.413	-0.263	0.000
13				0.014	0.049	0.031	-0.263	0.421	-0.075
14				0.005	-0.043	-0.024	0.000	-0.075	0.200
15									
16						**S_2**			
17				0.124	0.032	0.024	-0.101	0.019	0.012
18				0.032	0.065	0.047	-0.024	-0.012	-0.005
19				0.024	0.047	0.089	-0.019	0.000	0.034
20				-0.101	-0.024	-0.019	1.281	-0.490	0.238
21				0.019	-0.012	0.000	-0.490	0.404	-0.022
22				0.012	-0.005	0.034	0.238	-0.022	0.311
23									
24						**S_p**			
25				0.137	0.045	0.041	-0.022	0.017	0.009
26				0.045	0.099	0.066	0.016	0.019	-0.024
27				0.041	0.066	0.108	0.020	0.015	0.005
28				-0.022	0.016	0.020	0.847	-0.377	0.119
29				0.017	0.019	0.015	-0.377	0.413	-0.049
30				0.009	-0.024	0.005	0.119	-0.049	0.256
31									
32						**inverse(S_p)**			
33				8.931	-3.498	-1.315	0.551	0.247	-0.810
34				-3.498	19.849	-10.711	-1.004	-0.973	2.486
35				-1.315	-10.711	16.562	-0.233	-0.445	-1.278
36				0.551	-1.004	-0.233	2.189	1.939	-0.759
37				0.247	-0.973	-0.445	1.939	4.220	-0.191
38				-0.810	2.486	-1.278	-0.759	-0.191	4.519
39									
40						Ecart des moyennes			
41				length	left.width	right.width	bot.margin	top.margin	diag.length
42				0.146	-0.357	-0.473	-2.225	-0.965	2.067
43									
44									
45				T^2	2412.4507				
46									
47				F	391.9217				
48									
49				ddl1	6				
50				ddl2	193				
51									
52				F_0.95(6,193)	2.1458				
53									
54				p-value	0.00000				

Fig. 6.2. Comparaison des billets de 1000 francs suisses - Hypothèse d'homoscédasticité

- En tant que telle, cette valeur n'est pas très utilisable, mieux vaut passer par la transformée F

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \times T^2 = \frac{100 + 100 - 6 - 1}{6(100 + 100 - 2)} \times 2412.4507 = 391.9217$$

- Les degrés de libertés sont $ddl_1 = p = 6$ et $ddl_2 = n_1 + n_2 - p - 1 = 100 + 100 - 6 - 1 = 193$
- Le seuil critique du test au risque $\alpha = 5\%$ est le quantile d'ordre 0.95 de la loi de Fisher c.-à-d. $F_{0.95}(6, 193) = 2.1458$.
- Nous nous situons dans la zone de rejet de H_0. Il est possible de distinguer les billets à l'aide des caractéristiques mesurées.

C'est très bien tout cela, mais la question qui vient automatiquement derrière est : sur quelles dimensions les billets diffèrent ? Quelles sont les mesures à effectuer en priorité pour distinguer les faux billets ?

6.1.3 Évaluer les écarts sur une des variables

Dans le cas où nous rejetons l'hypothèse nulle du test, nous savons que les moyennes diffèrent au moins sur une des variables X_j. Pour détecter les variables responsables de l'écart, il ne faut surtout pas procéder à des tests individuels, nous ne tiendrions pas compte des interdépendances entre les variables. La bonne démarche repose sur l'ellipsoïde de confiance de l'écart entre les moyennes [1].

L'équation de l'ellipsoïde de confiance de l'écart est définie par l'égalité entre la statistique du test F et le seuil critique du test d'égalité des moyennes $F_{1-\alpha}(p, n_1+n_2-p-1)$ (pour simplifier l'écriture, nous noterons $F_{1-\alpha}$) c.-à-d.

$$\frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \times T^2 = F_{1-\alpha}$$

L'ellipsoïde de confiance permet de déduire les intervalles de confiance simultanés des écarts. Pour la variable X_j, elles est définie de la manière suivante

$$\bar{x}_{j1} - \bar{x}_{j2} \pm \sqrt{\frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{1-\alpha}} \times \sqrt{(\frac{1}{n_1}+\frac{1}{n_2})s_j^2} \tag{6.4}$$

$s_j^2 = s_{ll}$ est lue sur la diagonale principale de matrice de variance covariance intra classes.

Si l'intervalle de confiance contient la valeur 0, cela veut dire que l'écart n'est pas significatif au risque α que l'on s'est choisi.

Comparer la longueur des billets de banques

Nous avons conclu à la différence significative des dimensions des billets de banque légaux et illicites (section 6.1.2). Voyons si elle est imputable à leur longueur (variable "length" - X_1). Nous utiliserons l'équation 6.4 en nous appuyant sur les résultats produits précédemment (Figure 6.2) :

- Nous avons $p = 6$, $n_1 = n_2 = 100$, $\bar{x}_{11} = 214.969$, $\bar{x}_{12} = 214.823$
- $s_1^2 = s_{11} = 0.137$ est lue dans la première cellule de la matrice S_p
- $F_{0.95} = 2.1458$ est le quantile d'ordre 0.95 de la loi de Fisher
- $\sqrt{\frac{p(n_1+n_2-2)}{n_1+n_2-p-1}F_{1-\alpha}} = \sqrt{\frac{6(100+100-2)}{100+100-6-1}2.1458} = \sqrt{13.2084} = 3.6343$
- $\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})s_j^2} = \sqrt{(\frac{1}{100}+\frac{1}{100})0.137} = \sqrt{0.0027} = 0.0524$

1. Pour plus de détails, notamment pour savoir comment tracer l'ellipse en pratique, voir `http://www.stat.psu.edu/online/development/stat505/10_2sampHotel/05_2sampHotel_differ.html`. Le texte est réellement instructif.

– L'intervalle de confiance de l'écart au niveau $1 - \alpha = 95\%$ entre les moyennes pour la variables X_1 s'écrit alors

$$(214.969 - 214.823) - 3.6343 \times 0.0524 \ ; (214.969 - 214.823) + 3.6343 \times 0.0524$$
$$(-0.0443 \ ; 0.3363)$$

La différence entre les billets n'est pas imputable à la longueur ("length"), l'intervalle de confiance recouvre la valeur 0.

En réitérant les calculs sur l'ensemble des variables, nous obtenons le tableau 6.1 (Source : `http://www.stat.psu.edu/online/development/stat505/10_2sampHotel/06_2sampHotel_example.html`). Tous les intervalles qui ne contiennent pas la valeur 0 sont responsables du rejet de l'hypothèse nulle du test de comparaison de moyennes.

Variable	95% I.C.
Length	$(-0.044; 0.336)$
Left.Width	$(-0.519; -0.195)$
Right.Width	$(-0.642; -0.304)$
Bottom.Margin	$(-2.698; -1.752)$
Top.Margin	$(-1.295; -0.635)$
Diagonal.Length	$(1.807; 2.327)$

Tableau 6.1. Intervalle de confiance des écarts entre moyennes - Billets de 1000 francs suisses

6.1.4 Tester une combinaison linéaire des écarts

La formulation du test individuel (section 6.1.3) provient d'un schéma plus général. Nous pouvons tester la significativité une combinaison linéaire des écarts. Il faut définir judicieusement les coefficients pour obtenir la confrontation désirée.

Le test d'hypothèses devient :

$$H_0 : a'(\mu_1 - \mu_2) = 0$$
$$H_1 : a'(\mu_1 - \mu_2) \neq 0$$

a est un vecteur de dimension $(p, 1)$, avec $a' = (a_1, a_2, \ldots, a_p)$.

L'ellipsoïde de confiance de la combinaison linéaire au niveau $1 - \alpha$ s'écrit alors

$$\sum_j a_j(\bar{x}_{j1} - \bar{x}_{j2}) \pm \sqrt{\frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_{1-\alpha}} \times \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \sum_l \sum_c a_l a_c s_{lc}}$$

s_{lc} est lue dans la matrice de variance covariance intra classes S_p.

Un vecteur a composé de valeurs nulles, sauf pour le coefficient $a_j = 1$, permet de retrouver le test unitaire de significativité de l'écart portant sur la variable X_j (section 6.1.3).

6.1.5 Tester tous les écarts

Le cadre ci-dessus convient si nous nous intéressons plus particulièrement à un écart sur une des variables, ou à une combinaison linéaire d'écarts. En revanche, si le véritable objectif est de tester tous les écarts (comme nous le suggérons dans le tableau 6.1), une procédure qui tient compte de la multiplicité des tests est plus appropriée : l'intervalle de confiance est calculée en introduisant la correction de Bonferroni c.-à-d.

$$\bar{x}_{j1} - \bar{x}_{j2} \pm t_{1-\frac{\alpha}{2p}} \sqrt{s_j^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

où $t_{1-\frac{\alpha}{2p}}$ est le quantile d'ordre $1-\frac{\alpha}{2p}$ de la loi de Student à (n_1+n_2-2) degrés de liberté (voir `http://www.stat.psu.edu/online/development/stat505/10_2sampHotel/08_2sampHotel_CI.html`).

6.2 Comparaison de moyennes - 2 échantillons indépendants, hétéroscédasticité

Lorsque les effectifs sont déséquilibrés c.-à-d. n_1 très différent de n_2, le test précédent n'est pas très robuste lorsque l'on s'écarte, même très faiblement, de l'hypothèse d'homoscédasticité. Plutôt que de se lancer dans un test d'égalité des matrices de variance covariance (que nous présenterons par ailleurs) pour assurer l'affaire, nous avons tout intérêt à utiliser la variante que nous présentons dans cette section. Le parallèle avec le cas unidimensionnel est frappant (voir la section 1.2.3).

6.2.1 Statistique du test

Pour le test d'égalité des barycentres lorsque les matrices de variance covariances sont différentes, on utilisera la statistique

$$T^2 = (\bar{X}_1 - \bar{X}_2)' \left[\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right]^{-1} (\bar{X}_1 - \bar{X}_2) \tag{6.5}$$

Il n'est plus question de calculer une matrice commune des dispersions dans ce cas. Pour retomber sur les distributions usuelles, 2 transformations sont possibles, selon la taille de l'échantillon de données.

6.2.2 Région critique pour les grands effectifs

Lorsque les effectifs sont élevés, nous pouvons utiliser directement T^2, elle suit asymptotiquement une loi du χ^2 à p degrés de liberté. La région critique au risque α est donc

$$T^2 \geq \chi^2_{1-\alpha}(p)$$

6.2.3 Région critique pour les petits effectifs

L'affaire se corse lorsque les effectifs sont faibles, on doit passer par une transformation qui suit une loi de Fisher, plus précise dans ce cas, mais diablement plus complexe aussi [2].

Sous H_0, la quantité F suit une loi de Fisher à (p, ν) degrés de liberté, avec

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \times T^2$$

et

$$\frac{1}{\nu} = \sum_{k=1}^{2} \frac{1}{n_k - 1} \left[\frac{(\bar{X}_1 - \bar{X}_2)' S_T (\frac{1}{n_k} S_k)^{-1} S_T (\bar{X}_1 - \bar{X}_2)}{T^2} \right]^2$$

où

$$S_T = \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2$$

La différence est significative pour les grande valeurs de F c.-à-d.

$$R.C. : F \geq F_{1-\alpha}(p, \nu)$$

6.2.4 Application aux billets de 1000 francs suisses

Nous appliquons la variante adaptée aux échantillons de taille suffisamment grande, basée sur la loi du χ^2 sur notre exemple des billets de banque. L'organisation de la feuille de calcul est assez similaire à la précédente (section 6.1.2), les différences surviennent lorsque nous aurons à estimer la matrice de variance covariance commune (Figure 6.3) :

- De la même manière que pour le cas homoscédastique (Figure 6.2), nous calculons les moyennes conditionnelles, les écarts, et les matrices de variance covariance conditionnelles S_k
- Principale nouveauté, nous calculons maintenant la matrice $S_T = \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2$ (en **T25..Y30**), que nous inversons (**T33..Y38**).
- Nous appliquons alors la formule 6.5, nous obtenons $T^2 = 2412.45$
- Au risque $\alpha = 5\%$, nous le comparons au quantile $\chi^2_{0.95}(6) = 12.59$. Nous rejetons l'hypothèse d'égalité des moyennes, T^2 est largement supérieur au seuil critique du test.

Remarque 9 (Une étrange similitude entre les valeurs du T^2). Non sans surprises, nous constatons que le T^2 est identique que l'on s'appuie ou non l'hypothèse d'homoscédasticité (Figure 6.2 et 6.3). La raison est que nous nous situons dans un cadre bien particulier dans cet exemple, les effectifs sont parfaitement équilibrés $n_1 = n_2$. Autrement, de manière générale, les T^2 diffèrent.

2. oui, bon ben on va se contenter de la première formule dans notre tableur...

	Q	R	S	T	U	V	W	X	Y
1			Données						
2		statut	Nombre de length	Moyenne de length	Moyenne de left.width	Moyenne de right.width	Moyenne de bot.margin	Moyenne de top.margin	Moyenne de diag.length
3		real	100	214.969	129.943	129.720	8.305	10.168	141.517
4		fake	100	214.823	130.300	130.193	10.530	11.133	139.450
5				**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
6									
7									
8						**S_1**			
9				0.150	0.058	0.057	0.057	0.014	0.005
10				0.058	0.133	0.086	0.057	0.049	-0.043
11				0.057	0.086	0.126	0.058	0.031	-0.024
12				0.057	0.057	0.058	0.413	-0.263	0.000
13				0.014	0.049	0.031	-0.263	0.421	-0.075
14				0.005	-0.043	-0.024	0.000	-0.075	0.200
15									
16						**S_2**			
17				0.124	0.032	0.024	-0.101	0.019	0.012
18				0.032	0.065	0.047	-0.024	-0.012	-0.005
19				0.024	0.047	0.089	-0.019	0.000	0.034
20				-0.101	-0.024	-0.019	1.281	-0.490	0.238
21				0.019	-0.012	0.000	-0.490	0.404	-0.022
22				0.012	-0.005	0.034	0.238	-0.022	0.311
23									
24						**S_T**			
25				0.003	0.001	0.001	0.000	0.000	0.000
26				0.001	0.002	0.001	0.000	0.000	0.000
27				0.001	0.001	0.002	0.000	0.000	0.000
28				0.000	0.000	0.000	0.017	-0.008	0.002
29				0.000	0.000	0.000	-0.008	0.008	-0.001
30				0.000	0.000	0.000	0.002	-0.001	0.005
31									
32						**inverse(S_T)**			
33				446.572	-174.900	-65.753	27.557	12.367	-40.518
34				-174.900	992.461	-535.573	-50.198	-48.668	124.322
35				-65.753	-535.573	828.096	-11.635	-22.235	-63.917
36				27.557	-50.198	-11.635	109.426	96.971	-37.957
37				12.367	-48.668	-22.235	96.971	211.020	-9.549
38				-40.518	124.322	-63.917	-37.957	-9.549	225.926
39									
40						Ecart des moyennes			
41				length	left.width	right.width	bot.margin	top.margin	diag.length
42				0.146	-0.357	-0.473	-2.225	-0.965	2.067
43									
44									
45				T²	2412.45				
46									
47				ddl	6				
48									
49				KHI2_0.95(6)	12.59				
50									
51				p-value	0.00000				

Fig. 6.3. Comparaison des billets de 1000 francs suisses - Hypothèse d'hétéroscédasticité

6.2.5 Tester un des écarts

De la même manière que pour le cas homoscédastique, nous avons la possibilité de tester les écarts individuellement, spécifiquement sur une des variables, en construisant les intervalles de confiance simultanés des écarts entre les moyennes observées. La formule à privilégier dépend des effectifs.

Lorsque les effectifs sont élevés, la formule suivante suffit largement pour tester l'écart imputable à la variable X_j

$$\bar{x}_{j1} - \bar{x}_{j2} \pm \sqrt{\chi^2_{1-\alpha}\left(\frac{s^2_{j1}}{n_1} + \frac{s^2_{j2}}{n_2}\right)} \tag{6.6}$$

où $\chi^2_{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de $\chi^2(p)$.

Lorsque les effectifs sont faibles, on préférera l'approximation plus précise suivante

$$\bar{x}_{j1} - \bar{x}_{j2} \pm \sqrt{\frac{p(n_1+n_2-2)}{n_1+n_2-p-1}F_{1-\alpha}} \times \sqrt{\left(\frac{s^2_{j1}}{n_1} + \frac{s^2_{j2}}{n_2}\right)} \tag{6.7}$$

où $F_{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de $F(p,\nu)$.

Les s^2_{jk} sont lues sur la diagonale principale des matrices de variance covariance conditionnelles S_k.

Exemple : le cas de la variable X_1 (length)

Les effectifs sont suffisamment élevés, nous utilisons l'approximation du χ^2 (Equation 6.6). L'intervalle de confiance de l'écart entre les moyennes s'obtient facilement (toutes les valeurs peuvent être retrouvées dans la feuille de calcul, figure 6.3) :

$$214.969 - 214.823 \pm \sqrt{12.59\left(\frac{0.15024}{100} + \frac{1.12401}{100}\right)}$$
$$(-0.040 \ ; \ 0.332)$$

L'intervalle contient la valeur 0, on peut considérer que l'écart global n'est pas imputable à la variable "length" (X_1).

6.3 Comparaison de moyennes - 2 échantillon appariés

6.3.1 Principe, statistique du test et région critique

L'objectif de l'appariement est de réduire la variabilité due aux observations. Le test de comparaison est plus puissante. Nous pouvons la mettre en oeuvre dans les schémas d'expérimentation "avant-après" (ex. mesurer la fièvre avant et après la prise d'un médicament), ou lorsque nous avons la possibilité en définissant l'échantillon de créer des couples d'observations, aussi semblables que possibles, que nous opposerons (voir le chapitre 4 pour plus de détails).

S'agissant du test de comparaison de moyennes, la démarche sera identique au test univarié (section 4.2) : nous créons un nouveau groupe de variables D formée à partir de l'écart entre les groupes X_1 et X_2, le test de comparaison de moyennes entre les variables initiales devient un test de comparaison à un standard de la variable transformée. Précisons cette idée dans le cadre multivarié.

Nous avons affaire à n observations décrites par deux groupes de p variables. Les variables sont comparables deux à deux c.-à-d. la variable $n^o j$ du premier groupe est directement comparable à la variable $n^o j$ du second groupe. Les matrices de données X_k sont donc de dimension (n,p). Soit μ_k le vecteur moyenne de la population $n^o k$, le test d'hypothèses s'écrit toujours

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Nous formons la matrice $D = X_1 - X_2$ avec $d_{ij} = x_{ij1} - x_{ij2}$, toujours de dimension (n, p). Le vecteur espérance de D est $E(D) = \mu_D$. Le test de comparaison ci-dessus peut alors s'écrire comme un test de comparaison à un standard, à savoir

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

Nous retrouvons des choses connues (voir section 5.3), nous donnons directement les résultats. Si $\bar{X}_D$ et S_D sont respectivement le vecteur moyenne empirique et la matrice de variance covariance empirique, la statistique du test s'écrit

$$T^2 = n\bar{X}_D' S_D^{-1} \bar{X}_D \tag{6.8}$$

La transformée F

$$F = \frac{n-p}{p(n-1)} T^2 \tag{6.9}$$

suit une loi de Fisher à $(p, n-p)$ degrés de liberté sous H_0. Nous rejetons l'hypothèse nulle pour les grandes valeurs de F, excédant le seuil critique du test au risque α.

6.3.2 Un exemple : la passion dans les ménages

Voilà un exemple bien singulier, on a posé une série de question relatives à la perception de leur passion réciproque à l'homme et à la femme de $n = 30$ ménages. Les valeurs varient sur une échelle de 1 à 5, une valeur élevée indique une évaluation très positive (Figure 6.4 ; voir `http://www.stat.psu.edu/online/development/stat505/data.htm` pour l'accès et la description des données, `http://www.stat.psu.edu/online/development/stat505/09_pHotel/05_pHotel_example.html` pour le traitement de l'exemple).

L'objectif est de vérifier si l'homme et la femme, dans le même couple, ont la même perception de leur passion commune.

Après la construction de la matrice des différences D, la construction de la feuille de calcul est très proche du test de comparaison à un standard (Figure 6.5) :

- Nous avons $n = 30$ observations et $p = 4$ variables.
- Nous formons la matrice des écarts D en **J3..M32**. Nous calculons alors le vecteur moyenne $\bar{D} = (0.0667; -0.1333; -0.3000; -0.1333)'$.

	A	B	C	D	E	F	G	H
1	Mari (X1)				Epouse (X2)			
2	q1	q2	q3	q4	q1	q2	q3	q4
3	2	3	5	5	4	4	5	5
4	5	5	4	4	4	5	5	5
5	4	5	5	5	4	4	5	5
6	4	3	4	4	4	5	5	5
7	3	3	5	5	4	4	5	5
8	3	3	4	5	3	3	4	4
9	3	4	4	4	4	3	5	4
10	4	4	5	5	3	4	5	5
11	4	5	5	5	4	4	5	4
12	4	4	3	3	3	4	4	4
13	4	4	5	5	4	5	5	5
14	5	5	4	4	5	5	5	5
15	4	4	4	4	4	4	5	5
16	4	3	5	5	4	4	4	4
17	4	4	5	5	4	4	5	5
18	3	3	4	5	3	4	4	4
19	4	5	4	4	5	5	5	5
20	5	5	5	5	4	5	4	4
21	5	5	4	4	3	4	4	4
22	4	4	4	4	5	3	4	4
23	4	4	4	4	5	3	4	4
24	4	4	4	4	4	5	4	4
25	3	4	5	5	2	5	5	5
26	5	3	5	5	3	4	5	5
27	5	5	3	3	4	3	5	5
28	3	3	4	4	4	4	4	4
29	4	4	4	4	4	4	5	5
30	3	3	5	5	3	4	4	4
31	4	4	3	3	4	4	5	4
32	4	4	5	5	4	4	5	5

Fig. 6.4. Données "Perception des sentiments réciproques entre époux"

- Pour obtenir la matrice de variance covariance de D, nous construisons la matrice des données centrées $\overset{o}{D}$ en **O3..R32**. Nous obtenons alors

$$S_D = \frac{1}{30-1} \overset{o}{D}{}' \overset{o}{D}$$

  Que nous inversons pour obtenir S_D^{-1} en **043..R46**.
- La statistique T^2 s'obtient avec

$$T^2 = 30 \times \bar{X}_D' S_D^{-1} \bar{X}_D = 13.2178$$

- Nous la transformons $F = \frac{30-4}{4(30-1)} 13.2178 = 2.9424$
- Au risque 5%, nous devons comparer F avec le seuil critique $F_{0.95}(4, 26) = 2.7426$. Nous sommes dans la région critique. Les écarts sont significatifs. L'homme et la femme à l'intérieur du couple n'ont pas la même perception de leur passion réciproque.
- La probabilité critique du test est p-value = 0.0394. Au risque 1%, nous aurions conclu à l'égalité des moyennes.

6.3.3 Significativité de l'écart sur une des variables en particulier

Si on s'intéresse à une des variables en particulier (ou plus généralement sur une combinaison linéaire des écarts), nous construisons l'intervalle de confiance simultanée de l'écart, on regarde si elle contient la valeur 0. La formule s'écrit pour le niveau de confiance $1 - \alpha$

$$\bar{D}_j \pm \sqrt{\frac{p(n-1)}{n-p} F_{1-\alpha}} \times \sqrt{\frac{s_{Dj}^2}{n}}$$

	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		D					D centré						
2		q1	q2	q3	q4		q1	q2	q3	q4			
3		-2	-1	0	0		-2.0667	-0.8667	0.3000	0.1333			
4		1	0	-1	-1		0.9333	0.1333	-0.7000	-0.8667			
5		0	1	0	0		-0.0667	1.1333	0.3000	0.1333			
6		0	-2	-1	-1		-0.0667	-1.8667	-0.7000	-0.8667		T^2	13.1278
7		-1	-1	0	0		-1.0667	-0.8667	0.3000	0.1333			
8		0	0	0	1		-0.0667	0.1333	0.3000	1.1333		F	2.9424
9		-1	1	-1	0		-1.0667	1.1333	-0.7000	0.1333			
10		1	0	0	0		0.9333	0.1333	0.3000	0.1333		ddl1	4
11		0	1	0	1		-0.0667	1.1333	0.3000	1.1333		ddl2	26
12		1	0	-1	-1		0.9333	0.1333	-0.7000	-0.8667			
13		0	-1	0	0		-0.0667	-0.8667	0.3000	0.1333		F_0.95(4,26)	2.7426
14		0	0	-1	-1		-0.0667	0.1333	-0.7000	-0.8667			
15		0	0	-1	-1		-0.0667	0.1333	-0.7000	-0.8667		p-value	0.0394
16		0	-1	1	1		-0.0667	-0.8667	1.3000	1.1333			
17		0	0	0	0		-0.0667	0.1333	0.3000	0.1333			
18		0	-1	0	1		-0.0667	-0.8667	0.3000	1.1333			
19		-1	0	-1	-1		-1.0667	0.1333	-0.7000	-0.8667			
20		1	0	1	1		0.9333	0.1333	1.3000	1.1333			
21		2	1	0	0		1.9333	1.1333	0.3000	0.1333			
22		-1	1	0	0		-1.0667	1.1333	0.3000	0.1333			
23		-1	1	0	0		-1.0667	1.1333	0.3000	0.1333			
24		0	-1	0	0		-0.0667	-0.8667	0.3000	0.1333			
25		1	-1	0	0		0.9333	-0.8667	0.3000	0.1333			
26		2	-1	0	0		1.9333	-0.8667	0.3000	0.1333			
27		1	2	-2	-2		0.9333	2.1333	-1.7000	-1.8667			
28		-1	-1	0	0		-1.0667	-0.8667	0.3000	0.1333			
29		0	0	-1	-1		-0.0667	0.1333	-0.7000	-0.8667			
30		0	-1	1	1		-0.0667	-0.8667	1.3000	1.1333			
31		0	0	-2	-1		-0.0667	0.1333	-1.7000	-0.8667			
32		0	0	0	0		-0.0667	0.1333	0.3000	0.1333			
33													
34	Moyenne D	0.0667	-0.1333	-0.3000	-0.1333								
35							S_D						
36							0.8230	0.0782	-0.0138	-0.0598			
37							0.0782	0.8092	-0.2138	-0.1563			
38							-0.0138	-0.2138	0.5621	0.5103			
39							-0.0598	-0.1563	0.5103	0.6023			
40													
41													
42							Inverse(S_D)						
43							1.2558	-0.1501	-0.4511	0.4679			
44							-0.1501	1.4115	0.9283	-0.4351			
45							-0.4511	0.9283	8.4243	-6.9420			
46							0.4679	-0.4351	-6.9420	7.4760			

Fig. 6.5. Traitement des données "Perception des sentiments réciproques entre époux"

Pour la première variable de notre exemple (Figure 6.5), nous aurions l'intervalle $(-0.51271; 0.64604)$ à partir du calcul

$$0.0667 \pm \sqrt{\frac{4(30-1)}{30-4} \times 2.7426} \times \sqrt{\frac{0.8230}{30}}$$

Remarque 10 (Tester toutes les variables). On se tournera sur l'intervalle de confiance corrigée de Bonferroni si on veut tester toutes les moyennes (voir `http://www.stat.psu.edu/online/development/stat505/09_pHotel/06_pHotel_CI.html`).

7

Comparaison de $K > 2$ populations

7.1 Λ de Wilks - MANOVA ou la généralisation de l'ANOVA à 1 facteur

7.1.1 Principe et statistique de test

Il s'agit maintenant de comparer les barycentres de K échantillons gaussiens indépendants dans un espace à p dimensions. Nous généralisons l'analyse de variance à 1 facteur (voir section 1.3), on parle d'ailleurs de MANOVA pour *Multivariate Analysis of Variance*.

Le test d'hypothèses s'écrit

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$
$$H_1 : \text{deux au moins des vecteurs moyennes sont différents}$$

La procédure repose sur la décomposition de la variance. Dans un espace à p variables, la version multivariée de l'équation d'analyse de variance $SCT = SCE + SCR$ devient :

$$\overset{o}{X}{}' \overset{o}{X} = \sum_k n_k (\bar{X}_k - \bar{X})'(\bar{X}_k - \bar{X}) + \sum_k \overset{o}{X}{}_k{}' \overset{o}{X}{}_k \tag{7.1}$$

Avec :

- SCT, SCE et SCR sont maintenant des matrices de dimensions (p, p) ;
- $\bar{X}$ est le vecteur des moyennes de dimension $(p, 1)$, calculé sur l'ensemble de l'échantillon ;
- $\overset{o}{X}$ est la matrice de données centrées de taille (n, p) (les données ont été centrées avec la moyenne globale $\bar{X}$) ;
- $\bar{X}_k$ est le vecteur des moyennes pour l'échantillon Ω_k ;
- $\overset{o}{X}{}_k$ est la sous matrice des données centrées de taille (n_k, p), correspondant au sous échantillon Ω_k (les données ont été centrées avec la moyenne locale $\bar{X}_k$).

La statistique du test de comparaison des K moyennes, connue sous l'appellation Λ de Wilks, est alors définie par le rapport

$$\Lambda = \frac{\det(SCR)}{\det(SCT)} \tag{7.2}$$

Rappelons que le déterminant d'une matrice de variance covariance (à un facteur près ici) se lit comme une variance généralisée (voir section 5.2.1, "Distance de Mahalanobis et Variance généralisée"). Le Λ de Wilks peut donc s'interpréter comme le rapport entre la variabilité intra-classes et la variabilité totale, avec les propriétés suivantes

- $0 \leq \Lambda \leq 1$
- $\Lambda \to 1$: l'appartenance aux groupes n'explique en rien la variabilité des données, $SCR \to SCT$, et par conséquent $SCE \to 0$, les barycentres conditionnels sont confondus avec le barycentre global. On a tendance à accepter l'hypothèse nulle.
- $\Lambda \to 0$: pour chaque sous population (groupe), les points sont agglutinés autour de leur barycentres respectifs, $SCE \to 1$, les moyennes conditionnelles sont très différentes les unes des autres. On sera emmené à rejeter l'hypothèse nulle.

La distribution de la statistique de Wilks est compliquée, difficilement accessible. Nous ne pouvons donc pas définir la région critique du test à partir de Λ. Nous verrons plus loin qu'il est commode d'utiliser des transformations qui permettent de retrouver les lois de répartition usuelles (χ^2 et loi de Fisher).

7.1.2 Un exemple : distinguer les poteries selon leur composition

Des poteries ont été échantillonnées sur 4 îles britanniques différentes. On cherche à savoir si leurs compositions en Aluminium, Fer (Iron), Magnésium, Calcium et Sodium sont identiques selon le site de collecte. Le tableau de données comporte $n = 26$ observations (poteries) et 6 variables, la première étant une variable catégorielle indiquant le nom de l'île. Nous calculons directement les effectifs des sous échantillons, la moyenne globale et les moyennes conditionnelles (Figure 7.1). L'objectif est de statuer sur le caractère significatif de l'écart entre ces moyennes.

La première étape est de construire les matrices de l'équation d'analyse de variance (équation 7.1).

SCT Pour obtenir la matrice des "sommes des carrés totaux" (SCT), nous centrons les données, la matrice $\overset{o}{X}$ est ainsi construire, puis nous produisons le produit matriciel $SCT = \overset{o}{X}{}' \overset{o}{X}$ (Figure 7.2).

SCE A partir des écarts entre les moyennes conditionnelles et la moyenne globale, nous construisons les matrices $n_k(\bar{X}_k - \bar{X})'(\bar{X}_k - \bar{X})$. Il ne reste plus qu'à faire la somme pour produire SCE (Figure 7.3).

SCR Enfin, la matrice SCR peut être obtenue par différence (SCR = SCT - SCE), ou construite explicitement à partir de $\sum_k \overset{o}{X}_k{}' \overset{o}{X}_k$. Nous choisissons la première solution (Figure 7.4).

Le Λ de Wilks est obtenu via le rapport

Site	Aluminium	Iron	Magnesiur	Calcium	Sodium
Llanedryn	14.4	7	4.3	0.15	0.51
Llanedryn	13.8	7.08	3.43	0.12	0.17
Llanedryn	14.6	7.09	3.88	0.13	0.2
Llanedryn	11.5	6.37	5.64	0.16	0.14
Llanedryn	13.8	7.06	5.34	0.2	0.2
Llanedryn	10.9	6.26	3.47	0.17	0.22
Llanedryn	10.1	4.26	4.26	0.2	0.18
Llanedryn	11.6	5.78	5.91	0.18	0.16
Llanedryn	11.1	5.49	4.52	0.29	0.3
Llanedryn	13.4	6.92	7.23	0.28	0.2
Llanedryn	12.4	6.13	5.69	0.22	0.54
Llanedryn	13.1	6.64	5.51	0.31	0.24
Llanedryn	12.7	6.69	4.45	0.2	0.22
Llanedryn	12.5	6.44	3.94	0.22	0.23
Caldicot	11.8	5.44	3.94	0.3	0.04
Caldicot	11.6	5.39	3.77	0.29	0.06
Isle.Thornes	18.3	1.28	0.67	0.03	0.03
Isle.Thornes	15.8	2.39	0.63	0.01	0.04
Isle.Thornes	18	1.5	0.67	0.01	0.06
Isle.Thornes	18	1.88	0.68	0.01	0.04
Isle.Thornes	20.8	1.51	0.72	0.07	0.1
Ashley.Rails	17.7	1.12	0.56	0.06	0.06
Ashley.Rails	18.3	1.14	0.67	0.06	0.05
Ashley.Rails	16.7	0.92	0.53	0.01	0.05
Ashley.Rails	14.8	2.74	0.67	0.03	0.05
Ashley.Rails	19.1	1.64	0.6	0.1	0.03

	Données					
Site	Nombre de	X1	X2	X3	X4	X5
Ashley.Rails	5	17.32	1.51	0.61	0.05	0.05
Caldicot	2	11.70	5.42	3.86	0.30	0.05
Isle.Thornes	5	18.18	1.71	0.67	0.03	0.05
Llanedryn	14	12.56	6.37	4.83	0.20	0.25
Total	26	14.49	4.47	3.14	0.15	0.16

Fig. 7.1. Données "Poterie", effectifs et moyennes conditionnelles

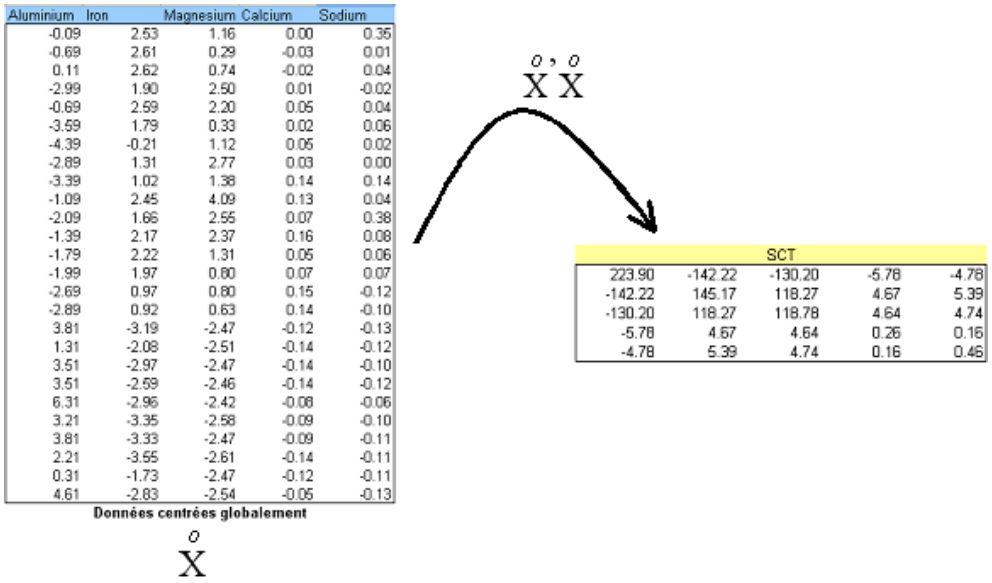


Fig. 7.2. Données "Poterie" - Construction de la matrice SCT

$$\Lambda = \frac{\det(SCR)}{\det(SCT)} = \frac{50.02016}{4066.38026} = 0.01230$$

$\Lambda = 0.01230$ est très petit, proche de 0. A vue de nez, nous nous écartons résolument de l'hypothèse nulle. On devrait la rejeter. Mais une démarche statistique doit inscrire la décision dans un cadre probabi-

$\overline{X}_1 - \overline{X}$

Ecarts à la moyenne globale

Ashley.Rail	2.83	-2.96	-2.54	-0.09	-0.11
Caldicot	-2.79	0.95	0.71	0.15	-0.11
Isle.Thornes	3.69	-2.76	-2.47	-0.12	-0.10
Llanedryn	-1.93	1.90	1.68	0.06	0.09

$n_1(\overline{X}_1 - \overline{X})'(\overline{X}_1 - \overline{X})$

Formation de SCE

Ashley.Rail	39.98	-41.79	-35.85	-1.34	-1.56
	-41.79	43.68	37.47	1.40	1.63
	-35.85	37.47	32.14	1.20	1.40
	-1.34	1.40	1.20	0.04	0.05
	-1.56	1.63	1.40	0.05	0.06
Caldicot	15.59	-5.29	-3.98	-0.83	0.61
	-5.29	1.79	1.35	0.28	-0.21
	-3.98	1.35	1.02	0.21	-0.15
	-0.83	0.28	0.21	0.04	-0.03
	0.61	-0.21	-0.15	-0.03	0.02
Isle.Thornes	68.00	-50.81	-45.50	-2.22	-1.93
	-50.81	37.97	34.00	1.66	1.44
	-45.50	34.00	30.44	1.49	1.29
	-2.22	1.66	1.49	0.07	0.06
	-1.93	1.44	1.29	0.06	0.05
Llanedryn	52.04	-51.41	-45.48	-1.50	-2.49
	-51.41	50.78	44.92	1.48	2.46
	-45.48	44.92	39.74	1.31	2.18
	-1.50	1.48	1.31	0.04	0.07
	-2.49	2.46	2.18	0.07	0.12

SCE

175.61	-149.30	-130.81	-5.89	-5.37
-149.30	134.22	117.75	4.82	5.33
-130.81	117.75	103.35	4.21	4.71
-5.89	4.82	4.21	0.20	0.15
-5.37	5.33	4.71	0.15	0.26

Fig. 7.3. Données "Poterie" - Construction de la matrice SCE

SCR

48.29	7.08	0.61	0.11	0.59
7.08	10.95	0.53	-0.16	0.07
0.61	0.53	15.43	0.44	0.03
0.11	-0.16	0.44	0.05	0.01
0.59	0.07	0.03	0.01	0.20

SCT

223.90	-142.22	-130.20	-5.78	-4.78
-142.22	145.17	118.27	4.67	5.39
-130.20	118.27	118.78	4.64	4.74
-5.78	4.67	4.64	0.26	0.16
-4.78	5.39	4.74	0.16	0.46

SCE

175.61	-149.30	-130.81	-5.89	-5.37
-149.30	134.22	117.75	4.82	5.33
-130.81	117.75	103.35	4.21	4.71
-5.89	4.82	4.21	0.20	0.15
-5.37	5.33	4.71	0.15	0.26

Fig. 7.4. Données "Poterie" - Construction de la matrice SCR = SCT - SCE

liste. Pour cela, il nous faut associer à Λ, ou à une de ses transformations, une distribution de probabilité connue et d'usage courant (si possible).

7.1.3 Transformations usuelles et régions critiques du test

Transformation de Bartlett

Lorsque les effectifs sont élevés, la transformation de Bartlett est suffisante, elle a le mérite de la simplicité (voir [2], page 331) :

$$\chi^2 = -\left(n - 1 - \frac{p+K}{2}\right) \ln \Lambda \tag{7.3}$$

Sous H_0, elle suit une loi du χ^2 à $[p(K-1)]$ degrés de liberté. On rejette l'hypothèse nulle si la statistique calculée dépasse le seuil critique.

Application aux données "Poterie". Appliquons directement cette formule 7.3 sur nos données :

$$\chi^2 = -\left(26 - 1 - \frac{5+4}{2}\right) \ln(0.01230) = 90.1607$$

Les degrés de liberté sont $p \times (K-1) = 5 \times (4-1) = 15$. La p-value du test est < 0.00001. Nous rejetons l'hypothèse nulle. Néanmoins, l'effectif étant assez faible, l'approximation de Bartlett n'est pas très bonne, la conclusion est à prendre avec prudence.

Transformation de Rao

De manière générale, et plus particulièrement sur les petits effectifs, nous avons intérêt à utiliser l'approximation de Rao [1]. Elle est plus performante, en revanche elle est assez rédhibitoire si on doit la calculer manuellement.

$$F = \left(\frac{1-\Lambda^{1/b}}{\Lambda^{1/b}}\right)\left(\frac{ab-c}{p(K-1)}\right) \tag{7.4}$$

Sous H_0, elle suit une loi de Fisher à $[p(K-1), ab-c]$ degrés de liberté. Détaillons les éléments qui composent F :

$$\begin{aligned}
a &= n - K - \frac{p-K+2}{2} \\
b &= \begin{cases} \sqrt{\frac{p^2(K-1)^2-4}{p^2+(K-1)^2-5}} & \text{si} \quad p^2+(K-1)^2 > 5 \\ 1 & \text{si} \quad p^2+(K-1)^2 \leq 5 \end{cases} \\
c &= \frac{p(K-1)-2}{2}
\end{aligned}$$

Les degrés de liberté peuvent être fractionnaires. Nous pouvons : soit utiliser une interpolation pour obtenir le bon seuil critique du test ; soit, plus simplement, arrondir la valeur à l'entier le plus proche (ce que font la très grande majorité des logiciels pour calculer la p-value).

Application aux données "Poterie". Calculons les différentes quantités ci-dessus :

$$\begin{aligned}
a &= 26 - 4 - \frac{5-4+2}{2} = 20.5 \\
b &= \sqrt{\frac{5^2(4-1)^2-4}{5^2+(4-1)^2-5}} = 2.7606 \quad \text{puisque} \quad 5^2+(4-1)^2 = 34 > 5 \\
c &= \frac{5(4-1)-2}{2} = 6.5 \\
\Lambda^{1/b} &= 0.01230^{1/2.7606} = 0.2033 \\
ddl_1 &= 5(4-1) = 15 \\
ddl_2 &= 20.5 \times 2.7606 - 6.5 = 50.09
\end{aligned}$$

Et enfin

$$F = (\frac{1-0.2033}{0.2033}) \times (\frac{20.5 \times 2.7606 - 6.5}{5(4-1)}) = 13.08854$$

1. Voir `http://www.stat.psu.edu/online/development/stat505/11_1wMANOVA/04_1wMANOVA_prob.html` ; la distribution est exacte si $\min(p, K-1) \leq 2$

F suit une loi de Fisher à $(15, 50)$ degrés de liberté, la p-value est < 0.00001. On rejette l'hypothèse nulle.

Les matrices et les ratios récapitulatifs sont présentés dans la figure 7.5. Dans notre feuille EXCEL, nous avions produit SCR indépendamment de SCE et SCT, la vérification $SCT = SCE + SCR$ nous a permis de valider la succession des calculs.

SCT				
223.90	-142.22	-130.20	-5.78	-4.78
-142.22	145.17	118.27	4.67	5.39
-130.20	118.27	118.78	4.64	4.74
-5.78	4.67	4.64	0.26	0.16
-4.78	5.39	4.74	0.16	0.46

SCR				
48.29	7.08	0.61	0.11	0.59
7.08	10.95	0.53	-0.16	0.07
0.61	0.53	15.43	0.44	0.03
0.11	-0.16	0.44	0.05	0.01
0.59	0.07	0.03	0.01	0.20

SCE				
175.61	-149.30	-130.81	-5.89	-5.37
-149.30	134.22	117.75	4.82	5.33
-130.81	117.75	103.35	4.21	4.71
-5.89	4.82	4.21	0.20	0.15
-5.37	5.33	4.71	0.15	0.26

SCT = SCE + SCR (vérification)				
223.90	-142.22	-130.20	-5.78	-4.78
-142.22	145.17	118.27	4.67	5.39
-130.20	118.27	118.78	4.64	4.74
-5.78	4.67	4.64	0.26	0.16
-4.78	5.39	4.74	0.16	0.46

det(SCR)	50.02016
det(SCT)	4066.38026

LAMBDA	0.0123009

n	26
p	5
K	4

Transf. Bartlett	
KHI2	90.1607
ddl	15

p-value	9.257E-13

Trans. Rao	
a	20.5
b	2.7606
c	6.5

Lambda^(1/b)	0.2033

F	13.08854
ddl1	15
ddl2	50.09

p-value	1.894E-12

Fig. 7.5. Données "Poterie" - χ^2 de Bartlett et F de Rao

7.1.4 D'autres statistiques de test : la trace de Pillai, la trace de Hotelling-Lawley, etc

D'autres statistiques sont utilisées pour répondre au test d'hypothèses de la MANOVA. Toutes exploitent d'une manière ou d'une autres les matrices SCT, SCE et SCR. Après c'est une question de préférences. Le Λ de Wilks est la plus connue. Il semble néanmoins que quelques unes des statistiques présentées dans cette section, la trace de Pillai notamment, soient préférables car plus robustes dans certaines configurations [2].

La trace de Pillai

La trace de Pillai, que l'on appelle aussi la trace de Pillai-Bartlett, est calculée de la manière suivante

2. Voir la documentation du logiciel R pour la classe MANOVA, citant l'ouvrage de D. Hand et C. Taylor, *Multivariate Analysis of Variance and Repeated Measures*, 1987. ; voir aussi `http://www2.chass.ncsu.edu/garson/PA765/manova.htm`

$$V = \mathrm{Tr}(SCE \times SCT^{-1}) \tag{7.5}$$

Rappelons que la trace d'une matrice carrée est la somme de ses éléments diagonaux [3].

Soient

- $s = \min(p, K-1)$
- $t = \frac{|p-K+1|-1}{2}$
- $u = \frac{n-K-p-1}{2}$

Alors la quantité

$$F = \left(\frac{2u+s+1}{2t+s+1}\right)\left(\frac{V}{s-V}\right) \tag{7.6}$$

Suit sous H_0 une loi de Fisher avec les degrés de liberté suivants

$$ddl_1 = s(2t+s+1)$$
$$ddl_2 = s(2u+s+1)$$

Application sur les données "Poterie". Dans la feuille EXCEL (Figure 7.6), les matrices SCT^{-1} et $SCE \times SCT^{-1}$ ont été calculées. Nous calculons alors les différents indicateurs :

- La trace de Pillai

$$V = 0.2426 + 0.5061 + 0.1775 + 0.4317 + 0.1961 = 1.55394$$

  C'est notre principale statistique, il nous faut la transformer pour passer à la loi de Fisher.
- $s = \min(p, K-1) = \min(5, 4-1) = 3$
- $t = \frac{|p-K+1|-1}{2} = \frac{|5-4+1|-1}{2} = 0.5$
- $u = \frac{n-K-p-1}{2} = \frac{26-4-5-1}{2} = 8$
- $ddl_1 = s(2t+s+1) = 3 \times (2 \times 0.5 + 3 + 1) = 15$
- $ddl_2 = s(2u+s+1) = 3 \times (2 \times 8 + 3 + 1) = 60$
- Et enfin

$$F = \left(\frac{2u+s+1}{2t+s+1}\right)\left(\frac{V}{s-V}\right) = \left(\frac{2 \times 8 + 3 + 1}{2 \times 0.5 + 3 + 1}\right)\left(\frac{1.55394}{3 - 1.55394}\right) = 4.29839$$

Que l'on doit comparer avec $F_{0.95} = 1.83644$ pour un test à 5%. Manifestement, nous sommes dans la région critique, on doit rejeter l'hypothèse nulle d'égalité des moyennes conditionnelles. La p-value du test qui est < 0.00001 confirme (bien évidemment) ce résultat.

3. http://fr.wikipedia.org/wiki/Trace_(algèbre)

Trace de Pillai

INV(SCT)

0.0147	0.0073	0.0052	0.1232	-0.0314
0.0073	0.0427	-0.0303	0.0116	-0.1176
0.0052	-0.0303	0.0682	-0.4911	-0.1178
0.1232	0.0116	-0.4911	14.7932	0.9069
-0.0314	-0.1176	-0.1178	0.9069	4.1365

SCE x INV(SCT)

0.2426	-0.5689	0.0422	-7.8430	-0.1117
-0.1654	0.5061	0.1903	1.4950	1.4373
-0.1462	0.4385	0.1775	1.0302	1.3889
-0.0087	0.0196	-0.0085	0.4317	-0.0520
-0.0042	0.0170	0.0255	-0.3896	0.1961

V	1.55394

s	3
t	0.5
u	8

F	4.29839

ddl1	15
ddl2	60

F_0.95(15,60)	1.83644

p-value	2.4129E-05

Fig. 7.6. Données "Poterie" - Trace de Pillai

La trace de Hotelling-Lawley

On retrouve également la trace de Hotelling-Lawley dans les logiciels. Sa formule est la suivante

$$U = \mathrm{Tr}(SCE \times SCR^{-1}) \tag{7.7}$$

Elle doit être transformée pour être compatible avec la loi de Fisher. En reprenant les indicateurs s, t et u développés dans la section précédente :

$$F = \frac{2(su+1)}{s^2(2t+s+1)}U \tag{7.8}$$

Sous H_0, F suit une loi de Fisher à (ddl_1, ddl_2) degrés de liberté, avec

$$ddl_1 = s(2t+s+1)$$
$$ddl_2 = 2(su+1)$$

Application sur les données "Poterie". Dans la feuille EXCEL, les matrices SCR^{-1} et $SCE \times SCR^{-1}$ ont été produites. Reste alors à calculer les résultats intermédiaires (Figure 7.7) :

Trace de Hotelling Lawley

INV(SCR)				
0.0242	-0.0170	0.0030	-0.1145	-0.0603
-0.0170	0.1106	-0.0178	0.5212	-0.0107
0.0030	-0.0178	0.0881	-0.8099	0.0259
-0.1145	0.5212	-0.8099	28.3061	-1.1553
-0.0603	-0.0107	0.0259	-1.1553	5.2544

SCE x INV(SCR)				
7.3922	-20.1884	-3.7165	-152.4787	-33.8135
-6.4138	17.7498	3.7751	122.0291	33.0385
-5.6224	15.5559	3.3352	106.3521	29.2004
-0.2446	0.6638	0.1058	5.3942	0.9896
-0.2397	0.6747	0.1856	3.6591	1.5674

U	35.4388

s	3
t	0.5
u	8

F	39.3764

ddl1	15
ddl2	50

F_0.95(15,50)	1.8714

p-value	1.95787E-22

Fig. 7.7. Données "Poterie" - Trace de Hotelling-Lawley

- La trace de Hotelling Lawley est la somme des valeurs sur la diagonale principale de $SCE \times SCR^{-1}$

$$U = 7.3922 + 17.7498 + 3.3352 + 5.3942 + 1.5674 = 35.4388$$

- De la même manière que précédemment, nous obtenons $s = 3$, $t = 0.5$ et $u = 8$.
- Calculons les degrés de liberté

$$ddl_1 = s(2t + s + 1) = 3(2 \times 0.5 + 3 + 1) = 15$$
$$ddl_2 = 2(su + 1) = 2(3 \times 8 + 1) = 50$$

- Reste à produire F

$$F = \frac{2(su+1)}{s^2(2t+s+1)}U = \frac{2(3 \times 8 + 1)}{3^2(2 \times 0.5 + 8 + 1)}35.4388 = 39.3764$$

- Le seuil critique du test au risque $\alpha = 5\%$ est $F_{0.95}(15, 50) = 1.8714$. F est largement supérieur au seuil, nous concluons au rejet de l'hypothèse nulle. Les vecteurs des moyennes conditionnelles diffèrent significativement.

La plus grande valeur propre de Roy

La statistique de Roy est similaire au test de Hotelling-Lawley, à la différence qu'on n'utilise que l'information la plus caractéristique de la matrice $SCE \times SCR^{-1}$. En effet, la statistique du test est la première (la plus grande) valeur propre λ_1 de cette matrice.

Notons que le niveau de signification calculé avec la procédure de Roy correspond à la borne inférieure du *véritable* niveau. En d'autre termes, le risque de première espèce réel est plus élevé que celui que l'on s'est choisi lors de la définition du test [4].

On utilise alors la transformation

$$F = \frac{n-r-1}{r} \times \lambda_1 \tag{7.9}$$

où $r = \max(p, K-1)$.

Sous H_0, F suit une loi de Fisher à (ddl_1, ddl_2) degrés de liberté, avec

$$ddl_1 = r$$
$$ddl_2 = n - r - 1$$

Application sur les données "Poterie". Calculer les valeurs propres dans EXCEL sans une macro complémentaire dédiée reste un peu ardu [5]. Nous récupérons directement le résultat du logiciel R, que nous présenterons de manière détaillée dans la section suivante, nous trouvons $\lambda_1 = 34.161$

Nous introduisons les expressions ci-dessus :

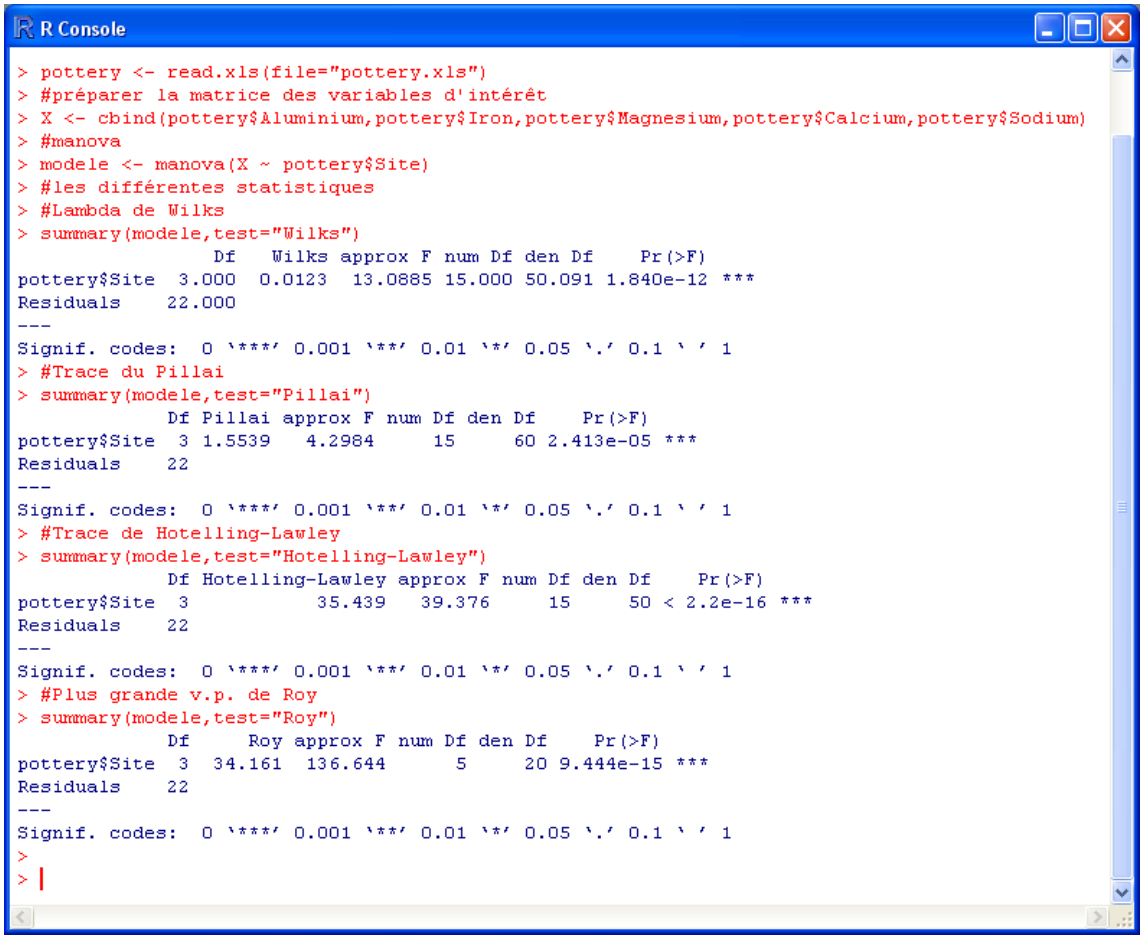
- $r = \max(p, K-1) = \max(5, 4-1) = 5$
- $ddl_1 = r = 5$
- $ddl_2 = n - r - 1 = 26 - 5 - 1 = 20$
- Et
$$F = \frac{n-r-1}{r} \times \lambda_1 = \frac{26-5-1}{5} \times 34.161 = 136.644$$
- Que l'on comparera avec le seuil critique $F_{0.95}(5, 20) = 2.7109$. On rejette l'hypothèse nulle d'égalité des moyennes car F est largement supérieur au seuil.

7.1.5 MANOVA avec le logiciel R

Les formules sont nombreuses et complexes dans cette section. Nous ne sommes pas à l'abri des coquilles malgré les recoupements entre différentes sources. Une (autre) bonne manière de s'assurer de l'exactitude des expressions est de les introduire dans un tableur pour suivre pas à pas la formation des indicateurs. C'est ce que nous nous évertuons à faire systématiquement dans ce support. Mais cela ne suffit pas, les équations peuvent comporter des erreurs, leur implémentation dans un tableur peut introduire d'autres types d'erreurs. Pour s'assurer de la qualité de notre texte, nous recoupons nos résultats avec ceux fournis par les outils qui font référence : les logiciels commerciaux comme les logiciels libres.

Parmi les logiciels libres, nous avons utilisé **R** (`http://www.r-project.org/`) dans cette section. Nous avons mis en oeuvre la fonction **manova**. Les différents indicateurs ci-dessus (Λ de Wilks, Trace de Pillai, etc., Figure 7.8) ont été calculés... et fort heureusement, c'est toujours rassurant, tous les résultats sont cohérents. Ils concordent également avec les résultats des logiciels commerciaux très répandus dont le nom commence par S...

4. Voir `http://www2.chass.ncsu.edu/garson/PA765/manova.htm`
5. On pourrait le faire avec l'outil SOLVEUR mais ça nous distrairait de notre véritable propos.



```
R Console
> pottery <- read.xls(file="pottery.xls")
> #préparer la matrice des variables d'intérêt
> X <- cbind(pottery$Aluminium,pottery$Iron,pottery$Magnesium,pottery$Calcium,pottery$Sodium)
> #manova
> modele <- manova(X ~ pottery$Site)
> #les différentes statistiques
> #Lambda de Wilks
> summary(modele,test="Wilks")
                 Df   Wilks approx F num Df den Df    Pr(>F)
pottery$Site  3.000  0.0123  13.0885 15.000 50.091 1.840e-12 ***
Residuals    22.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Trace du Pillai
> summary(modele,test="Pillai")
             Df Pillai approx F num Df den Df    Pr(>F)
pottery$Site  3 1.5539   4.2984     15     60 2.413e-05 ***
Residuals    22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Trace de Hotelling-Lawley
> summary(modele,test="Hotelling-Lawley")
             Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
pottery$Site  3           35.439   39.376     15     50 < 2.2e-16 ***
Residuals    22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Plus grande v.p. de Roy
> summary(modele,test="Roy")
             Df     Roy approx F num Df den Df    Pr(>F)
pottery$Site  3  34.161  136.644      5     20 9.444e-15 ***
Residuals    22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
```

Fig. 7.8. Données "Poterie" - Traitement MANOVA avec le logiciel R

7.2 Comparaison des matrices de variances covariances - Généralisation du test de Bartlett

7.2.1 Test, statistique du test et région critique

Le test d'égalité de variances peut être généralisé dans le cadre multidimensionnel, on parle toujours de test de Bartlett [6]. L'hypothèse nulle est l'égalité des matrices de variance covariances conditionnelles :

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_K \tag{7.10}$$

$$H_1 : \Sigma_k \neq \Sigma_{k'} \text{pour au moins une paire} k \neq k' \tag{7.11}$$

La procédure peut être mise en oeuvre pour sonder l'hypothèse d'homoscédasticité nécessaire à d'autres tests (la MANOVA par exemple), elle peut avoir aussi sa propre finalité, on cherche à savoir si la variabilité est la même dans différentes sous populations au regard des variables d'intérêt.

6. Bien que dans certains logiciels, on parle de *Box's M test*. Mais les formules et les distributions sont les mêmes

Avant de présenter la statistique de test elle-même, rappelons deux matrices importantes :

- $S_k = \frac{1}{n_k - 1} \overset{o}{X}_k{}' \overset{o}{X}_k$ est l'estimation non biaisée de la matrice de variance covariance dans le sous échantillon Ω_k
- $S_p = \frac{1}{n-k} \sum_k (n_k - 1) S_k$ est la matrice de variance covariance intra-classes, l'estimateur non biaisé de la matrice de variance covariance globale.

La statistique de Bartlett s'écrit

$$T = c \times \left[(n-K) \ln[\det(S_p)] - \sum_{k=1}^{K} (n_k - 1) \ln[\det(S_k)] \right] \tag{7.12}$$

où c est un coefficient correcteur,

$$c = 1 - \frac{2p^2 + 3p - 1}{6(K-1)(p+1)} \left(\sum_k \frac{1}{n_k - 1} - \frac{1}{n-K} \right)$$

Il s'agit d'une véritable généralisation. Dans un espace à une dimension ($p = 1$), cette expression sera confondue avec le test de Bartlett pour l'égalité des variances conditionnelles (section 2.3).

Sous H_0, T suit une loi du χ^2 à ν degrés de liberté, avec

$$\nu = \frac{(K-1)p(p+1)}{2}$$

La zone de rejet de l'hypothèse nulle correspond aux valeurs élevés de T.

Notons qu'à l'instar de son homologue univarié, ce test est très sensible aux écarts, mêmes faibles, par rapport à la distribution multinormale des données. Nous devons être très prudents quant à son utilisation.

7.2.2 Un exemple : analyser les clients selon la garantie contractée

Nous traitons notre fichier des demandeurs de crédits (Figure 0.1) pour le test multivarié d'homogénéité des variances. Nous cherchons à savoir si les matrices de variances covariances relatives aux variables "Salaire.Homme", "Salaire.Femme", "Rev.Tete" et "Age" sont identiques dans les groupes définies par la variable "Garantie.Supplémentaire".

Les calculs sont résumés dans la feuille Excel suivante (Figure) :

- Nous avons $K = 3$, avec $n_1 = 5$, $n_2 = 29$ et $n_3 = 16$. Ainsi $n = \sum_k n_k = 50$. La matrice de données comporte $p = 4$ variables.
- Les degrés de libertés sont ne plus facile à calculer. Réglons cela tout de suite

$$\nu = \frac{(K-1)p(p+1)}{2} = \frac{(3-1)4(4+1)}{2} = 20$$

- Les matrices conditionnelles S_k sont estimées à partir des sous échantillons concernés ($S_{caution}$, $S_{hypotheque}$ et S_{non}).

- Nous en déduisons la matrice de variance covariance intra classes S_p.
- Pour chacune des matrices, nous calculons le déterminant et le logarithme du déterminant
- Nous produisons alors la quantité M

$$\begin{aligned} M &= \left[(n-K)\ln[\det(S_p)] - \sum_{k=1}^{K}(n_k-1)\ln[\det(S_k)]\right] \\ &= [(50-3)\times(-8.77353) - ((5\times(-12.87193) + 29\times(-8.74984) + 16\times(-9.82761))] \\ &= 31.541630 \end{aligned}$$

- Puis le coefficient de correction c

$$\begin{aligned} c &= 1 - \frac{2p^2+3p-1}{6(K-1)(p+1)}\left(\sum_k \frac{1}{n_k-1} - \frac{1}{n-K}\right) \\ &= 1 - \frac{2\times 4^2+3\times 4-1}{6(3-1)(4+1)}\left(\frac{1}{5-1}+\frac{1}{29-1}+\frac{1}{16-1}-\frac{1}{50-3}\right) \\ &= 0.762709 \end{aligned}$$

- Nous multiplions les deux termes pour obtenir la statistique du test

$$T = c\times M = 0.762709\times 31.541630 = 24.057071$$

- Le seuil critique du test au risque $\alpha = 5\%$ est $\chi^2_{0.95}(20) = 31.4104$. La statistique calculée est inférieure au seuil, nous ne pouvons pas rejeter l'hypothèse d'homogénéité des matrices de variance covariance, elle est compatible avec les données.
- La p-value est égale à 0.2399

Ce test est important dans le cadre multivarié. En effet, la MANOVA sur des petits effectifs, avec des tailles de sous échantillon différents, n'est pas très robuste par rapport à un non respect de l'hypothèse d'homoscédasticité.

Dans la pratique, on constate souvent que l'hétéroscédasticité et la violation de l'hypothèse de multinormalité des distributions vont de pair.

A

Gestion des versions

Voici les versions successives de ce document :

1. La première version (1.0) de ce support a été finalisée et mise en ligne en Juillet 2008. Elle est accompagnée du fichier EXCEL qui contient tous les exemples traités - `http://eric.univ-lyon2.fr/~ricco/cours/cours/comp_pop_tests_parametriques.xls`
2. Version 1.1 : des tutoriels pour le logiciel Tanagra sont été élaborés. Ils sont référencés.
3. Version 1.2 : quelques coquilles très mineures ont été corrigées.

B

Tutoriels pour le logiciel Tanagra

Tanagra est un logiciel gratuit de statistique, d'analyse de données et de data mining. Mon idée est de développer de concert les supports de cours et les implémentation dans ce logiciel. L'utilisateur pourra ainsi, d'une part, reproduire les calculs dans Excel, ce qui est très intéressant pédagogiquement, mais aussi d'autre part, mettre en oeuvre les techniques à l'aide d'un logiciel libre qui respecte au mieux les standards des logiciels du marché.

Le logiciel est accessible sur le site suivant `http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html`. Les didacticiels sont regroupés dans un blog `http://tutoriels-data-mining.blogspot.com/`. Ils sont classés par thèmes, il est également possible de réaliser des recherches par mot-clés.

Concernant les tests paramétriques de comparaison de populations, deux didacticiels sont en prise directe avec les techniques que nous avons présentées dans ce support :

1. Les tests univariés
   `http://tutoriels-data-mining.blogspot.com/2008/07/comparaison-de-populations-tests.html`
2. Les tests multivariés
   `http://tutoriels-data-mining.blogspot.com/2008/07/comparaison-de-populations-tests_22.html`

D'autres sont plus ou moins en relation avec la comparaison de populations. Consulter à ce sujet la Catégorie "Statistiques et tests" sur le site (Voir **Catégories des tutoriels**).

Littérature

1. H. Abdi, *Introduction au traitement statistique des données expérimentales*, PUG, 1987.
2. S. Aïvazian, I. Enukov, L. Mechalkine, *Éléments de modélisation et traitement primaire des données*, Mir, 1986.
3. P. Dagnelie, *Principes d'expérimentation - Planification des expériences et analyse de leurs résultats*, Les Presses Agronomiques du Gembloux, 2003. Cet ouvrage est disponible en version électronique sur le site `http://www.dagnelie.be/extextes.html`
4. G. Garson, *Univariate GLM, ANOVA and ANCOVA*, from Statnotes : Topics in Multivariate Analysis, `http://www2.chass.ncsu.edu/garson/pa765/statnote.htm`.
5. B. Grais, *Méthodes statistiques*, Dunod, 2003.
6. W. Guenther, *Analysis of Variance*, Prentice-Hall, 1964.
7. D. Howell, *Méthodes statistiques en sciences humaines*, De Boeck Université, 1998.
8. J. McDonald, *Handbook of Biological Statistics*, `http://udel.edu/~mcdonald/statintro.html`
9. NIST/SEMATECH, *e-Handbook of Statistical Methods*, `http://www.itl.nist.gov/div898/handbook/`
10. S. Rathbun, A. Wiesner, *Applied Multivariate Statistical Analysis (STAT 505)*, The Pennsylvania State University, `http://www.stat.psu.edu/online/development/stat505/`
11. G. Saporta, *Probabilités, Analyse des données et Statistique*, Dunod, 2006.
12. M. Tenenhaus, *Statistique : Méthodes pour décrire, expliquer et prévoir*, Dunod, 2006.
13. R. Veysseyre, *Aide-mémoire - Statistique et probabilités pour l'ingénieur*, Dunod, 2006.
14. C. Wendorf, *Manuals for univariate and multivariate statistics*, `http://www.uwsp.edu/psych/cw/statmanual/index.html`